ОБЗОРНАЯ СТАТЬЯ
REVIEW PAPER

УДК 612.84 004.9 004.946

PAST AND FUTURE APPLICATIONS OF 3-D (VIRTUAL REALITY) TECHNOLOGY

N. Foreman[a, b], L. Korallo[a, b]

[a] Middlesex University, London, NW4 4BT, UK, n.foreman@mdx.ac.uk
[b] ITMO University, Saint Petersburg, 197101, Russian Federation

Abstract. Virtual Reality (virtual environment technology, VET) has been widely available for twenty years. In that time, the benefits of using virtual environments (VEs) have become clear in many areas of application, including assessment and training, education, rehabilitation and psychological research in spatial cognition. The flexibility, reproducibility and adaptability of VEs are especially important, particularly in the training and testing of navigational and way-finding skills. Transfer of training between real and virtual environments has been found to be reliable. However, input device usage can compromise spatial information acquisition from VEs, and distances in VEs are invariably underestimated. The present review traces the evolution of VET, anticipates future areas in which developments are likely to occur, and highlights areas in which research is needed to optimise usage.

Keywords: virtual reality technology, applications, benefits and drawbacks, future applications, future research.

Acknowledgements. The authors are grateful for support from the Russian Scientific Foundation; grant number RNF 14-15-00918, “Optimization of technologies and restoration of cognitive functions of humans in virtual environments”.
The authors would also like to thank several colleagues and friends for interesting and valuable conversations related to Virtual Reality and spatial cognition in Saint Petersburg, in ITMO, the Faculty of Psychology of the State University, and the I.P. Pavlov Institute RAS, particularly Professor Yuri Evgenievich Shelepin, also in the course of a recent lecture visit to Aktobe, Kazakhstan, the President of the Russian-Kazakh University, Berdimuratov Temirkhan Baybosynovich, Dr. Danna Naurzalina, the Head of the Psychology Department Yapparova Gulfiya Muratovna, and members of the Aktobe regional Education Department.

ПРОШЛОЕ И БУДУЩЕЕ 3-D ТЕХНОЛОГИЙ ВИРТУАЛЬНОЙ РЕАЛЬНОСТИ

Н. Форман[a, b] Л. Коралло[a, b]

[a] Миддлсекский университет, Лондон, NW4 4BT, Великобритания, n.foreman@mdx.ac.uk
[б] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

Аннотация. Виртуальная реальность (метод виртуальной среды, VET) широко используется на протяжении двадцати лет. За этот период стали очевидными преимущества использования технологий виртуальной окружающей среды во многих областях, включая оценку рисков и обучение, образование, реабилитацию и психологические исследования познавательной способности в пространстве. Гибкость, воспроизводимость и адаптируемость методов виртуальной среды особенно важны при тренировке способности операторов к ориентации на местности. Однако, к сожалению, использование оператором пространственной информации от входных устройств виртуальной реальности приводит к неизменному преуменьшению оценки расстояния до наблюдаемых объектов. В данной работе выполнен обзор эволюции технологий создания виртуальной среды и прогнозирования вероятных сфер их применения. Подчеркиваются перспективные направления оптимизации использования человеком новых технологий.

Ключевые слова: технологии виртуальной реальности, применение 3D технологий, преимущества и недостатки виртуальной среды, эволюция 3D технологий

Благодарности. Работа выполнена при поддержке Российского научного фонда, грант РНФ 14-15-00918, «Технологии оптимизации и восстановления когнитивных функций человека виртуальной средой».
Авторы выражают благодарность за интересные и содержательные обсуждения вопросов, касающихся виртуальной реальности и пространственных когнитивных способностей, коллегам и друзьям из Университета ИТМО, Санкт-Петербург, факультета психологии СПбГУ и Института физиологии имени И.П. Павлова РАН, особенно профессору Юрию Евгеньевичу Шелепину, а также Президенту Российско-Казахского университета Темирхану Байбосуновичу Бердимуратову, доктору Данне Наурзалиной, заведующей кафедрой психологии Гульфие Муратовне Яппаровой и членам регионального отдела образования г. Актобе.

Introduction

Virtual Reality (VR) is an objective rather than an achievement; the concept of generating a realistic (“real”) environment in which an individual feels a sense of presence is not easily achieved, even using the best software and hardware such as head-immersion helmets. Jaron Lanier first coined the term Virtual Reality some 30 years ago, and although VR has improved in that time, the thought that an individual could place a VR helmet on their head, let alone view a computer screen, and feel “transported” to a computer generated world is not imminently in prospect, however authentic the graphics.

Virtual Reality is, nevertheless, in widespread use (though usage is not as widespread as originally envisaged at the height of VR fever in the 1990’s; see below.) Not surprisingly, VR was especially well

developed by the gaming industry, where financial returns are guaranteed on substantial investments and for military applications where national funding is guaranteed to promote training of personnel involved in defence. There are many reasons why VR is preferred over other types of media. For example, the programmer/designer of a virtual environment is in complete control. Anything that can be imagined can be modelled using the right software. VR is of interest for this reason too: when the senior author ran training courses with Paul Wilson in VR in the 1990s for businesses in Italy and Portugal, we would typically introduce complete novices to very simple VR packages (the forerunners of later packages such as SuperScape) and instruct them on how to create simple objects using a relatively small number of polygons that could be adjusted, coloured and placed in a small VE. Our own experience had been in creating small school environments, rooms and other small spaces that could be navigated using a mouse or keyboard. However, we were amazed at the design skills of "students" working in those industries, who would often create really interesting and original models -- of virtual computers, for example -- using these primitive tools. Therefore, the ability to use VR successfully is determined by both programming skills and also the artistic skills required to make best use of it. Also, since we were also interested to know what were the particular benefits of using VEs in experimental research (to create and to use virtual environments, to navigate and way-find) psychologists could also make good use of VEs to investigate the ways in which people use landmarks and allocentric cues to find their way within a virtual world. VR is essentially "spatial"; indeed some referred to "VR soft-where". This enables all kinds of research possibilities [1]. If a psychologist wishes to investigate participants' establishment of a virtual "cognitive spatial map" of Paris, they could do it in Paris using maps and compasses, and having participants remember routes through the city or make pointing judgments to landmarks (see [2], [3] and [4] for methodologies). However, in a virtual environment one can manipulate the cue layout of the environment in any way, such as swapping the positions of the Eiffel tower and Arc de Triomphe, to determine the importance of each of those landmarks for Parisian way-finding, though this is clearly not a manipulation that it would be possible to use in the real Paris! A recent study has employed VR to investigate activity and passivity in relation to understanding an area of Bordeaux [5].

What is special about VR and why are its unique features useful?

As indicated above, the great benefit of using VR in research is its flexibility. Real worlds can be and artificial environments created from nothing. Moreover, the use of a VE can be repeated over and over again, so that an individual can be trained in an identical but repeated situation; of course difficulty can be titrated to provide a progressive learning experience. In Psychology, it is particularly useful to be able to test groups of people individually knowing that they have experienced precisely the same test situation, by virtue of their having viewed the same display. Some past studies (e.g., [6]; see below) have used passive and active groups for experimental comparisons, the latter group in control of their exploration, while the former group consisted of "yoked" controls, each individually matched to an active participant but observing the screen and thus watching exactly the same spatial displacements as made by the active participant, but without controlling the interface. Doing such a thing in reality is possible but it is difficult.

When VR first appeared, the senior author was personally doubtful about the authenticity of viewing a virtual world for conducting spatial research, not being convinced that people would acquire good quality spatial information from a virtual display particularly since some experimenters in the USA had emphasized the importance of inner ear mechanisms for generating comprehensive "spatial maps" when moving about in space [7]. These concerns were quickly dispelled after a few early experiments, conducted with colleagues in Leicester University [2, 8, 9, 10, 11], showed that both adults and children could view a VE and afterwards make accurate pointing judgements to indicate the directions of distant spatial locations, from various pointing locations to which they were taken within the virtual world. This ability relies on having a "spatial cognitive map" of the environment [12]; they should also be able to take short cuts and make detours in a VE using spatial mapping abilities (see [13] and [14]), and this is just the sort of thing that participants in VR studies have been shown to be able to do. Also, in the early studies and since, there was clearly very good transfer of learning between virtual and real equivalent environments [15, 16], in children and adults, but also, perhaps surprisingly, in children with disabling conditions (that were usually thought to render their spatial judgements poor) and in adults of advanced age whose spatial brain systems, such as the hippocampus [17], were conventionally thought to be compromised [2, 8; see below].

In other words, by and large, people behave in a VE much as they would in the real equivalent environment. This poses the question – one that has not often been addressed in spatial research – of what changes occur in the brain during navigation, changes which can kick in and operate effectively in both the real world but also in a virtual world? Clearly, when sitting passively in a chair, spatial systems in the brain are generally in abeyance; there is no point in computing relations among objects and between self and self-movement and environmental objects, since these relationships and computations only become important when movement occurs, especially movement initiated by the person concerned [18]. The matter could be addressed by recording brain activity using imaging procedures; it would be interesting to know which brain areas "light up" when an individual changes from their stationary state and begins to move about autonomously, for example

on foot. However, since we now realise that virtual movement control is almost as effective as real movement in space, these brain areas should be detectable using timed self-directed movement in a VE. This application of VEs has yet to be investigated, though combining VEs with brain recordings has already been illuminating [19, 20].

Activity and passivity in virtual worlds

VR is clearly a unique medium, and paradoxical insofar as it is essentially passive (since participants are typically seated and view a screen or head immersion screens in a helmet) but it is active in that it engages the participant in self-initiated displacements so that they make active decisions about where to go and what to see and do, just as when they are walking about or driving car autonomously. Activity versus passivity (in both reality and in VR) has attracted a great deal of research attention. The general assumption was always that active exploration would always produce better spatial learning, and some found evidence for that. However the results were always mixed and controversial [6, 21]. What appeared to have escaped their notice is that use of VEs – particularly desktop VEs -- always involves some kind of abnormal participant activity such as moving a mouse or joystick or pressing keyboard keys. The most recent findings have led to the conclusion that interfacing with a VE is effectively a secondary task – a competing parallel task – that occupies some cognitive capacity leaving less available for spatial (working) memory [15, 22, 23]. This should not be a great surprise since "dual task" methodology has been used for many years to investigate spatial working memory capacity [22, 23] but the surprise is that it applies to VE usage and interactivity.

What applications have been successfully demonstrated?

VR has been beneficial in areas where visualisation in reality is not possible, such as illustrating structure and function in the brain or throughout the body, and using virtual manipulations to create novel and illuminating test protocols [24]. Here VR has developed in parallel with other 3-D media and procedures including functional Magnetic Resonance Imaging (fMRI). In early work in Leicester University (see above) the senior author's own research group was able to show that VEs could be used successfully to convey spatial knowledge of a building (such as a school) to children who might otherwise be permanently spatially impaired. Research had shown [25] that children with a variety of conditions that affected their mobility (including spina bifida, cerebral palsy and childhood arthritis) tended to make inaccurate pointing judgments when asked to point (with eyes closed) in the direction of prominent landmarks on their school campus. However, after several days of VR training in a virtual version of a novel school we showed that even children with debilitating conditions could learn a great deal about the novel environment and practically, find their way about successfully when they arrived there. Transfer from virtual to real environments was not quite equivalent, but sufficiently equivalent for all practical purposes (see [2, 9, 10]. In a similar study, using a virtual model of the Astley Clarke Building of Leicester University (at that time, the Psychology Department), we found that students in the department who used a wheelchair for their mobility, and were denied access to the basement area of the department (which had, at that time, no lift installed), told us what a relief it was to be able to visit the basement and to know "what was there". It made them feel that they understood the building in which they were learning. In truth, the basement area was not especially attractive or enlightening, but the example illustrates how limited experience from mobility restriction can have surprising consequences and that this can be overcome using a VE. The latter example illustrates another important point: some of those we tested had brain damage that must have affected spatial areas of the brain (such as hippocampal damage in cerebral palsy) but others did not (childhood arthritis), suggesting that spatial deficits could arise not from brain damage per se but rather secondarily from the consequences of brain damage that restricted mobility and undermined normal exploration and spatial decision-making [25]. Akhutina et al. [26] found that training with VEs successfully enhanced improvement in several cognitive skills in children with a range of brain insults, and in adults Brooks et al. [27] have found that residual route learning skills can be successfully trained in an amnesic patient in a virtual reconstruction of their hospital environment. Mobility training in a wheelchair can benefit from the use of VEs [28]. A great benefit of a VE in training is its safety; participants can make virtual errors and have virtual collisions without injuring anyone, or themselves.

Applications in education

Within education, VR has many potential benefits, for example linking children in widely different locations in mutual social interaction [29]. Special needs education may benefit because children of all abilities can generally access VR software given its intuitive nature [30]. It has been used successfully to enhance the learning of historical chronology. Learning about the passage of time and the correct sequencing of events is arguably an important part of understanding history generally. Time and space are clearly closely related – we might say that the distance between two locations is 4 kilometers or alternatively 20 minutes' walk. Historical timelines can be set up as spatial displays. By setting up a timeline like a row of shops, it is possible to represent successive events in history as individual items (each a "shop frontage") and these can be remembered in sequence just as we might remember a row of shops [31, 32]. This has proved beneficial compared with learning

from booklets or equivalent PowerPoint displays when the design of the environment is appropriate, and suitable for the age of the child [33, 34].

Moderating variables

VR is arguably a sensitive means of investigating subtle influences on behaviours, including sex differences in information processing [35], and especially anxiety effects on spatial behaviours (see [36]). In research just published, Schoenfeld et al. [37] have found that when adult participants with ages ranging from 20-80 years were tested in a sophisticated virtual version of the familiar Water Maze, age was a major predictor of spatial performance, although this was moderated by depression and anxiety scores, while personality traits (particularly Extraversion) predicted weaker spatial perseveration. VR has been used to assess human behaviours in situations such as mazes originally designed for animals [38], and can enable direct comparisons of animal and human performance [39].

Are there drawbacks to using VR?

Students with learning difficulties have been trained in a VE to take a bus route to their college or workplace, getting on and off the bus at the appropriate points and identifying significant landmarks en route and guides to their progress and location. In that case the training was successful. VR training has proved useful in training road crossing skills in children, including children with cognitive challenges [40, 41]. However, although children with autistic spectrum disorders have been trained to identify safe intervals between cars when crossing a busy road, this raised concerns that a child who is somewhat "detached" from reality, they might, after VR training, subsequently fail to distinguish virtual from real and may attempt dangerous road crossings as a result.

Darken and Silbert [42] argued that long periods of exposure were needed to become familiar with an environment in VR, although others have found short periods to be adequate even when using unsophisticated technology such as desk-top VR presentation (e.g., [2, 43, 44]. It should be noted that desk-top presentation is often the medium of choice, given that end users of VR systems are likely to be schools, hospitals and individuals, many unable to afford the luxury of head-immersion and other hardware.

What are the future potential applications?

There are many. Future applications of VEs are dependent upon our better understanding of the way in which the brain processes information from VEs compared to reality. Also, VR may have been held back in the past by the fact that no universal software has been available. In the 1980s and 90's, easy-to-use packages such as SuperScape, easily programmable with draw-down menus, were generally available at a low cost and with support from the company. Many universities adopted SuperScape for its VR research. However, SuperScape withdrew from the academic market, lured by Japanese and American applications in architecture and modelling. Thus some of the anticipated benefits of VEs have never materialized. Applications of VEs must be cost-effective and so for small projects such as on-line visits by potential buyers to houses which they cannot easily visit; this is only cost-effective if the building is very expensive. On the other hand, VEs have been used in industry; motor companies now design vehicles in virtual form before putting them on the production line; totally authentic virtual engines can be modelled and used to give designers and mechanics the most intimate understanding of their working parts and spatial relationships among components as they (virtually) operate. Safety training in buildings, virtual building evacuations, and real-time virtual evacuations from aircraft have all been used, and in a situation that can be anxiety provoking but without asking participants to adopt any real risk. Aircraft evacuation is a good example of the need for better understanding of spatial cognition in virtual and real environments and transfer between the two. Clearly, evacuation from an aircraft is likely to occur in darkness, and underestimation of virtual distances [45], for example between doors and safety equipment, could be crucial. Other applications such as fire fighter training, military training [46], integration of team responses to emergencies, and 3-D training of surgeons using game-like environments [47] have all been used successfully, and again benefit from understanding of the spatial cognitive factors involved. There are many clinical applications (see [48–52], the most successful of which has been the treatment of neurotic disorders such as agoraphobia [53, 54]. VR has been shown to have adjunctive benefits to exercise in stress management [55].

Virtual museums have been created, an application that points up both the limitations and the benefits of VR usage. On one hand, it is unlikely that, even in a future populated by individuals who have grown up surrounded by technological media, VR will effectively or acceptably replace the experience of visiting an actual museum. Nevertheless, it could be especially useful where virtual artefacts cannot usually be investigated at close quarters or handled; studies have used datagloves to allow the manipulation of virtual objects that would usually be viewed from behind glass cases. Various forms of augmented reality become possible, limited only by our imagination.

Conclusions

VET has brought a range of benefits, and has great potential for future development. Although there are issues to be considered when using VEs, it can generally be concluded that information acquisition from VE simulations is reliable and authentic, equivalent to that gained from experience within real environments. Assuming that VE technology remains affordable, there are likely to be many important future applications, in situations where training in reality is dangerous, where real spatial environmental cues cannot be easily manipulated and varied, and where the augmentation of real experience is beneficial.

References

1. Rose F.D., Foreman N.P. Virtual reality. *Psychologist*, 1999, vol. 12, no. 11, pp. 550–554.
2. Foreman N., Stanton D., Wilson P., Duffy H. Spatial knowledge of a real school environment acquired from virtual or physical models by able-bodied children and children with physical disabilities. *Journal of Experimental Psychology: Applied*, 2003, vol. 9, no. 2, pp. 67–74. doi: 10.1037/1076-898X.9.2.67
3. *Handbook of Spatial Research Paradigms and Methodologies. Volume 1: Spatial Cognition in Child and Adult*. Eds. N.P. Foreman, R. Gillett. East Sussex: Psychology Press, 1997.
4. *Handbook of Spatial Research Paradigms and Methodologies. Volume 2: Clinical and Comparative Studies*. Eds. N.P. Foreman, R. Gillett. East Sussex: Psychology Press, 1998, 278 p.
5. Taillaude M., Sauzeon H., Arvind Pala P., Dejos M., Larrue F., Gross C., N'Kaoua B. Age-related wayfinding differences in real large-scale environments: detrimental motor control effects during spatial learning are mediated by executive decline? *PLOS ONE*, 2013, vol. 8, no. 7, art. e67193. doi: 10.1371/journal.pone.0067193
6. Sandamas G., Foreman N. Spatial reconstruction following virtual exploration in children aged 5-9 years: effects of age, gender and activity-passivity. *Journal of Environmental Psychology*, 2007, vol. 27, no. 2, pp. 126–134. doi: 10.1016/j.jenvp.2007.03.001
7. Potegal M., Day M.J., Abraham L. Maze orientation, visual and vestibular cues in two-maze spontaneous alternation of rats. *Physiological Psychology*, 1977, vol. 5, no. 4, pp. 414–420.
8. Foreman N.P., Stanton-Fraser D.E., Wilson P.N., Duffy H., Parnell R. Transfer of spatial knowledge to a two-level shopping mall in older people, following virtual exploration. *Environment and Behavior*, 2005, vol. 37, no. 2, pp. 275–292. doi: 10.1177/0013916504269649
9. Wilson P.N., Foreman N.P., Tlauka M. Transfer of spatial information from a virtual to a real environment in physically disabled children. *Disability and Rehabilitation*, 1996, vol. 18, no. 12, pp. 633–637.
10. Wilson P.N., Foreman N.P., Tlauka M. Transfer of spatial information from a virtual to a real environment. *Human Factors*, 1997, vol. 39, no. 4, pp. 526–531.
11. Wilson P.N., Foreman N., Gillett R., Stanton D. Active versus passive processing of spatial information in a computer-simulated environment. *Ecological Psychology*, 1997, vol. 9, no. 3, pp. 207–222.
12. O'Keefe J., Nadel L. *The Hippocampus as a Cognitive Map*. Oxford University Press, 1978.
13. Tolman E.C. Cognitive maps in rats and men. *Psychological Review*, 1948, vol. 55, no. 4, pp. 189–208. doi: 10.1037/h0061626
14. Tolman E.C., Ritchie B.F., Kalish D. Studies in spatial learning I. Orientation and the short-cut. *Journal of Experimental Psychology*, 1946, vol. 36, no. 1, pp. 13–24. doi: 10.1037/h0053944
15. Foreman N.P. Spatial cognition and its facilitation in special populations. In: *Applied Spatial Cognition: From Research to Cognitive Technology*. Ed. G. Allen. Mahwah, NJ, Lawrence Erlbaum, 2006, pp. 129–178.
16. McComas J., Pivik J., Laflamme M. Children's transfer of spatial learning from virtual reality to real environments. *Cyberpsychology and Behavior*, 1998, vol. 1, no. 2, pp. 121–128.
17. Maguire E.A., Frackiowiak R.S.J., Frith C.D. Recalling routes around London: activation of the right hippocampus in taxi drivers. *Journal of Neuroscience*, 1997, vol. 17, no. 18, pp. 7103–7110.
18. Gibson J.J. *The Ecological Approach to Visual Perception*. Boston, Houghton Mifflin, 1979, 332 p.
19. Gron G., Wunderlich A.P., Spitzer M., Tomczak R., Riepe M.W. Brain activation during human navigation: gender-different neural networks as substrate of performance. *Nature Neuroscience*, 2000, vol. 3, no. 4, pp. 404–408. doi: 10.1038/73980
20. Hoffman H. G., Richards T., Coda B., Richards A., Sharar S.R. The illusion of presence in immersive virtual reality during an fMRI brain scan. *Cyberpsychology and Behavior*, 2003, vol. 6, no. 2, pp. 127–131.
21. Sandamas G., Foreman N., Coulson M. Interface familiarity restores active advantage in a virtual exploration and reconstruction task in children. *Spatial Cognition and Computation*, 2009, vol. 9, no. 2, pp. 96–108. doi: 10.1080/13875860802589202
22. Sandamas G., Foreman N.P. Spatial demands of concurrent tasks can compromise spatial learning of a virtual environment: Implications for active input control. *Sage Open*, 2014, vol. 4, no. 1. doi: 10.1177/2158244014525424

23. Sandamas G., Foreman N. Driver status and spatial acquisition from a virtual environment // *Sage Open*, 2014. (in press).
24. Stirk J.A., Foreman N. Assessment of visual-spatial deficits in patients with early Parkinson's disease and closed head injuries using virtual environments. *Cyberpsychology and Behavior*, 2005, vol. 8, no. 5, pp. 431–440. doi: 10.1089/cpb.2005.8.431
25. Foreman N.P., Gell M. Kids in space: handicapped children's spatial knowledge of their mainstream school environment. *Special Children*, 1990, no. 1.
26. Akhutina T., Foreman N., Krichevets A., Matikka L., Narhi V., Pylaeva N, Vahakuopus J. Improving spatial functioning in children with cerebral palsy using computerised and traditional game tasks. *Disability and Rehabilitation*, 2005, vol. 25, no. 24, pp. 1361–1371. doi: 10.1080/09638280310001616358
27. Brooks B.M., McNeil J.E., Rose F.D., Greenwood R.J., Attree E.A., Leadbetter A. G. Route learning in a cased of amnesia: a preliminary investigation into the efficacy of training in a virtual environment. *Neuropsychological Rehabilitation*, 1999, vol. 9, no. 1, pp. 63–76.
28. Harrison A., Derwent G., Enticknap A., Rose F.D., Attree E.A. The role of virtual reality technology in the assessment and training of inexperienced powered wheelchair users. *Disability and Rehabilitation*, 2002, vol. 24, no. 11–12, pp. 599–606. doi: 10.1080/09638280110111360
29. Bailey F., Moar M. The VERTEX project: designing and populating shared 3D virtual worlds in the primary (elementary) classroom. *Computers and Graphics (Pergamon)*, 2003, vol. 27, no. 3, pp. 353–359. doi: 10.1016/S0097-8493(03)00030-X
30. Foreman N.P. Finding a place for virtual reality in special needs education. *Themes in Education*, 2000, vol. 1, pp. 391–408.
31. Foreman N., Boyd-Davis S., Moar M., Korallo L., Chappell E. Can virtual environments be used to enhance the learning of historical chronology? *Instructional Science*, 2008, vol. 36, no. 2, pp. 155–173. doi: 10.1007/s11251-007-9024-7
32. Foreman N., Korallo L., Newson D., Sarantos N. The incorporation of challenge enhances the learning of chronology from a virtual display. *Journal of Virtual Reality*, 2008, vol. 12, no. 2, pp. 107–113. doi: 10.1007/s10055-007-0078-2
33. Korallo L., Foreman N., Boyd-Davis S., Moar M., Coulson M. Do challenge, task experience and computer familiarity influence the learning of historical chronology from virtual environments in 8-9 year old children? *Computers and Education*, 2012, vol. 58, no. 3, pp. 1106–1116. doi: 10.1016/j.compedu.2011.12.011
34. Korallo L., Foreman N., Boyd-Davis S., Moar M., Coulson M. Can multiple “spatial” virtual timelines convey the relatedness of chronological knowledge across parallel domains? *Computers and Education*, 2012, vol. 58, no. 2, pp. 856–862. doi: 10.1016/j.compedu.2011.10.011
35. Dabbs Jr. J.M., Chang E.-L., Strong R.A., Milun R. Spatial ability, navigational strategy and geographic knowledge among men and women. *Evolution and Human Behavior*, 1998, vol. 19, no. 2, pp. 89–98.
36. Robillard G., Bouchard S., Fournier T., Renaud P. Anxiety and presence during VR immersion: a comparative study of the reactions of phobic and non-phobic participants in therapeutic virtual environments derived from computer games. *Cyberpsychology and Behavior*, 2003, vol. 6, no. 5, pp. 467–476. doi: 10.1089/109493103769710497
37. Schoenfeld R., Foreman, N.P., Leplow B. Ageing and spatial reversal learning in humans: findings from a virtual water maze. *Behavioural Brain Research*, 2014, vol. 270, pp. 47–55. doi: 10.1016/j.bbr.2014.04.036
38. Foreman N.P., Stirk J., Pohl J., Mandelkow L., Lehnung M., Herzog A., Leplow B. Spatial information transfer from virtual to real versions of the Kiel locomotor maze. *Behavioral Brain Research*, 2000, vol. 112, no. 1–2, pp. 53–61.
39. Schoenfeld R. A comparison of the performance of animals on the Water Maze and humans on a virtual version (paper in preparation).
40. Thomson J.A., Tolmie A.K., Foot H.C., Whelan K.M., Sarvary P., Morrison S. Influence of virtual reality training on the roadside crossing judgments of child pedestrians. *Journal of Experimental Psychology: Applied*, 2005, vol. 11, no. 3, pp. 175–186. doi: 10.1037/1076-898X.11.3.175
41. Clancy T.A., Rucklidge J.J., Owen D. Road crossing safety in virtual reality: a comparison of adolescents with and without ADHD. *Journal of Clinical Child and Adolescent Psychology*, 2006, vol. 35, no. 2, pp. 203–215. doi: 10.1207/s15374424jccp3502_4
42. Darken R.P., Silbert J.L. Navigating large virtual spaces. *Plastics, Rubber and Composites Processing and Applications*, 1996, vol. 8, no. 1, pp. 49–71.
43. Tlauka M., Brolese A., Pomeroy D., Hobbs W. Gender differences in spatial knowledge acquired through simulated exploration of a virtual shopping centre. *Journal of Environmental Psychology*, 2005, vol. 25, no. 1, pp. 111–118. doi: 10.1016/j.jenvp.2004.12.002

44. Waller D.I. Individual differences in spatial learning from computer-simulated environments. *Journal of Experimental Psychology: Applied*, 2000, vol. 6, no. 4, pp. 307–321.
45. Foreman N.P., Sandamas G., Newson D. Distance underestimation in virtual space is sensitive to gender but not activity-passivity or mode of interaction. *Cyberpsychology and Behavior*, 2004, vol. 7, no. 4, pp. 451–457. doi: 10.1089/cpb.2004.7.451
46. Lampton D.R., Clark B.R., Knerr B.W. Urban combat: the ultimate extreme environment. *Journal of Society for Human Performance in Extreme Environments*, 2003, vol. 7, pp. 57–62.
47. Gallagher A.G., Ritter E.M., Champion H., Higgins G., Fried M.P., Moses G., Smith C.D, Satava R.M. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Annals of Surgery*, 2005, vol. 241, no. 2, pp. 364–372. doi: 10.1097/01.sla.0000151982.85062.80
48. Riva G. Virtual reality in neuroscience: a survey. *Studies in Health Technology and Informatics*, 1998, vol. 58, pp. 191–199. doi: 10.3233/978-1-60750-902-8-191
49. Riva G., Gaggioli A., Villani D., Preziosa A., Morganti F., Corsi R., Faletti G.,Vezzadini L. NeuroVR: an open-source virtual reality tool for research and therapy. *Proc. 15th Annual Medicine Meets Virtual Reality Conference*. Long Beach, California, 2007.
50. Rizzo A.A., Schultheis M.T. Expanding the boundaries of psychology: the application of virtual reality. *Psychological Enquiry*, 2002, vol. 13, no. 2, pp. 134–140.
51. Wiederhold B.K., Wiederhold M.D. Lessons learned from 600 virtual reality sessions. *Cyberpsychology and Behavior*, 2000, vol. 3, no. 3, pp. 393–400. doi: 10.1089/10949310050078841
52. Wiederhold B.K., Wiederhold M.D. *Virtual Reality Therapy for Anxiety Disorders: Advances in Evaluation and Treatment*. US: American Psychological Association, 2005, 225 p.
53. Cardenas G., Munoz S., Gonzalez M., Uribarren G. Virtual reality applications to agoraphobia: a protocol. *Cyberpsychology and Behavior*, 2006, vol. 9, no. 2, pp. 248–250. doi: 10.1089/cpb.2006.9.248
54. Gershon J., Anderson P., Graap K., Zimand E., Hodges L., Rothbaum B.O. Virtual reality exposure therapy in the treatment of anxiety disorders. *The Scientific Review of Mental Health Practice*, 2004, vol. 1, pp. 76–81.
55. Plante T.G., Cage C., Clements S., Stover A. Psychological benefits of exercise paired with virtual reality: outdoor exercise energizes whereas indoor virtual exercise relaxes. *International Journal of Stress Management*, 2006, vol. 13, pp. 108–117. doi: 10.1037/1072-5245.13.1.108

Nigel Foreman – BSc (Bradford), PhD (Nottingham), Hon. Doc (Novi Sad), is Professor emeritus in Psychology at Middlesex University, London, UK. He has over 200 publications in international books and journals and has held over £1.5 million in research grants. He was the Chair of the International Committee of the British Psychological Society (BPS) for 8 years and sat on several other Boards and Committees, including the BPS Council, Ethics Committee, President's Award and Lifelong Achievement Award Committees. He led a delegation of international psychologists to China and Vietnam in 2004. He has represented BPS on the Scientific Affairs Board of the European Federation of Psychologists Associations (EFPA) and chaired the Board between 2007 and 2012. He has held the post of Professor in Psychology in Saint Petersburg University, Russia, and holds an Honorary Doctorate from the University of Novi Sad, Serbia. He has supervised student exchanges between the UK and America, Russia, Serbia, Macedonia and The Czech Republic, and arranged publishing contracts between the BPS and Russian publishers. Nigel's established contacts with influential academics across the world and his high reputation allows him of giving authoritative advice on commercial exchanges, also educational exchanges and integration, across Europe, Russia and beyond.

Найджел Форман – BSc (Bradford), PhD (Nottingham), Hon. Doc (Novi Sad). Удостоен степени почетного профессора психологии Миддлсекского университета. Результаты его исследовательской деятельности активно публикуются в научных периодических изданиях, а также в образовательных пособиях. Его деятельность активно поддерживалась академическими и государственными сообществами в виде грантов на исследования, которые суммарно составили 1,5 миллионами фунтов стерлингов. В течение 8 лет он занимал пост председателя Международного комитета Британского психологического общества и являлся членом нескольких других советов и комитетов, включая Совет Британского психологического общества, Комитет по этике, комитеты Премии премьер-министра и Премии «За дело всей жизни». Он был главой международной делегации психологов в Китае и Вьетнаме в 2004 г., а также представлял Британское психологическое общество в Ученом совете Европейской федерации ассоциаций психологов и председательствовал в Совете с 2007 по 2012 гг. Он занимал пост профессора психологии в Санкт-Петербургском университете в России, имеет почетную докторскую степень университета г. Нови Сад,

Сербия. Он выступал в качестве инспектора при обмене студентами между Великобританией и США, Россией, Сербией, Македонией и Чешской Республикой, а также обеспечил заключение издательских договоров между Британским психологическим обществом и российскими издателями. Профессиональные контакты Найджела с влиятельными академиками по всему миру и его высокая репутация позволяет ему давать авторитетные советы по вопросам коммерческого обмена, образовательного обмена (обмена студентами) и научно-исследовательской интеграции между академическими центрами по всему миру.

Liliya Korallo – PhD, Practitioner Psychologist, Chartered Psychologist, Cognitive Behavioural Therapist, has been an active contributor to educational activities in the UK for the last 15 years. She has graduated from Middlesex University and University College in London (UCL). Liliya has completed her PhD entitled “The use of Virtual Environments in the teaching of historical chronology”. Her research has been financed by the Leverhulme Trust and Halle University, Germany; her findings have been published in international peer-reviewed journals (Instructional Science, Journal of Virtual Reality, Computers and Education and many others). Moreover, her papers have been presented at international and national conferences in Great Britain, Vietnam, Germany, Greece, the USA, Norway, Sweden, Serbia, Ukraine and Turkey. In June 2012 her research received a high accolade from such authoritative editions as the Times Educational Supplement and The Psychologist. She has just contributed a chapter to an influential book "On Human Visualization", which is due to be published by the leading Springer publishing house. Liliya has been collaborating on projects with Germany (“Use of VEs in clinical psychology”), Croatia (“Invited young academics to the UK”), Vietnam (“Culture and mental health”), Slovenia, and Ukraine (Integration of the Ukrainian Psychological Society into EFPA). She is constantly been invited as a reviewer for different S&R projects of West Europe.

Лилия Коралло – PhD. Научный сотрудник, психолог, Миддлсекский университет. Практикующий психолог высшей квалификации, специалист в области терапии когнитивных и поведенческих нарушений. На протяжении последних 15 лет Лилия активно занимается образовательной деятельностью в Великобритании. Она окончила Миддлсекский университет и Университетский колледж в Лондоне. Лилия защитила кандидатскую диссертацию на тему «Использование виртуального пространства при изучении исторической хронологии». Ее работы финансировались Фондом Leverhulme и Университетом Галле (Германия); результаты ее экспериментов были широко освещены международными рецензируемыми журналами: Instructional Science, Journal of Virtual Reality, Computers and Education и многими другими. Помимо этого, результаты ее исследований были представлены на многочисленных международных и национальных конференциях в Великобритании, Вьетнаме, Германии, Греции, Норвегии, Сербии, США, Турции и Украине. В июне 2012 исследования д-ра Коралло получили высокую оценку таких авторитетных изданий, как Times Educational Supplement и The Psychologist. Она является соавтором книги «О зрительном восприятии человека», опубликованной ведущим издательством Springer. Г-жу Коралло регулярно приглашают в качестве эксперта в различные научно-исследовательские проекты Западной Европы.

Nigel Foreman – BSc (Bradford), PhD (Nottingham), Hon. Doc (Novi Sad), Emeritus Professor, Middlesex University, London, NW4 4BT, UK; International Laboratory of Neurophysiology of Virtual Reality, ITMO University, Saint Petersburg, 197101, Russian Federation; n.foreman@mdx.ac.uk

Liliya Korallo – PhD, Practitioner Psychologist, Chartered Psychologist, Cognitive Behavioural Therapist, Psychologist, Middlesex University, London, NW4 4BT, UK; International Laboratory of Neurophysiology of Virtual Reality, ITMO University, Saint Petersburg, 197101 Russian Federation; lkorallo@yahoo.co.uk

Найджел Форема́н – доктор наук, профессор, почетный профессор, Миддлсекский университет, NW4 4BT, Лондон, Великобритания; Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; n.foreman@mdx.ac.uk

Лилия Коралло – PhD, научный сотрудник, психолог, Миддлсекский университет, Лондон, NW4 4BT, Великобритания; Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация;lkorallo@yahoo.co.uk

Принято к печати 30.06.14
Accepted 30.06.14

1 ФОТОНИКА И ОПТОИНФОРМАТИКА
PHOTONICS AND OPTOINFORMATICS

УДК 535.3

ЭФФЕКТ ПЁРСЕЛЛА В ПРЕДЕЛЬНО АНИЗОТРОПНЫХ ЭЛЛИПТИЧЕСКИХ МЕТАМАТЕРИАЛАХ

А.В. Чебыкин[a], А.А. Орлов[a], Ф. Хайслер[b], К.В. Барышникова[a], П.А. Белов[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, chebykin.alexandr@gmail.com

[b] Йенский университет имени Фридриха Шиллера, Йена, 07737, Германия

Аннотация. Теоретически продемонстрирован эффект Пёрселла в предельно анизотропных метаматериалах с эллиптической изочастотной поверхностью. В отличие от гиперболических метаматериалов, данный эффект не связан с расходящейся плотностью состояний. Показано, что большой фактор Пёрселла можно наблюдать и без возбуждения мод с большими волновыми векторами в одном из направлений, при этом нормальная слоям компонента волнового вектора будет меньше k_0. Это дает возможность получить в данных материалах увеличение не только мощности, излучаемой в среду, но и мощности, передаваемой в свободное пространство через границу среды, расположенную поперечно слоям материала. Методами анализа являлись построение изочастотных контуров, а также расчет аналитической зависимости фактора Пёрселла от частоты для бесконечной структуры слоистого метаматериала. В диапазоне видимого света сильная пространственная дисперсия не позволяет получить усиление спонтанного излучения в метаматериале с двухслойной элементарной ячейкой. Эффект может быть реализован в периодических слоистых металлодиэлектрических наноструктурах с элементарной ячейкой, содержащей два металлических и два диэлектрических слоя. Анализ полученных зависимостей фактора Пёрселла от частоты показывает, что спонтанное излучение усиливается на порядок и более только для случая ориентации возбуждающего диполя вдоль слоев метаматериала, а для случая поперечной ориентации излучение максимально может усиливаться лишь в 2–3 раза. Результаты работы могут быть использованы для создания нового типа метаматериалов с эллиптическими изочастотными контурами, обеспечивающих более эффективное излучение света в дальнее поле.

Ключевые слова: метаматериалы, эффект Пёрселла, спонтанное излучение, поверхностные плазмоны.

Благодарности. Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01), фонда РФФИ (проект 14-02-31720), а также стипендии Президента Российской Федерации (грант СП-2154.2012.1). Авторы приносят отдельную благодарность анонимному рецензенту за ценные замечания и помощь в подготовке статьи.

PURCELL EFFECT IN EXTREMELY ANISOTROPIC ELLIPTIC METAMATERIALS

A.V. Chebykin[a], A.A. Orlov[a], F. Heisler[b], K.V. Baryshnikova[a], P.A. Belov[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, chebykin.alexandr@gmail.com

[b] Friedrich-Schiller-Universität Jena, Jena, 07737, Germany

Abstract. The paper deals with theoretical demonstration of Purcell effect in extremely anisotropic metamaterials with elliptical isofrequency surface. This effect is free from association with divergence in density of states unlike the case of hyperbolic metamaterials. It is shown that a large Purcell factor can be observed without excitation of modes with large wave vectors in one direction, and the component of the wave vector normal to the layers is less than k_0. For these materials the possibility is given for increasing of the power radiated in the medium, as well as the power radiated from material into free space across the medium border situated transversely to the layers. We have investigated isofrequency contours and the dependence of Purcell factor from the frequency for infinite layered metamaterial structure. In the visible light range strong spatial dispersion gives no possibility to obtain enhancement of spontaneous emission in metamaterial with unit cell which consists of two layers. This effect can be achieved in periodic metal-dielectric layered nanostructures with a unit cell containing two different metallic layers and two dielectric ones. Analysis of the dependences for Purcell factor from the frequency shows that the spontaneous emission is enhanced by a factor of ten or more only for dipole orientation along metamaterial layers, but in the case of the transverse orientation radiation can be enhanced only 2-3 times at most. The results can be used to create a new type of metamaterials with elliptical isofrequency contours, providing a more efficient light emission in the far field.

Keywords: metamaterials, Purcell effect, spontaneous radiation, surface plasmons.

Acknowledgements: The work is partially financially supported by the Government of the Russian Federation (grant 074-U01),the Russian Foundation for Basic Research (project 14-02-31720),and the Russian Federation President scholarship (grant СП-2154.2012.1). The authors express their special thanks to an anonymous reviewer for valuable remarks and rendering assistance in paper preparation.

Введение

В настоящее время гиперболические метаматериалы рассматриваются в качестве наиболее подходящих структур для увеличения силы взаимодействия света с веществом в широком диапазоне [1–4]. Такие среды характеризуются тем, что имеют эффективные диэлектрические проницаемости разных знаков, т.е. $\varepsilon_{xx} = \varepsilon_{yy} < 0$ и $\varepsilon_{zz} > 0$ (где ε_{xx}, ε_{yy} и ε_{zz} – диагональные компоненты тензора диэлектрической проницаемости) [5]. Плотность состояний фотонов в таких материалах расходится благодаря гиперболическому

виду изочастотных поверхностей, что обеспечивает сильный эффект Пёрселла [6]. Обычно эффект Пёрселла связан и со скоростью излучательного затухания, и с мощностью, излучаемой в окружающее пространство [7]. Тем не менее, переход от эллиптического режима к гиперболическому сопровождается сокращением времени жизни в возбужденном состоянии и уменьшением наблюдаемой интенсивности излучения в свободное пространство [8, 9]. Причина заключается в том, что спонтанное излучение в гиперболических метаматериалах возникает в основном из-за возбуждения мод с такими волновыми векторами, у которых обе компоненты (вдоль и поперек оптической оси) превышают волновое число свободного пространства. Такие пространственные гармоники при преломлении через границу материала в свободное пространство становятся эванесцентными в обоих практических случаях расположения границы – как вдоль оптической оси, так и ортогонально ей. Эванесцентные волны ввиду экспоненциального затухания при удалении от границы не могут быть детектированы в дальнем поле [10].

В настоящей работе мы теоретически рассматриваем возможность реализации сильного эффекта Пёрселла в предельно анизотропных эллиптических метаматериалах, где выполняется соотношение

$$0 < \varepsilon_{xx} = \varepsilon_{yy} << \varepsilon_{zz} . \quad (1)$$

В таких средах изочастотная поверхность имеет вид сильно сплющенного эллипсоида (рис. 1). Условие (1) выполняется по одну сторону от топологического перехода между гиперболическим и эллиптическим режимами одного и того же материала (под эллиптическим и гиперболическим режимами здесь понимается форма изочастотных контуров материала). Данный переход происходит на определенной частоте. Далее мы покажем, что анизотропный эллиптический метаматериал может обладать большими значениями фактора Пёрселла, несмотря на конечную величину плотности фотонных состояний. Более того, в этом случае по одному из направлений распространения (по оси *Z*) волны с большими волновыми векторами не вносят вклад в спонтанное излучение, поэтому излученная волна может покинуть структуру. В результате этого увеличивается как скорость излучательного затухания (что есть и в гиперболических метаматериалах), так и излучаемая через границу по плоскости *XZ* в свободное пространство мощность (чего в гиперболических материалах с плоской границей при ее обычном расположении относительно оптической оси не наблюдается ни экспериментально, ни теоретически).

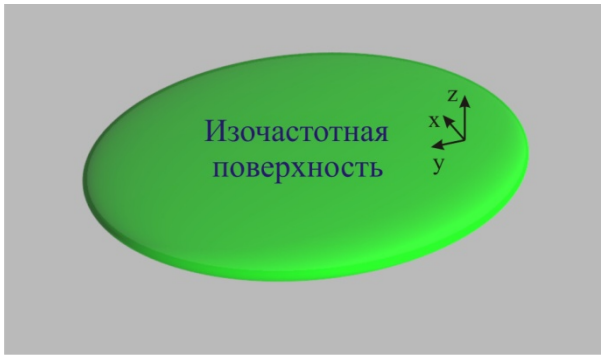


Рис. 1. Изочастотная поверхность материала в предельно анизотропном эллиптическом режиме

Физика эффекта Пёрселла в эллиптических метаматериалах

Причину усиления спонтанного излучения в среде, удовлетворяющей условию (1), можно наиболее просто описать при помощи золотого правила Ферми:

$$\frac{1}{\tau} = \frac{2\pi}{\hbar}\sum_{k,\sigma}\left|d \cdot E_{k,\sigma}\right|^2 \delta(\hbar\omega_{k,\sigma} - \hbar\omega_0) . \quad (2)$$

Здесь τ – скорость излучательного затухания; ω_0 – частота излучения двухуровневой системы; d – дипольный матричный элемент; $E_{k,\sigma}$ – оператор амплитуды электрического поля, соответствующий единичному кванту излучения; $\hbar$ – постоянная Планка, δ – дельта-функция. Суммирование осуществляется по волновому вектору k излучаемой волны. Символом σ обозначена поляризация (TE либо TM). Электрическое поле может быть записано в следующем виде:

$$E_{k,\mathrm{TM}} = \sqrt{\frac{2\pi\hbar\omega_{k,\mathrm{TM}}}{V_{mode,\mathrm{TM},k}}}\left(\cos\theta_k \,\hat{\varphi} - \frac{\varepsilon_{xx}}{\varepsilon_{zz}}\sin\theta_k \,\hat{z}\right) ,$$

где V_{mode} – это эффективный объем моды, θ и φ – сферические координаты волнового вектора k. V_{mode} определяется из условия квантования для плоской волны $VE_k(\hat{\varepsilon}E_k) = 2\pi\hbar\omega_{k,\mathrm{TM}}$. Здесь V – нормированный объем. Эффективный объем моды может быть выражен через эффективный показатель преломления n_{TM}:

$$V_{\mathrm{mode,TM},k} = V\varepsilon_{xx}^2 / n_{k,\mathrm{TM}}^2(\theta_k) , \quad (3)$$

n_{TM}, в свою очередь, определяется дисперсией ТМ-мод: $\omega_{k,\mathrm{TM}} = ck / n_{\mathrm{TM}}(\theta_k)$, где c – скорость света в вакууме.

Выражение для эффективного показателя преломления может быть записано в виде

$$n_{k,\mathrm{TM}}(\theta_k) = \left(\frac{\sin^2 \theta_k}{\varepsilon_{zz}} + \frac{\cos^2 \theta_k}{\varepsilon_{xx}} \right)^{-1/2} . \quad (4)$$

Подставив уравнения (3) и (4) в уравнение (2) и осуществив интегрирование по k и φ, получим вклад ТМ-мод в скорость излучательного затухания для поляризации по оси X:

$$\frac{1}{\tau_{x,\mathrm{TM}}} = \frac{d^2}{2\hbar}\left(\frac{\omega_0}{c}\right)^3 \int_0^{\pi} \frac{d\theta_k \sin\theta_k \cos^2\theta_k n_{\mathrm{TM}}^5(\theta_k)}{\varepsilon_{xx}^2} = \frac{d^2}{3\hbar}\left(\frac{\omega_0}{c}\right)^3 \frac{\varepsilon_{zz}}{\sqrt{\varepsilon_{xx}}} . \quad (5)$$

Уравнение (5) расходится при $\varepsilon_{xx} = 0$. Эта расходимость имеет место на границе гиперболического ($\varepsilon_{xx} < 0$) и эллиптического ($0 < \varepsilon_{xx} << \varepsilon_{zz}$) режима, поэтому природа возникновения усиления излучения в данном случае совершенно отличается от гиперболического случая. В частности, в предельно анизотропном эллиптическом режиме ($\varepsilon_{xx} << \varepsilon_{zz}$) в спонтанном излучении преобладают волны, распространяющиеся в направлении плоскости симметрии, $\theta_k \approx \pi/2$:

$$\frac{1}{\tau_{x,\mathrm{TM}}} = \frac{d^2}{2\hbar}\left(\frac{\omega_0}{c}\right)^3 \sqrt{\varepsilon_{xx}} \int_{-\infty}^{\infty} \frac{\psi^2 d\psi}{|\psi^2 + \varepsilon_{xx}/\varepsilon_{zz}|^{5/2}} , \quad (6)$$

где $\psi = \theta_k - \pi/2$. Резкий максимум в уравнении (6) при $\psi \sim \sqrt{\varepsilon_{xx}/\varepsilon_{zz}} << 1$ обусловлен расхождением эффективного показателя преломления в уравнении (4), что соответствует стремлению эффективного объема моды в уравнении (3) к 0. Малый эффективный объем моды приводит к сильному эффекту Пёрселла аналогично данному эффекту в резонаторе.

Учтем теперь вклад от ТЕ-мод (он определяется эффективным показателем преломления $n_{\mathrm{TE}} = \sqrt{\varepsilon_{xx}}$) и нормируем результат на скорость излучательного затухания в свободном пространстве $1/\tau = 4d^2\omega^3/(3\hbar c^3)$. Тогда выражения для фактора Пёрселла в случае излучателей, поляризованных по осям X и Z, можно записать в следующей форме:

$$f_{purc,x} = f_{purc,y} = \frac{\varepsilon_{zz}}{\sqrt{\varepsilon_{xx}}} + \frac{3\sqrt{\varepsilon_{xx}}}{4} , \quad (7)$$

$$f_{purc,z} = \sqrt{\varepsilon_{xx}} . \quad (8)$$

Первый член в уравнении (7) представляет собой вклад от ТМ-волн в скорость спонтанного излучения, и он может быть существенно больше в случае предельно анизотропного эллиптического режима. Стоит также заметить, что относительно большой рост фактора Пёрселла по сравнению с единицей будет наблюдаться только для излучателей, поляризованных в плоскости XY.

Эффект Пёрселла в предельно анизотропном металлодиэлектрическом метаматериале

Далее перейдем к рассмотрению возможного варианта реализации данного эффекта. Наиболее естественным подходом является применение плазмонного многослойного металлодиэлектрического метаматериала [11], состоящего из периодически чередующихся слоев металла и диэлектрика. Обычно такие структуры описываются в эффективной модели следующими диэлектрическими проницаемостями [12]:

$$\varepsilon_{xx}^{(eff)} = \varepsilon_{yy}^{(eff)} = \langle \varepsilon(z) \rangle = \frac{\varepsilon_{Me} d_{Me} + \varepsilon_{diel} d_{diel}}{d_{Me} + d_{diel}} ,$$

$$\varepsilon_{zz}^{(eff)} = \langle \varepsilon^{-1}(z) \rangle^{-1} = \left(\frac{d_{Me}/\varepsilon_{Me} + d_{diel}/\varepsilon_{diel}}{d_{Me} + d_{diel}} \right)^{-1} ,$$

где угловые скобки обозначают пространственное усреднение; d_{diel}, d_{Me} – толщины диэлектрического и металлического слоев; ε_{diel}, ε_{Me} – их диэлектрические проницаемости. Анализ данных уравнений показал, что условие в уравнении (1) удовлетворяется на частоте, немного превышающей частоту перехода от эллиптического к гиперболическому режиму ω^*,

$$\varepsilon_{Me}(\omega^*) = -\frac{d_{diel}}{d_{Me}} \varepsilon_{diel}(\omega^*) ,$$

когда $\varepsilon_{xx}^{(eff)}$ обращается в 0. Из условия $\varepsilon_{xx}^{(eff)}(\omega^*) > 0$ следует, что для выполнения соотношения $d_{Me} < d_{diel}$ металлические слои должны быть тоньше диэлектрических. В частности, это значит, что абсолютное значение ε_{Me} больше ε_{diel} на частоте ω^*, а также то, что частота ω^* ниже частоты поверхностного плазмо-

на, определяемой условием $\varepsilon_{Me} = -\varepsilon_{diel}$. Это приводит к тому, что в данном случае изочастотный контур содержит не только эллиптическую, но также и дополнительную гиперболическую составляющую из-за эффектов сильной пространственной дисперсии [13]. Следовательно, в данном случае в оптическом диапазоне невозможно получить эффект усиления спонтанного излучения при эллиптической изочастоте метаматериала.

Для получения изолированного эллиптического изочастотного контура мы предлагаем использовать плазмонные многослойные структуры со сложной элементарной ячейкой, обладающей бипериодичностью [14]. В частности, мы рассмотрели систему с четырьмя слоями в элементарной ячейке – двумя разными металлическими слоями и двумя одинаковыми диэлектриками, как показано на рис. 2.

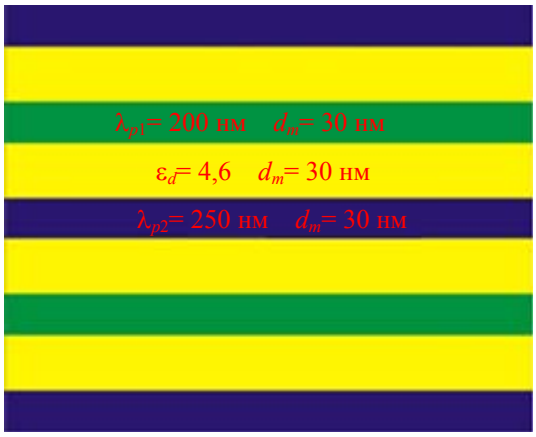


Рис. 2. Структура рассматриваемого металлодиэлектрического метаматериала с элементарной ячейкой, состоящей из двух чередующихся слоев диэлектрика с диэлектрической проницаемостью 4,6 и двух слоев из металлов с разными плазменными длинами волн

Диэлектрическая проницаемость диэлектриков ε_d была выбрана равной 4,6. Диэлектрическая проницаемость металлических слоев соответствовала модели Друде с разными плазменными длинами волн $\lambda_{p1} = 250$ нм и $\lambda_{p2} = 200$ нм. Толщины всех слоев равнялись 30 нм.

В настоящей работе мы аналитически построили спектральную зависимость фактора Пёрселла для излучателя, помещенного в центр диэлектрического слоя в бесконечной периодической структуре, описанной выше. В расчетах использовался стандартный математический аппарат функций Грина для слоистых систем [15]. Результаты представлены на рис. 3. Сплошная синяя кривая соответствует случаю, когда излучатель расположен вдоль слоев, красная пунктирная кривая – случаю перпендикулярной ориентации излучателя относительно слоев. На рис. 3, б–г, показаны изочастотные контуры на трех фиксированных частотах. Частота 570 ТГц соответствует главному результату: как видно из рисунка, на этой частоте изочастотный контур имеет сильно анизотропный эллиптический вид, а фактор Пёрселла достигает значения 24 (синяя линия). Для частот ниже 570 ТГц большая часть изочастотного контура находится в диапазоне волн, которые при преломлении через вертикальную границу среды (ориентированную поперечно слоям) оказываются распространяющимися в свободном пространстве: $k_z < k_0 = \omega/c$. Энергия этих волн может эффективно излучаться в дальнее поле через границу материала, а пространственный спектр таких волн в предельно анизотропном эллиптическом режиме довольно широк, потому что горизонтальная полуось эллипса $k_{y\max} >> k_0$, причем большая нормальная к границе компонента волнового вектора не препятствует преломлению (эффект полного внутреннего отражения полностью связан с k_z). Однако здесь требуется уточнение. Энергия пространственных гармоник, создаваемых оптически коротким диполем, распределена по изочастотному контуру неравномерно. Так, у дипольного излучателя, ориентированного нормально к слоям и создающего поэтому в основном вертикальную поляризацию электрического поля, превалируют пространственные гармоники с малыми значениями k_y, потому что именно у гармоник $k_y << k_z$ поле поляризовано практически вдоль z. Для такого излучателя вытянутый вдоль оси y изочастотный контур особого значения не имеет. Большие значения фактора Пёрселла по сравнению с единицей в эллиптическом режиме наблюдаются только для излучателя, ориентированного параллельно слоям. В полном соответствии с физикой в эллиптическом режиме, который наблюдается в диапазоне 550–570 ТГц, фактор Пёрселла для поперечно ориентированного диполя сравнительно мал (красный пунктир), а для тангенциального диполя весьма велик (15–20). Такая зависимость фактора Пёрселла от поляризации, полученная в результате точных симуляций, согласуется с предсказаниями эффективной среды в уравнениях (7), (8). На рис. 3, а, зеленым штрих-пунктиром показан результат расчета дипольного излучения в модели эффективной среды, полученный с использованием значений ε_{xx} и ε_{zz}, извлеченных из изочастотных контуров. Функция в уравнении (7) умножена на 1,4, что может быть объяснено поправ-

кой локального поля [16]. Данное полуаналитическое выражение хорошо описывает полученную численно частотную зависимость для фактора Пёрселла (см. синюю и зеленую кривую на рис. 3, а).

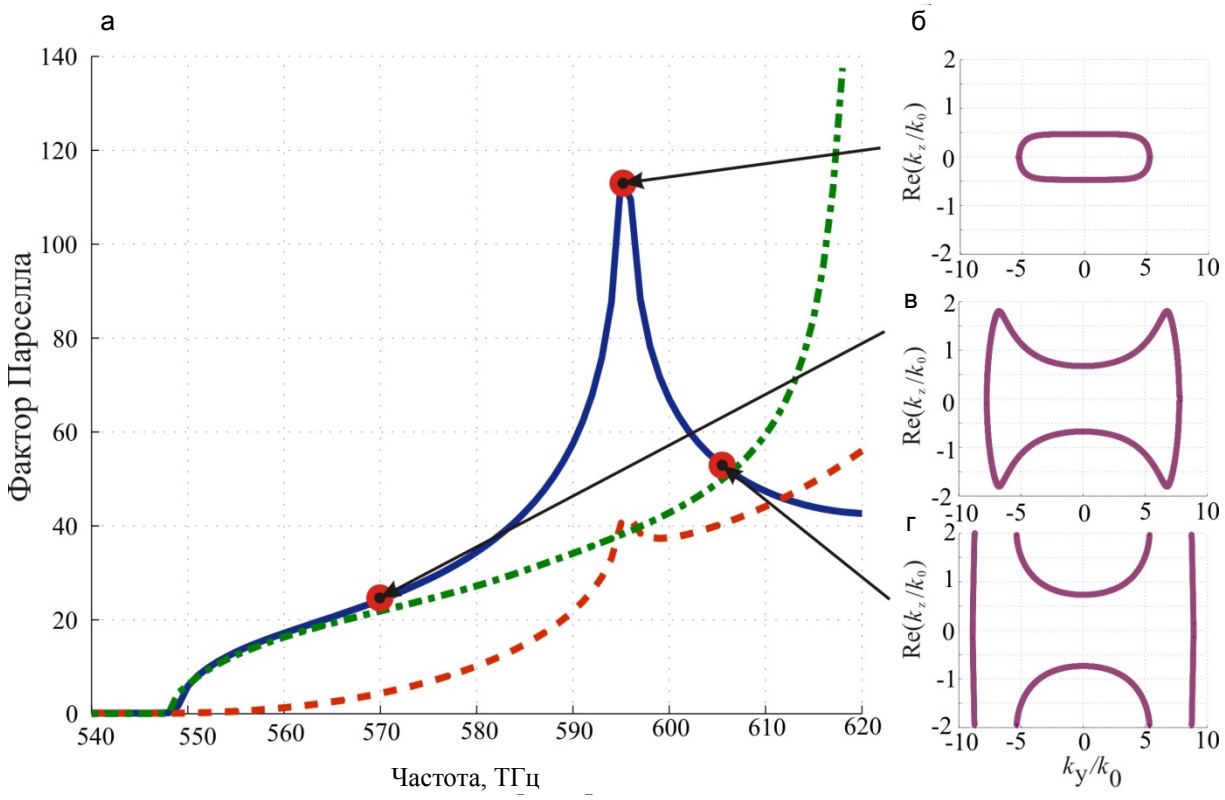


Рис. 3. Зависимость фактора Пёрселла от частоты. Синяя сплошная кривая соответствует случаю ориентации диполя вдоль слоев, красный пунктир – случаю поперечной ориентации. Зеленый штрих-пунктир – результат эффективной модели (7), скорректированный коэффициентом 1,4 (а). Изочастотные контуры структуры на частотах 570, 595, 605 ТГц соответственно (б)–(г).

В диапазоне 570–595 ТГц изочастотный контур превращается из эллипса в «гантелю», а на частотах свыше 595 ТГц распадается на две гиперболические ветви. В данном режиме скорость спонтанного излучения квантового излучателя велика, но в нем доминируют волны с большими значениями вертикальной компоненты волнового вектора ($k_z > k_0$), которые не могут покинуть структуру и распространяться в дальнем поле.

Заключение

В работе продемонстрирован сильный (порядка 20) эффект Пёрселла в физически реализуемых предельно анизотропных эллиптических метаматериалах. Такие метаматериалы могут быть реализованы в виде слоистых металло-диэлектрических с элементарной ячейкой, состоящей из двух разных металлических слоев (например, золото и серебро) и двух одинаковых диэлектриков. Результаты работы могут быть полезны для получения абсолютно нового типа метаматериалов, которые заслуживают дальнейших исследований, а также для усиления спонтанной эмиссии квантовых излучателей за счет эффекта Пёрселла.

Литература

1. Jacob Z., Shalaev V.M. Plasmonics goes quantum // Science. 2011. V. 334. N 6055. P. 463–464.
2. Cortes C.L., Newman W., Molesky S., Jacob Z. Quantum nanophotonics using hyperbolic metamaterials // Journal of Optics. 2012. V. 14. N 6. Art. 063001.
3. Drachev V.P., Podolskiy V.A., Kildishev A.V. Hyperbolic metamaterials: new physics behind a classical problem // Optics Express. 2013. V. 21. N 12. P. 15048–15064.
4. Poddubny A., Iorsh I., Belov P., Kivshar Y. Hyperbolic metamaterials // Nature Photonics. 2013. V. 7. N 12. P. 958–967.
5. Felsen L., Marcuvitz N. Radiation and Scattering of Waves. NY: Wiley, 2003. 464 p.
6. Jacob Z., Smolyaninov I.I., Narimanov E.E. Broadband Purcell effect: radiative decay engineering with metamaterials // Applied Physics Letters. 2009. V. 100. N 18. Art. 181105.
7. Kavokin A., Baumberg J.J., Malpuech G., Laussy F.P. Microcavities. Oxford University Press, 2007. 430 p.

8. Tumkur T., Zhu G., Black P., Barnakov Y.A., Bonner C.E., Noginov M.A. Control of spontaneous emission in a volume of functionalized hyperbolic metamaterial // Applied Physics Letters. 2011. V. 99. N 15. Art. 151115.
9. Kim J., Drachev V.P., Jacob Z., Naik G.V., Boltasseva A., Narimanov E.E., Shalaev V.M. Improving the radiative decay rate for dye molecules with hyperbolic metamaterials // Optics Express. 2012. V. 20. N 7. P. 8100–8116.
10. Lu D., Kan J.J., Fullerton E.E., Liu Z. Enhancing spontaneous emission rates of molecules using nanopatterned multilayer hyperbolic metamaterials // Nature Nanotechnology. 2014. V. 9. N 1. P. 48–53.
11. Orlov A.A., Zhukovsky S.V., Iorsh I.V., Belov P.A. Controlling light with plasmonic multilayers // Photonics and Nanostructures – Fundamentals and Applications. 2014. V. 12. N 3. P. 213–230.
12. Agranovich V.M., Kravtsov V.E. Notes on crystal optics of superlattices // Solid State Communications. 1985. V. 55. N 1. P. 85–90.
13. Orlov A.A., Voroshilov P.M., Belov P.A., Kivshar Y.S. Engineered optical nonlocality in nanostructured metamaterials // Physical Review B – Condensed Matter and Materials Physics. 2011. V. 84. N 4. Art. 045424.
14. Orlov A.A., Krylova A.K., Zhukovsky S.V., Babicheva V.E., Belov P.A. Multiperiodicity in plasmonic multilayers: general description and diversity of topologies // Physical Review A – Atomic, Molecular, and Optical Physics. 2014. V. 90. N 1. Art. 013812.
15. Tomas M.S., Lenac Z. Spontaneous-emission in an absorbing Fabry-Perot cavity // Physical Review A – Atomic, Molecular, and Optical Physics. 1999. V. 60. N 3. P. 2431–2437.
16. Poddubny A.N., Belov P.A., Ginzburg P., Zayats A.V., Kivshar Y.S. Microscopic model of Purcell enhancement in hyperbolic metamaterials // Physical Review B – Condensed Matter and Materials Physics. 2012. V. 86. N 3. Art. 035148.

Чебыкин Александр Васильевич – аспирант, инженер-исследователь, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, chebykin.alexandr@gmail.com

Орлов Алексей Анатольевич – аспирант, младший научный сотрудник, Санкт-Петербург, 197101, Российская Федерация, orlov.aleksei@phoi.ifmo.ru

Хайслер Фабиан – студент, Йенский Университет имени Фридриха Шиллера, Йена, 07737, Германия; стажер-исследователь, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, fabian.heisler@uni-jena.de

Барышникова Ксения Владимировна – аспирант, инженер-исследователь, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, strekkuku@gmail.com

Белов Павел Александрович – доктор физико-математических наук, главный научный сотрудник, заведующий лабораторией, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, belov@phoi.ifmo.ru

Alexander V. Chebykin – postgraduate, research engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, chebykin.alexandr@gmail.com

Alexei A. Orlov – postgraduate, junior scientific researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, orlov.aleksei@phoi.ifmo.ru

Fabian Heisler – student, Friedrich-Schiller-Universität Jena, Jena, 07737, Germany; trainee researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, fabian.heisler@uni-jena.de

Ksenia V. Baryshnikova – postgraduate, research engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, strekkuku@gmail.com

Pavel A. Belov – D.Sc., chief research fellow, head of laboratory, ITMO University, Saint Petersburg, 197101, Russian Federation, belov@phoi.ifmo.ru

Принято к печати 09.09.14
Accepted 09.09.14

УДК 535.42

ИССЛЕДОВАНИЕ ОБЪЕМА С ВЫСОКОЙ ПЛОТНОСТЬЮ ЧАСТИЦ НА ОСНОВЕ КОНТУРНОГО И КОРРЕЛЯЦИОННОГО АНАЛИЗА ИЗОБРАЖЕНИЙ

Т.Ю. Николаева[a], Н.В. Петров[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, paltanya@mail.ru

Аннотация. Предметом исследования являются техники определения статистики частиц, в частности, методы обработки изображений частиц, полученных при когерентной подсветке. Рассматривается задача распознавания и статистического учета индивидуальных изображений малых рассеивающих частиц в произвольном сечении объема в случае их высокой концентрации. Для автоматического распознавания изображений сфокусированных частиц использовался специальный алгоритм статистического анализа на основе оконтуривания и пороговой обработки. С использованием математического аппарата скалярной теории дифракции были смоделированы когерентные изображения частиц, сформированные оптической системой с высокой числовой апертурой. Проведена численная апробация предложенного метода для случаев различных концентраций и распределений частиц по объему. В результате получены распределения плотности и массовой доли частиц и определена эффективность метода при работе с изображениями частиц различной концентрации. При высоких концентрациях усиливается проявление эффекта когерентного наложения частиц из соседних плоскостей, что делает затруднительным распознавание изображений частиц с помощью рассмотренного в работе алгоритма. В этом случае мы предлагаем дополнить методику вычислением функции взаимной корреляции изображений частиц соседних сегментов объема и оценкой отношения высоты корреляционного пика к высоте пьедестала функции в случае различных характеров распределения. Рассмотренный в работе способ статистического учета частиц имеет важное практическое значение при исследовании объема с частицами различной природы, например, в задачах биологии и океанологии. Эффективная работа в режиме высоких концентраций расширяет пределы применимости рассматриваемых методов на практически важные случаи и позволяет оптимизировать время определения характера распределения и статистических характеристик частиц.

Ключевые слова: обработка изображений, численное моделирование, лазерная анемометрия по изображениям частиц.

Благодарности. Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01). Н.В. Петров благодарит за поддержку Министерство образования и науки Российской Федерации, проект № 2014/190 на выполнение государственных работ в сфере научной деятельности в рамках базовой части государственного задания.

VOLUME STUDY WITH HIGH DENSITY OF PARTICLES BASED ON CONTOUR AND CORRELATION IMAGE ANALYSIS

T.Yu. Nikolaeva[a], N.V. Petrov[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, paltanya@mail.ru

Abstract. The subject of study is the techniques of particle statistics evaluation, in particular, processing methods of particle images obtained by coherent illumination. This paper considers the problem of recognition and statistical accounting for individual images of small scattering particles in an arbitrary section of the volume in case of high concentrations. For automatic recognition of focused particles images, a special algorithm for statistical analysis based on contouring and thresholding was used. By means of the mathematical formalism of the scalar diffraction theory, coherent images of the particles formed by the optical system with high numerical aperture were simulated. Numerical testing of the method proposed for the cases of different concentrations and distributions of particles in the volume was performed. As a result, distributions of density and mass fraction of the particles were obtained, and the efficiency of the method in case of different concentrations of particles was evaluated. At high concentrations, the effect of coherent superposition of the particles from the adjacent planes strengthens, which makes it difficult to recognize images of particles using the algorithm considered in the paper. In this case, we propose to supplement the method with calculating the cross-correlation function of particle images from adjacent segments of the volume, and evaluating the ratio between the height of the correlation peak and the height of the function pedestal in the case of different distribution characters. The method of statistical accounting of particles considered in this paper is of practical importance in the study of volume with particles of different nature, for example, in problems of biology and oceanography. Effective work in the regime of high concentrations expands the limits of applicability of these methods for practically important cases and helps to optimize determination time of the distribution character and statistical characteristics of the particles.

Keywords: images processing, numerical simulation, particle image velocimetry.

Acknowledgements. The work has been partially financially supported by the Government of the Russian Federation (grant 074-U01). N.V. Petrov expresses thanks to the Russian Federation Ministry of Education and Science for federal projects support (project № 2014/190) in the sphere of scientific activity within a basic part of the government order.

Введение

Сложно переоценить важность методов цифровой обработки изображений частиц и их статистического учета. Область применения этих методов включает в себя фундаментальные научные исследования, такие как изучение динамики потоков частиц (газодинамики двухфазных потоков, сверхзвуковых потоков, гидродинамики течений, аэродинамики вертолетов) [1, 2], исследование частиц разной природы в задачах океанологии и биологии [3], исследование прозрачных сред [4]. Помимо этого, методы цифровой

обработки изображений также находят применение при решении многих практических задач, таких как впрыск топлива, сельскохозяйственные спреи, фармацевтика, распылительная сушка продуктов питания. Здесь особую важность приобретает знание статистических характеристик распределения частиц в объеме среды, таких как характер распределения, размеры и концентрация частиц.

Существует множество методов обработки изображений частиц, полученных при когерентной подсветке, среди которых можно выделить метод лазерной анемометрии по изображениям частиц – PIV (particle image velocimetry) [5] и методы цифровой голографии – DH (digital holography) [6]. PIV наиболее широко используется для визуализации различных потоков жидкости или газа. Однако в классических методах PIV можно получить только две компоненты скорости потока в двух пространственных координатах в плоскости измерения. В связи с этим в настоящее время большое применение в области обработки изображений частиц нашли методы цифровой голографии. Благодаря своему главному преимуществу, способности записывать и реконструировать 3D-изображения тестового объема в масштабе реального времени, эти методы используются при решении задач детектирования частиц и определения их размера и положения в объеме [7], а также для анализа движущихся частиц и измерения полей скоростей потоков (цифровые голографические методы лазерной анемометрии по изображениям частиц DH PIV) [8]. Однако возможности широкого практического применения DH PIV-методов для анализа как потоков частиц, так и частиц, взвешенных в объеме оптической среды, существенно ограничены вследствие того, что запись голограммы с опорной волной требует высокой стабильности опорного и предметного пучков. А это подразумевает необходимость работы в лабораторных условиях с виброизоляцией. В дополнение к этому, наличие опорной волны увеличивает количество оптических элементов в схеме, что усложняет работу с ней. Кроме того, в работах, использующих DH-методы, речь идет о малых концентрациях частиц [9, 10]. Особое внимание при этом уделяется именно качеству восстановления изображений частицы и точности определения ее 3D-координат в объеме. Однако с практической точки зрения не менее важным является исследование точности существующих методов в случае высоких плотностей частиц в объеме среды. К сожалению, опубликованных работ, посвященных этому вопросу, нам найти не удалось.

Целью настоящей работы является решение задачи распознавания и статистического учета индивидуальных изображений малых рассеивающих частиц в произвольном сечении объема в случае их высокой концентрации. Ввиду упомянутых выше особенностей DH-методов в качестве базового метода мы решили использовать одну из современных модификаций PIV-техник [11], которая заключается в использовании алгоритма статистического анализа частиц на основе оконтуривания и пороговой обработки. В результате численной апробации исследуемого метода с использованием специально разработанной имитационной модели были получены распределения плотности и массовой доли частиц для случаев различной плотности частиц и их распределений по объему, а также определена эффективность метода при работе с изображениями частиц различной концентрации. В случае, когда плотности частиц в объеме среды слишком высоки, чтобы эффективно распознавать их посредством предложенного метода, мы предлагаем дополнить методику вычислением функции взаимной корреляции изображений двух соседних сегментов объема для определения характера распределения частиц по нему.

Имитационная модель для исследования индивидуальных изображений частиц

Для апробации исследуемого метода и оценки его эффективности в качестве инструмента для статистического учета взвешенных в объеме рассеивающих частиц была разработана специальная имитационная модель в многофункциональной среде National Instruments LabVIEW.

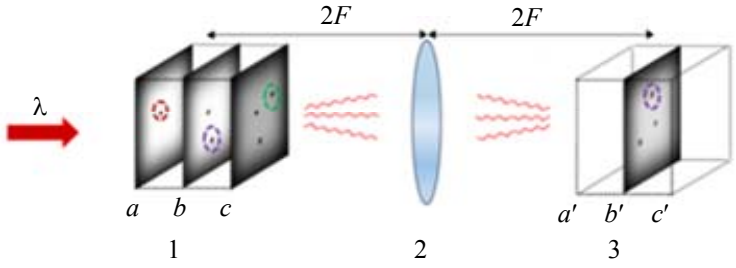


Рис. 1. Распределение интенсивности волнового поля, продифрагировавшего на трех частицах в объеме оптической среды, регистрируемое в плоскости *b′*: 1 – объем среды, разбитый на сегменты *a*, *b*, *c*; 2 – линза; 3 – плоскость изображения

На рис. 1 представлена схема оптической системы, формирующей изображение объема среды с частицами. Электромагнитное излучение, проходя через объем среды, разбитый предварительно на сегменты (*a*, *b*, *c* на рис. 1), дифрагирует на частицах, находящихся в этих сегментах. Сфокусированное изображение частиц, находящихся в определенном сегменте объема (сегмент *b*), переносится в плоскость регистрации с помощью линзы. Линза имеет высокую числовую апертуру, что позволяет различать изображения частиц, находящих в различных плоскостях объема. Полученное изображение частиц (*b′* на

рис. 1) является результатом когерентного наложения частиц, расположенных в различных сегментах объема. В численном эксперименте, меняя расстояния, на которое производился расчет волнового поля из плоскости сегмента объема в плоскость линзы, были получены сфокусированные изображения различных плоскостей из объема оптической среды. При этом частицы из соседних плоскостей принимали вид расфокусированных пятен.

Чтобы разрабатываемая модель объемной оптической среды максимально близко соответствовала условиям, наиболее часто имеющим место в задачах исследования частиц, рассматривались два распределения: случайное (равновероятное по сегментам) и нормальное, реализуемое в численной модели с помощью преобразования Бокса–Мюллера (случайные величины распределены по нормальному закону относительно оптической оси [12]). Для расчета распространения электромагнитного поля через объем оптической среды и оптическую систему использовался метод распространения углового спектра плоских волн, позволяющий рассчитать распространение волнового поля между двумя близко расположенными плоскостями (x_1, y_1) и (x_2, y_2), находящимися на расстоянии Δl [13]:

$$U(x_2, y_2, \Delta l) = \iint_{-\infty}^{+\infty} \hat{F}(f_x, f_y) \exp\left(ik\Delta l \sqrt{\left[1 - (\lambda f_x)^2 - (\lambda f_y)^2\right]} \right) \exp\left[i2\pi (f_x x + f_y y) \right] df_x df_y,$$

где $\hat{F}(f_x, f_y)$ – преобразование Фурье волнового поля во входной плоскости $U(x_1, y_1)$; f_x, f_y – пространственные частоты.

Принципы метода статистического учета частиц на основе автоматического оконтуривания и пороговой обработки

Обрабатываемые изображения представляют собой наборы сфокусированных и расфокусированных изображений частиц из всех сегментов объема, накладывающихся друг на друга. Для автоматического распознавания изображений сфокусированных частиц в заданном сегменте объема использовался специальный алгоритм статистического анализа на основе оконтуривания и пороговой обработки [1], подробно описанный в [14].

В изначальном варианте алгоритм был разработан для работы с протяженными изображениями треков частиц. В данной работе алгоритм был адаптирован для изображений, взвешенных в объеме частиц круглой формы: изменены параметры срабатывания механизма оконтуривания на основе априорной информации о строении частиц. Используемый алгоритм заключается в следующем: выделяются примерные области, внутри которых находятся изображения сфокусированных частиц; затем программа, сканируя выделенные фрагменты изображения вдоль и поперек, находит в них двойные перепады яркости от светлого к темному и обратно и тем самым определяет координаты частиц; далее, считывая результаты, полученные на предыдущих этапах, строятся радиальные распределения, распределения плотности и массовой доли исследуемых частиц в плоскости изображения.

Корреляционная обработка изображений частиц в случае высоких концентраций

В случае, когда концентрация частиц высока, их изображения накладываются таким образом, что использование алгоритма, основанного на оконтуривании и пороговой обработке, становится недостаточно эффективным. Здесь применение корреляционного анализа изображений частиц может служить дополнительным инструментом при исследовании частиц, распределенных по объему. В отличие от обычного корреляционного анализа, применяемого для оценки перемещений по координатам корреляционного пика [11], и метода, представленного в [15], где оценивается автокорреляционная функция распределения интенсивности поля в плоскости наблюдения (радиус корреляции), мы предлагаем использовать корреляционный анализ изображений частиц соседних сегментов объема. Функция взаимной корреляции для двух изображений соседних сегментов рассчитывалась с использованием серии двумерных преобразований Фурье [16] вида

$$c(\Delta x, \Delta y, \mathrm{t}) = \hat{F}^{-1}\left[H^*(\xi, \eta, 0) H(\xi, \eta, \mathrm{t}) \right],$$

где $\Delta x, \Delta y$ – относительные координаты в плоскости изображения; ξ, η – пространственно-частотные компоненты; $\hat{F}^{-1}$ – обратное преобразование Фурье; $H(\xi, \eta, 0)$ и $H(\xi, \eta, \mathrm{t})$ – двумерные преобразования Фурье от отклонений от средних распределений интенсивности фрагментов изображений, * – комплексное сопряжение.

По построенным сечениям функции взаимной корреляции изображений двух соседних сегментов объема по одной из координат изображения рассматривалось соотношение высоты корреляционного пика к высоте пьедестала функции, что позволило определять характер распределения частиц в случае их высокой концентрации.

Статистическое исследование когерентных изображений частиц, взвешенных в объеме оптической среды

Для получения качественных и количественных характеристик распределений частиц, взвешенных в объеме оптической среды, была проведена статистическая обработка сфокусированных изображений частиц. Для различных значений концентраций частиц в объеме рассматривалась тестовая выборка из 10 изображений, полученных при одинаковых параметрах. На рис. 2 представлены результаты моделирования сфокусированных изображений частиц на один поперечный сегмент объема для случайного и нормального распределений частиц по объему, а также обозначены координаты распознанных с помощью описанного выше алгоритма изображений частиц для всех 10 изображений из тестовой выборки. Исследуемый объем среды 3×3×3 мм в ходе численного эксперимента делился на 20 сегментов, шаг сетки Δx = 6 мкм, λ = 632,8 нм, фокусное расстояние формирующей изображение линзы f = 29 мм, диаметр частицы d = 0,05 мм. Можно заметить, что в обоих случаях попавшие на границы или вблизи границ объема частицы не распознаются, что обусловлено использованием квадратной аподизирующей диафрагмы при моделировании.

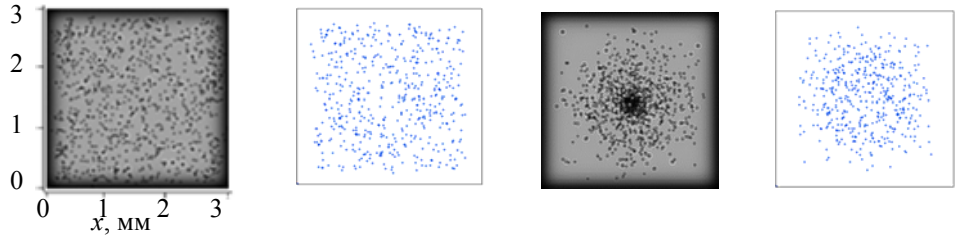


Рис. 2. Сфокусированные изображения частиц (концентрация частиц в одном сегменте (плоскости) объема K = 50): случайное распределение (а), нормальное распределение (в); результаты распознавания частиц: случайное распределение (б), нормальное распределение (г)

Были построены зависимости нормированных распределений плотности частиц в плоскости изображения и нормированные распределения массовой доли частиц (рис. 3), где n – концентрация частиц в единице объема; K/K_{max} – нормированные на максимальное значение плотности частиц; Kn/Kn_{max} – нормированные на максимальное значение массовые доли частиц. Из рисунка видно, что в случае равновероятного распределения частиц по объему полученные в результате численного моделирования распределения плотности хорошо соотносятся с пуассоновской статистикой. Однако при больших концентрациях частиц на сегмент объема распознаются также некоторые частицы, не принадлежащие исследуемому сегменту, и, как следствие, наблюдаются незначительные флуктуации массовой доли частиц (рис. 3, е). Когда частицы распределены по объему по нормальному закону, имеет место меньшее соответствие распределений частиц пуассоновской статистике; так как большинство частиц локализуется ближе к центру оптической оси, выделить частицы, находящиеся в исследуемом сегменте, становится затруднительным. На рис. 3, ж, з, четко видны флуктуации массовой доли частиц. Это говорит о том, что в случае нормального распределения эффект когерентного наложения частиц из соседних плоскостей проявляется уже при малых концентрациях частиц, что делает затруднительным распознавание изображений частиц с помощью рассмотренного в работе алгоритма при больших концентрациях.

Для оценки эффективности работы алгоритма была построена зависимость количества распознанных (пойманных) частиц от изначально заданного количества на один сегмент объема (рис. 4). Видно, что в случае равновероятного распределения частиц по объему при концентрации частиц больше 100 на один сегмент объема происходит довольно резкое ухудшение эффективности работы программы ввиду невозможности распознать более 10% от заданного количества частиц. В случае стандартного нормального распределения невозможность распознавания более 10% частиц возникает уже при меньших концентрациях.

Для случая больших концентраций мы предлагаем дополнить метод расчетом взаимной корреляционной функции для двух изображений соседних сегментов объема. В одном изображении мы имеем дело со сфокусированными частицами из одного сегмента, в другом изображении эти частицы принимают вид расфокусированных пятен, и в фокусе находятся частицы из соседнего сегмента. На рис. 5 представлены результаты расчета взаимной корреляционной функции для двух изображений, а также сечения этой функции по одной из координат изображения. Корреляционная функция вычислялась от изображения меньшего размера, что было обусловлено наличием аподизирующей диафрагмы, а также локализацией частиц преимущественно в непосредственной близости от оптической оси в случае нормального распределения.

Графики сечений функции взаимной корреляции двух изображений показывают, что в случае равновероятного распределения с увеличением концентрации частиц увеличивается корреляционные связи между элементами изображений соседних сегментов объема, и изображения частиц все меньше поддаются распознаванию методами компьютерной обработки. Это заметно по уменьшению пьедестала пика (a – на рис. 5, г) и увеличению пика корреляционной функции (b – на рис. 5, г). Однако в случае нормаль-

ного распределения наблюдается противоположная ситуация: увеличение концентрации частиц на один сегмент объема ведет к уменьшению корреляционных связей между элементами изображений. Изображения соседних сегментов объема становятся хорошо различимыми, но при этом, как обсуждалось выше, при таких больших концентрациях происходит значительная потеря информации о частицах.

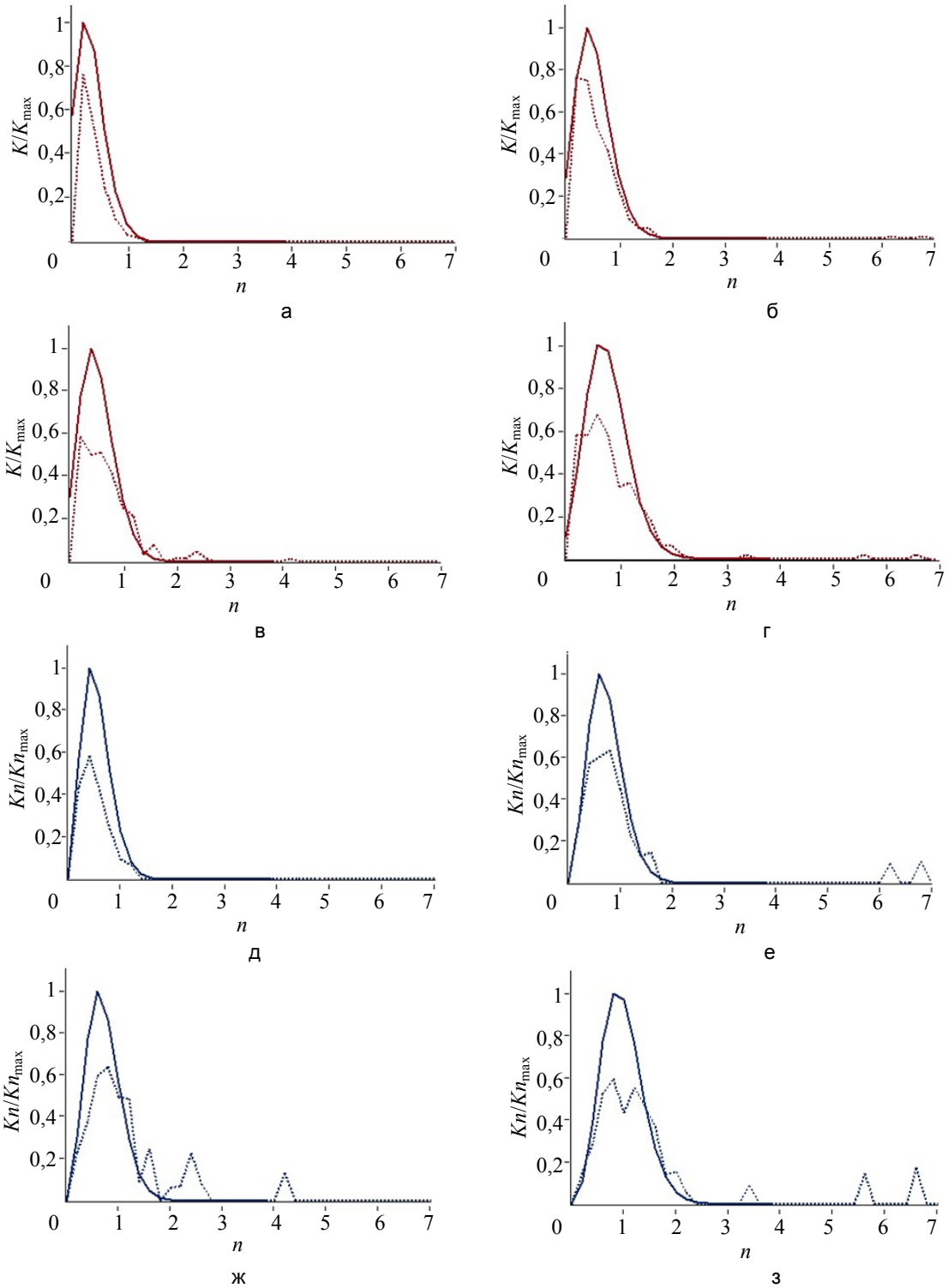


Рис. 3. Рассчитанные согласно пуассоновской статистике (сплошная кривая) и полученные в результате численного моделирования (пунктирная кривая) нормированные распределения плотности частиц в плоскости изображения (а)–(г): первая строка – случайное распределение (K = 50 и K = 150 соответственно), вторая строка – нормальное распределение (K = 50 и K = 150); рассчитанные согласно пуассоновской статистике (сплошная кривая) и полученные в результате численного моделирования (пунктирная кривая) нормированные распределения массовой доли частиц (д)–(з): третья строка – случайное распределение (K = 50 и K = 150), четвертая строка – нормальное распределение (K = 50 и K = 150)

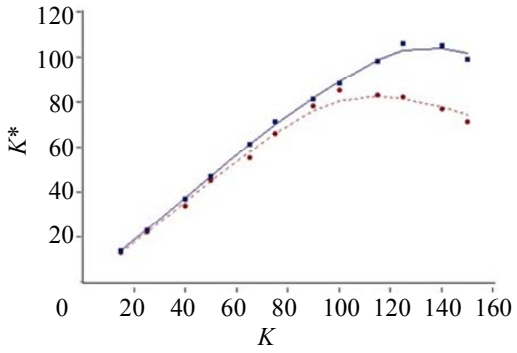


Рис. 4. Зависимость количества распознанных частиц K^* от изначально заданного количества частиц в одном сегменте объема K: случайное распределение (синяя кривая) и нормальное распределение (красная кривая)

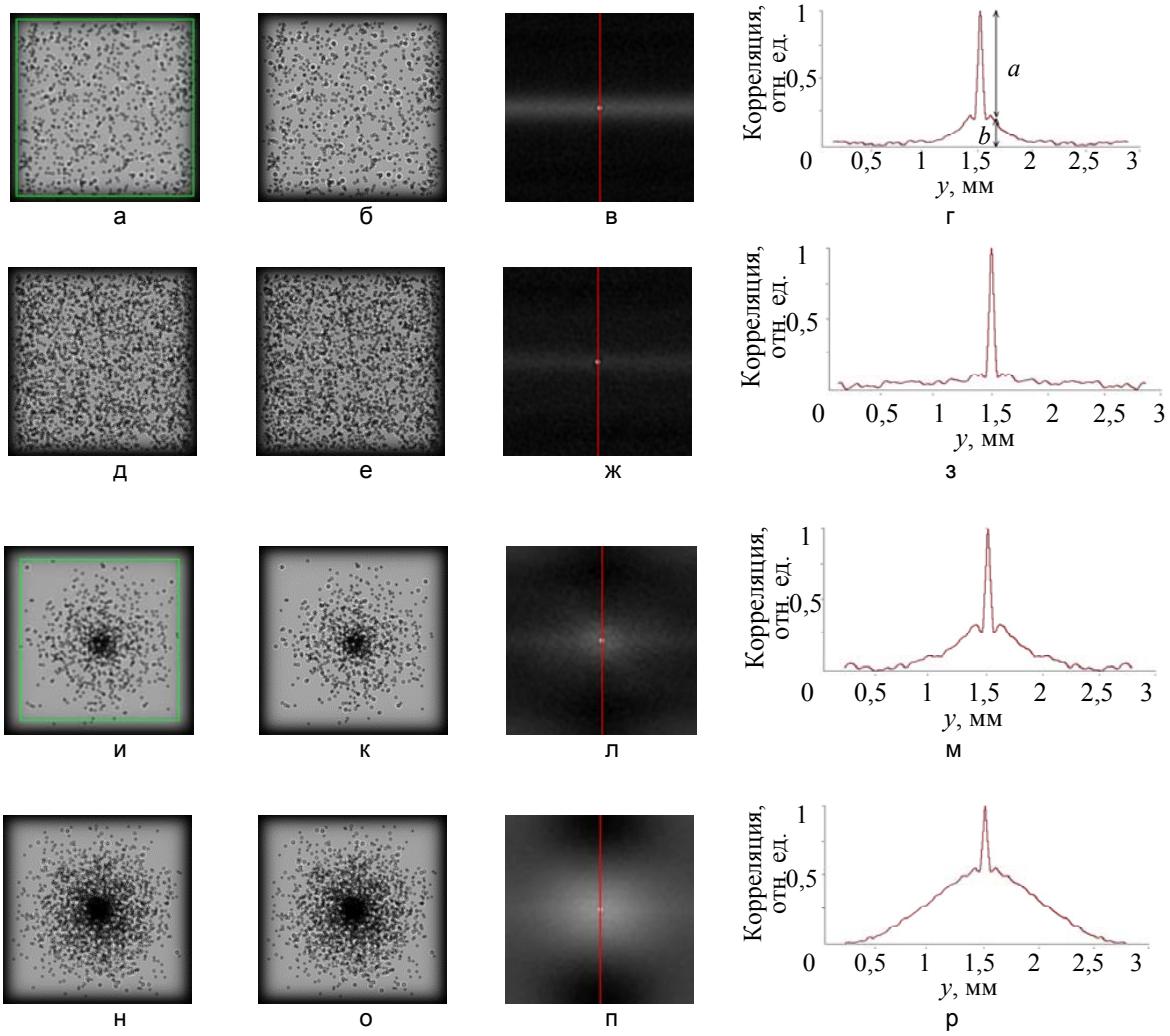


Рис. 5. Результаты корреляционного анализа изображений для случаев распределения частиц: по случайному закону (а)–(з), по нормальному закону (и)–(р): использованные изображения частиц (первые два столбца), нормированная функция взаимной корреляции (третий столбец) и сечение этой функции (последний столбец) для различных концентраций: K = 50 (а)–(г), (и)–(м); K = 150 (д)–(з), (н)–(р)

Заключение

В работе проведено исследование и предложено решение задачи распознавания и статистического учета индивидуальных изображений малых рассеивающих частиц в произвольном сечении объема в случае их высокой концентрации. Разработана специальная имитационная модель, позволяющая численно моделировать объем среды с распределенными по нему частицами, обладающими такими различными характеристиками, как размер, концентрация и характер распределения по объему. Для исследования ста-

тистики взвешенных частиц круглой формы был адаптирован алгоритм статистического анализа на основе оконтуривания и пороговой обработки. В результате численной апробации исследуемого метода с использованием специально разработанной имитационной модели были получены распределения плотности и массовой доли частиц для случаев различных концентраций и законов распределений частиц по объему. В работе были установлены предельные концентрации частиц, при которых удается распознать изображения отдельных частиц с помощью используемого алгоритма. В дополнение к совокупности таких техник, как автоматическое оконтуривание и пороговая обработка, предложено использование корреляционного анализа изображений частиц из двух соседних плоскостей объема. Оценка отношения высоты корреляционного пика к высоте пьедестала функции взаимной корреляции позволила определить характер распределения частиц даже при высоких концентрациях, когда стандартный метод оконтуривания дает ошибочные результаты. Предложенный подход к задаче статистического учета взвешенных в объеме частиц обладает рядом преимуществ перед распространенными DH-методами, в частности, простотой его практической реализации.

Литература

1. Воронецкий А.В., Михайлов В.Н., Петров Н.В., Стаселько Д.И. Измерение пространственно-временных параметров движения самосветящихся частиц в сверхзвуковом высокотемпературном потоке // Оптический журнал. 2012. Т. 79. № 1. С. 18–24.
2. Pereira F., Gharib M. Defocusing digital particle image velocimetry and the three-dimensional characterization of two-phase flows // Measurement Science and Technology. 2002. V. 13. N 5. P. 683–694.
3. Dyomin V.V., Olshukov A.S. Digital holographic video for studying biological particles // Journal of Optical Technology. 2012. V. 79. N 6. P. 344–347.
4. Johansson E.-L., Benckert L., Sjodahl M. Phase object data obtained from defocused laser speckle displacement // Applied Optics. 2004. V. 43. N 16. P. 3229–3234.
5. Peterson K., Regaard B., Heinemann S., Sick V. Single-camera, three-dimensional particle tracking velocimetry // Optics Express. 2012. V. 20. N 8. P. 9031–9037.
6. Pitkäaho T., Niemelä M., Pitkäkangas V. Partially coherent digital in-line holographic microscopy in characterization of a microscopic target // Applied Optics. 2014. V. 53. N 15. P. 3233–3240.
7. Malek M., Allano D., Coëtmellec S., Lebrun D. Digital in-line holography: influence of the shadow density on particle field extraction // Optics Express. 2004. V. 12. N 10. P. 2270–2279.
8. Zhang Y., Shen G., Schroder A., Kompenhans J. Influence of some recording parameters on digital holographic particle image velocimetry // Optical Engineering. 2006. V. 45. N 7. Art. 075801.
9. Yang W., Kostinski A.B., Shaw R.A. Depth-of-focus reduction for digital in-line holography of particle fields // Optics Letters. 2005. V. 30. N 11. P. 1303–1305.
10. Singh D.H., Panigrahi P.K. Improved digital holographic reconstruction algorithm for depth error reduction and elimination of out-of-focus particles // Optics Express. 2010. V. 18. N 3. P. 2426–2448.
11. Petrov N.V., Bespalov V.G., Zhevlakov A.P., Soldatov Yu.I. Determining the velocity of an object in water, using digital speckle-photography // Journal of Optical Technology. 2007. V. 74. N 11. P. 779–782.
12. Box G.E.P., Muller M.E. A note on the generation of random normal deviates // Ann. Math. Stat. 1958. V. 29. N 2. P. 610–611.
13. Goodman J.W. Introduction to Fourier Optics. NY: McGraw-Hill, 1961. 441 p.
14. Воронецкий А.В., Михайлов В.Н., Петров Н.В., Стаселько Д.И. Экспериментальное исследование пространственно-скоростных параметров частиц в сверхзвуковом двухфазном потоке // Труды НИЦ фотоники и оптоинформатики. СПб.: СПбГУ ИТМО, 2009. С. 347–359.
15. Павлов П.В., Петров Н.В., Малов А.Н. Определение параметров шероховатости и дефектация поверхностей деталей воздушного судна с применением спиральных пучков лазерного излучения // Научно-технический вестник СПбГУ ИТМО. 2011. № 6 (76). С. 84–88.
16. Synnergren P., Larsson L., Lundström S. Digital speckle photography: visualization of mesoflow through clustered fiber networks // Applied Optics. 2002. V. 41. N 7. P. 1368–1373.

Николаева Татьяна Юрьевна – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, paltanya@mail.ru

Петров Николай Владимирович – кандидат физико-математических наук, доцент, соискатель, Университет ИТМО, ИТМО, Санкт-Петербург, 197101, Российская Федерация, Nickolai.petrov@gmail.com

Tatyana Yu. Nikolaeva – student, ITMO University, Saint Petersburg, 197101, Russian Federation, paltanya@mail.ru

Nikolai V. Petrov – PhD, ITMO University, Saint Petersburg, 197101, Russian Federation, n.petrov@niuitmo.ru

Принято к печати 10.09.14
Accepted 10.09.14

УДК 539.26+535.34+535.37+620.179.152.1

САМООРГАНИЗАЦИЯ КВАНТОВЫХ ТОЧЕК СУЛЬФИДА СВИНЦА В СУПЕРСТРУКТУРЫ

Е.В. Ушакова[a], В.В. Голубков[b], Е.О. Осколков[a], А.П. Литвин[a], П.С. Парфенов[a], А.В. Баранов[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, el.ushakova@gmail.com

[b] Институт химии силикатов имени И.В. Гребенщикова Российской Академии наук, Санкт-Петербург, 199034, Российская Федерация

Аннотация. Методом рентгеновского структурного анализа (рассеяние рентгеновского излучения под малыми углами) показано, что структуры, полученные в результате самоорганизации на подложке квантовых точек сульфида свинца, представляют собой упорядоченные массивы. Самоорганизация квантовых точек происходит при медленном испарении растворителя из кюветы. Кювета представляет собой тонкий слой слюды с приклеенным на нее тефлоновым кольцом. По положению пиков рассеяния на дифрактограмме была рассчитана кристаллическая структура полученных упорядоченных структур. Такие структуры обладают ромбической сингонией с примитивной кристаллической решеткой. Вычисленные параметры решетки: a = 21,1 нм; b = 36,2 нм; c = 62,5 нм. Размеры структур составили десятки микрометров. Исследованы спектральные свойства полученных суперструктур из квантовых точек сульфида свинца и кинетические параметры их люминесценции. Полоса поглощения суперструктур уширена по сравнению с полосой поглощения квантовых точек в растворе, полоса люминесценции немного смещена в красную область спектра, при этом ширина полосы практически не изменилась. Время затухания люминесценции полученных структур значительно уменьшилось по сравнению с изолированными квантовыми точками в растворе, но совпадает для плотно упакованных ансамблей квантовых точек сульфида свинца. Такие суперструктуры могут быть использованы для создания элементов солнечных батарей с улучшенными параметрами.

Ключевые слова: квантовая точка, сульфид свинца, самоорганизация, суперкристалл, рентгеновский структурный анализ.

SELF-ORGANIZATION OF LEAD SULFIDE QUANTUM DOTS INTO SUPERSTRUCTURES

E.V. Ushakova[a], V.V. Golubkov[b], E.O. Oskolkov[a], A.P. Litvin[a], P.S. Parfenov[a], A.V. Baranov[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, el.ushakova@gmail.com

[b] Institute of Silicate Chemistry of the Russian Academy of Sciences, Saint Petersburg, 199034, Russian Federation

Abstract. The method of X-ray structural analysis (X-ray scattering at small angles) is used to show that the structures obtained by self-organization on a substrate of lead sulfide (PbS) quantum dots are ordered arrays. Self-organization of quantum dots occurs at slow evaporation of solvent from a cuvette. The cuvette is a thin layer of mica with teflon ring on it. The positions of peaks in SAXS pattern are used to calculate crystal lattice of obtained ordered structures. Such structures have a primitive orthorhombic crystal lattice. Calculated lattice parameters are: a = 21,1 (nm); b = 36,2 (nm); c = 62,5 (nm). Dimensions of structures are tens of micrometers. The spectral properties of PbS QDs superstructures and kinetic parameters of their luminescence are investigated. Absorption band of superstructures is broadened as compared to the absorption band of the quantum dots in solution; the luminescence band is slightly shifted to the red region of the spectrum, while its bandwidth is not changed much. Luminescence lifetime of obtained structures has been significantly decreased in comparison with the isolated quantum dots in solution, but remained the same for the lead sulfide quantum dots close-packed ensembles. Such superstructures can be used to produce solar cells with improved characteristics.

Keywords: quantum dot, lead sulfide, self-organization, super-crystal, X-ray structural analysis.

Введение

Коллоидные полупроводниковые нанокристаллы или квантовые точки (КТ) – это монодисперсные кластеры с абсолютными размерами меньше радиуса экситона Бора [1]. Поскольку энергетический зазор между верхним уровнем дырок и нижайшим уровнем электронов определяется размером КТ, открывается возможность управлять значением энергии фундаментального перехода за счет синтеза нанокристаллов различного диаметра. Благодаря своим уникальным оптоэлектронным свойствам и продолжительной фотостабильности, КТ и структуры на их основе привлекают интерес ввиду широких перспектив применений в различных областях, в том числе и в современном материаловедении [2, 3]. Среди множества полупроводниковых материалов одним из многообещающих для применения в устройствах инфракрасного (ИК) диапазона является класс коллоидных КТ халькогенидов свинца, таких как сульфид свинца (PbS), квантовые переходы которых лежат в ближней ИК области спектра (0,8–3,0 мкм). Они обладают высоким коэффициентом экстинкции в ближнем ИК диапазоне спектра, малой эффективной массой носителей заряда [4], возможностью осуществлять мультиэкситонную генерацию, большими значениями времен затухания люминесценции [5] и др. Исходя из этого, материалы, основанные на упорядоченных структурах из PbS КТ, являются перспективными для создания более эффективных преобразователей солнечной энергии.

Одним из способов получения упорядоченных структур является самоорганизация нанокристаллов на подложке. Самоорганизация – это самопроизвольное упорядочение частиц и расположение их в некую структуру. Многие исследования за последнее время посвящены таким структурам из полупроводниковых и металлических наночастиц, самоорганизованных в сверхрешетки [2]. Более того, некоторые КТ могут собираться в крупные кристаллы, так называемые суперкристаллы [6, 7], структура кото-

рых может быть определена с помощью рентгеновского структурного анализа. Несмотря на интенсивные исследования, материалы, основанные на самоорганизованных структурах, еще далеки от применения. Было установлено, что такие структуры крайне чувствительны к ряду параметров, например, к размеру КТ, типу растворителя и лигандов, находящихся на их поверхности [6]. Именно поэтому изучение влияния таких параметров процесса самоорганизации структур из нанокристаллов является актуальным. В связи с этим целью настоящей работы стало исследование процесса формирования объемных структур на основе PbS КТ и дальнейшее изучение оптических свойств полученных структур.

Экспериментальные методы исследования

Для определения пространственной структуры полученных структур на подложке, а также размера образующих ее КТ использовался метод рассеяния рентгеновских лучей под малыми углами (РМУ) [8]. Спектры поглощения КТ в растворе и в объемной структуре были исследованы при помощи спектрофотометра Shimadzu UV3600. Для исследования спектральных и кинетических параметров люминесценции нами была использована оригинальная установка, аналогичная описанной в работах [9–11].

Для обработки полученных экспериментальных данных использовались программы Origin 8.0. Индицирование дифрактограмм РМУ выполнялось с использованием программного комплекса PDWin 3.0, разработанного научно-производственным предприятием «Буревестник».

Приготовление образцов

Для изготовления наночастиц был применен высокотемпературный органометаллический синтез в органическом растворе – метод горячей инжекции [12]. Общим подходом в таком методе синтеза наночастиц халькогенидов свинца является использование реакции между солями свинца (с олеиновой и лауриновой кислотами) и металлоорганическими реагентами серы (комплексы серы с октадеценом, олеиламином или триоктилфосфином). В качестве реакционной среды использовались смеси октадецена, олеиламина и триоктилфосфиноксида. Реакция проводилась при температуре 70–170 ºС в атмосфере аргона с использованием стандартного химического оборудования для работы в инертной атмосфере, включая вакуумную сушку. В качестве растворителя для дальнейших исследований использовался четыреххлористый углерод. Таким образом, были получены PbS КТ, спектры поглощения и люминесценции которых приведены на рис. 1. Концентрация КТ в растворе составляла $4{,}76 \cdot 10^{-6}$ М/л. Пик поглощения находится на 1175 нм, пик люминесценции – на 1270 нм.

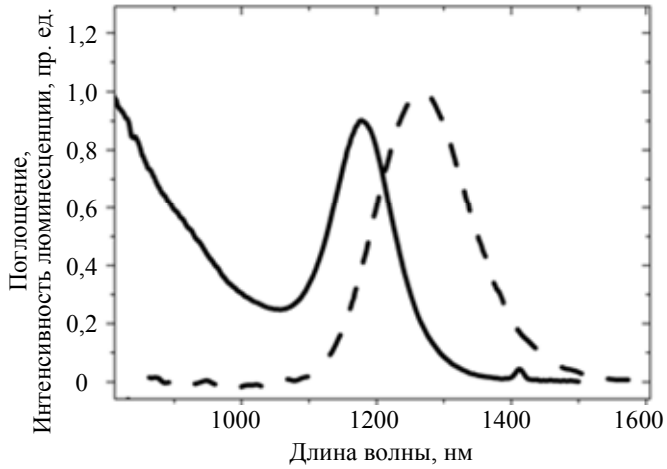


Рис. 1. Спектры поглощения (сплошная линия) и люминесценции (штриховая линия) PbS КТ в растворе четыреххлористого углерода

Кювета для РМУ измерений представляет собой тонкий слой слюды с приклеенным на нее тефлоновым кольцом. Объемные структуры изготавливались методом, основанным на испарении насыщенного раствора открытым способом. В нашем случае медленное испарение растворителя из стокового раствора КТ приводит к самоорганизации КТ на подложке тонкого слоя слюды. Объем стокового раствора в кювете составил 300 мкл.

Результаты и обсуждение

Были получены дифрактограммы от приготовленного образца через различные промежутки времени. Измерения проводились с интервалом 15 мин, чтобы посмотреть кинетику высыхания раствора и образования суперкристалла. На рис. 2 приведены полученные типичные угловые зависимости РМУ. В легенде на рис. 2 указаны время испарения раствора и интенсивность первичного пучка рентгеновского рассеяния.

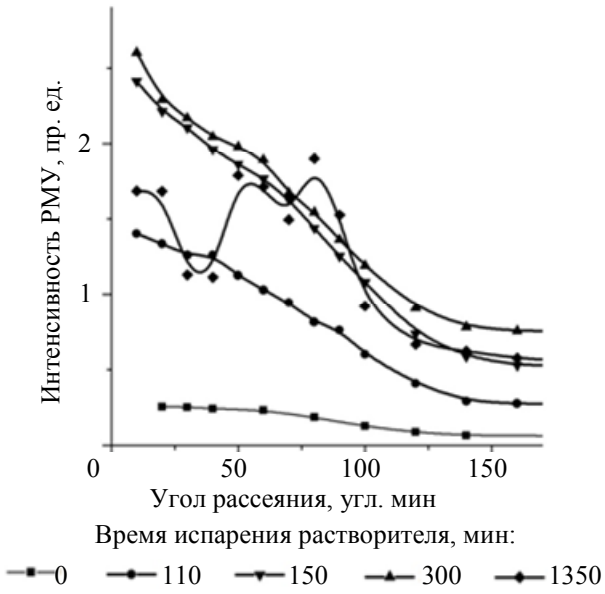


Рис. 2. Интенсивность рассеяния РМУ. В легенде указано время испарения четыреххлористого углерода

Из рис. 2 видно, что в самом начале эксперимента и вплоть до 110 мин испарения растворителя угловые зависимости сигнала РМУ представляют собой классические угловые зависимости интенсивности рентгеновского рассеяния изолированными частицами, находящимися в растворе. Такие зависимости описываются формулой Гинье [11]:

$$\ln\left(\frac{I}{I_0}\right) = \ln(Nn^2I_e) - \frac{4\pi^2}{3\lambda^2}\cdot R_0^2\cdot\varepsilon^2\,, \qquad (1)$$

где I – измеряемая интенсивность рассеяния рентгеновского излучения; I_0 – интенсивность первичного пучка; N – число частиц в системе; n – число электронов в частице; $\varepsilon = 2\theta$; $I_e = I_0\cdot\frac{e^4}{m^2c^2}\cdot\frac{1}{r^2}$ – интенсивность по Томсону. Отрезок, отсекаемый по оси ординат графика зависимости, позволяет определить число частиц в системе N, а угловой коэффициент наклона – электронный радиус инерции R_0. Анализ углового распределения интенсивности рассеяния позволяет определить размеры и форму неоднородностей, а также среднее расстояние между ними. Размер КТ, вычисленный по формуле (1), составил 4 нм.

При увеличении времени испарения четыреххлористого углерода из раствора, начиная со 150 мин, происходит увеличение интенсивности рассеяния около 10 угл. мин и наблюдается появление широкого пика, находящегося примерно на 60 угл. мин. Спустя 20 ч после приготовления образца в угловых зависимостях интенсивности РМУ наблюдается два широких пика, находящихся на 54 и 80 угл. мин. Такие пики свидетельствуют о плотной аморфной упаковке КТ [13]. Спустя почти 30 ч после приготовления образца была получена угловая зависимость интенсивности РМУ, схожая с дифрактограммами обычных кристаллов. Она приведена на рис. 3.

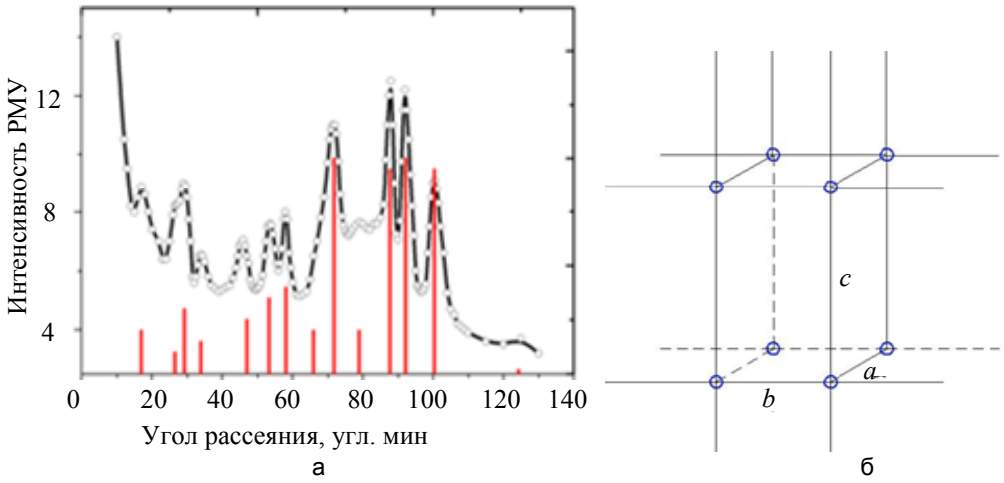


Рис. 3. Угловая зависимость интенсивности РМУ образца КТ PbS с размером 4 нм после высыхания растворителя (более 24 ч) (а); ячейка кристаллической решетки полученной суперструктуры (б)

Полученные данные были обработаны с помощью программы индицирования пиков рассеяния. Вычисленные положения пиков и их интенсивности показаны на рис. 3 красными линиями. Видно, что они полностью совпадают с полученными экспериментально данными. В результате индицирования экспериментальной дифрактограммы оказалось, что полученный суперкристалл имеет ромбическую сингонию с примитивной решеткой. Вычисленные параметры решетки составляют: $a = 21{,}1$ нм; $b = 36{,}2$ нм; $c = 62{,}5$ нм. Ячейка кристаллической решетки с полученными параметрами показана на рис. 3, б, в узлах решетки находятся КТ. На рис. 4 приведена фотография выращенных структур, полученная с помощью конфокального микроскопа.

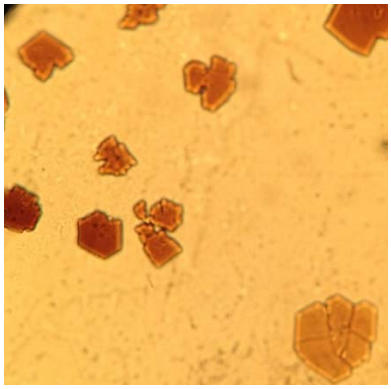


Рис. 4. Микрофотография суперкристаллов из PbS КТ, с использованием объектива с 100× увеличением. Размер изображения 100×100 мкм

На микрофотографии отчетливо видны разветвленные структуры, находящиеся в полимерной пленке. Мы предполагаем, что существует некий центр зарождения суперкристалла, из которого растут ветви, причем грани этих ветвей имеют довольно четкие границы.

На рис. 5 приведены спектры поглощения и люминесценции PbS КТ в растворе и организованные в суперкристалл.

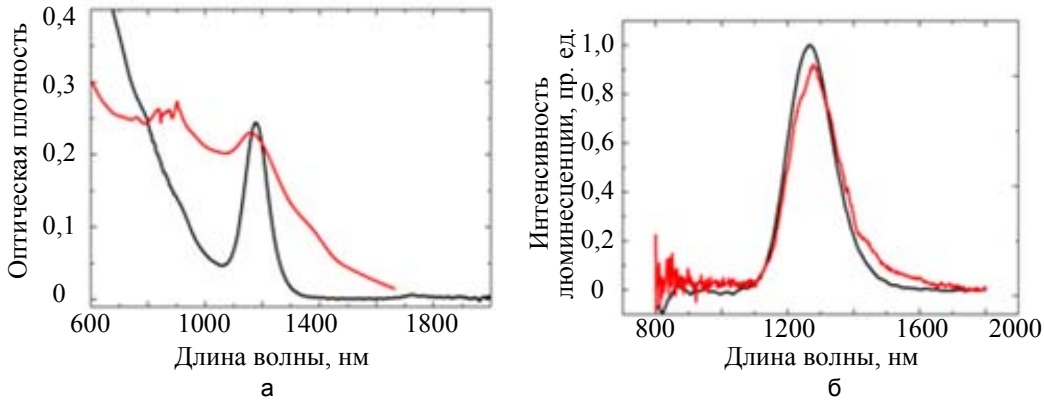


Рис. 5. Спектры поглощения (а) и люминесценции (б) PbS КТ с размером 4 нм в растворе четыреххлористого углерода (черная кривая) и организованные в суперкристалл (красная кривая)

Полоса поглощения суперкристалла находится на 1160 нм, также наблюдается плечо примерно на 1355 нм. Из сравнения спектров поглощения суперкристалла и коллоидного раствора КТ видно, что полоса поглощения суперкристалла уширена и сдвинута немного в коротковолновую область. Пик полосы люминесценции расположен на 1285 нм, ширина на половине высоты (FWHM) составляет 148 нм. Полоса люминесценции суперкристалла из КТ немного сдвинута в красную область.

Нами также была исследована кинетика затухания люминесценции суперкристалла. Кривая затухания люминесценции приведена на рис. 6. Затухание люминесценции КТ хорошо описывается биэкспоненциальной зависимостью:

$$I(t) = A_1 \cdot e^{-t/\tau_1} + A_2 \cdot e^{-t/\tau_2}.$$

Для анализа кинетики затухания люминесценции мы использовали среднее время релаксации люминесценции:

$$< \tau \geq {}^{\Sigma A_i \cdot \tau_i^2}\Big/_{\Sigma A_i \cdot \tau_i},$$

где A_i и τ_i – амплитуды и времена затухания i-го компонента. Среднее время затухания составило 233 нс, что практически совпадает с полученным ранее значением для плотно упакованных ансамблей КТ [14].

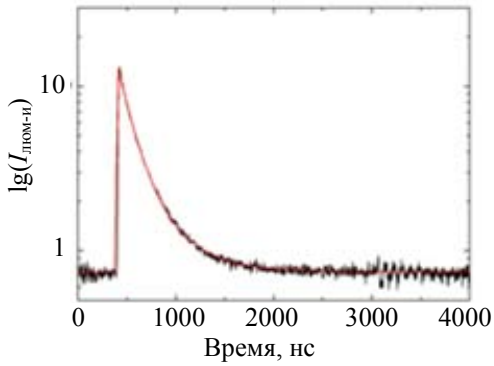


Рис. 6. Кривая затухания люминесценции суперкристалла из PbS КТ с размером 4 нм (черная линия); моделирование экспериментальных данных с помощью биэкспоненциальной зависимости (красная линия). Для удобства представления интенсивность люминесценции по оси ординат приведена в логарифмической шкале

Заключение

Полученные экспериментальные данные позволяют сделать вывод, что самоорганизованная структура из квантовых точек сульфида свинца, полученная путем испарения растворителя из коллоидного раствора квантовых точек на слюде, представляет собой суперкристалл. Полученные структуры имеют микронные размеры. Методом рассеяния рентгеновского излучения в области малых углов определено, что полученная структура обладает ромбической сингонией с примитивной кристаллической решеткой. При образовании суперкристалла из квантовых точек наблюдается уширение полосы поглощения по сравнению с коллоидными квантовыми точками, что связано с упорядочением квантовых точек и образованием объемных структур, как было ранее предсказано в [15]. При этом ширина полосы люминесценции практически не изменилась. Более глубокое понимание процессов самоорганизации квантовых точек в 3D-структуры требует проведения дальнейших исследований формирования объемных суперструктур на основе коллоидных квантовых точек разных размеров.

Полученные данные об объемных самоорганизованных структурах из нанокристаллов PbS представляют существенный интерес для разработки элементов солнечных батарей на основе тонких слоев квантовых точек халькогенидов свинца, поглощающих свет в прозрачной для кремниевых элементов ближней инфракрасной области спектра.

Литература

1. Федоров А.В., Баранов А.В. Оптика квантовых точек // Оптика наноструктур / Под ред. А.В. Федорова. СПб.: Недра. 2005. 326 с.
2. Collier C.P., Vossmeyer T., Heath J.R. Nanocrystal superlattices // Annual Review of Physical Chemistry. 1998. V. 49. N 1. P. 371–404.
3. Algar W.R., Tavares A.J., Krull U.J. Beyond labels: a review of the application of quantum dots as integrated components of assays, bioprobes, and biosensors utilizing optical transduction // Analytica Chimica Acta. 2010. V. 673. N 1. P. 1–25.
4. Giansante C., Carbone L., Giannini C., Altamura D., Ameer Z. et al. Colloidal arenethiolate-capped PbS quantum dots: optoelectronic properties, self-assembly, and application in solution-cast photovoltaics // The Journal of Physical Chemistry C. 2013. V. 117. N 25. P. 13305–13317.
5. Ushakova E.V., Litvin A.P., Parfenov P.S., Fedorov A.V., Artemyev M., Prudnikau A.V., Rukhlenko I.D., Baranov A.V. Anomalous size-dependent decay of low-energy luminescence from PbS quantum dots in colloidal solution // ACS Nano. 2012. V. 6. N 10. P. 8913–8921.
6. Scheele M., Hanifi D., Zherebetsky D., Chourou S.T., Axnanda S. et al. PbS nanoparticles capped with tetrathiafulvalenetetracarboxylate: utilizing energy level alignment for efficient carrier transport // ACS Nano. 2014. V. 8. N 3. P. 2532–2540. doi: 10.1021/nn406127s
7. Quan Z., Xu H., Wang C., Wen X., Wang Y. et al. Solvent-mediated self-assembly of nanocube superlattices // Journal of the American Chemical Society. 2014. V. 136. N 4. P. 1352–1359.
8. Small Angle X-Ray Scattering. Eds. Glatter O., Kratky O. NY-London: Academic Press, 1982. 515 p.
9. Parfenov P.S., Baranov A.V., Veniaminov A.V., Orlova A.O. A complex for the fluorescence analysis of macro-and microsamples in the near-infrared // Journal of Optical Technology. 2011. V. 78. N 2. P. 120–123.
10. Parfenov P.S., Litvin A.P., Baranov A.V., Ushakova E.V., Fedorov A.V., Prudnikov A.V., Artemyev M.V. Measurement of the luminescence decay times of PbS quantum dots in the near-IR spectral range // Optics and Spectroscopy. 2012. V. 112. N 6. P. 868–873.

11. Parfenov P.S., Litvin A.P., Baranov A.V., Veniaminov A.V., Ushakova E.V. Calibration of the spectral sensitivity of instruments for the near infrared region // Journal of Applied Spectroscopy. 2011. V. 78. N 3. P. 433–439.
12. de Mello Donegá C., Liljeroth P., Vanmaekelbergh D. Physicochemical evaluation of the hot-injection method, a synthesis route for monodisperse nanocrystals // Small. 2005. V. 1. N 12. P. 1152–1162.
13. Ушакова Е.В., Голубков В.В., Литвин А.П., Парфенов П.С., Баранов А.В. Самоорганизация квантовых точек сульфида свинца разного размера // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 6 (86). С. 127–132.
14. Litvin A.P., Parfenov P.S., Ushakova E.V., Fedorov A.V., Artemyev M.V., Prudnikov A.V., Golubkov V.V., Baranov A.V. PbS quantum dots in a porous matrix: optical characterization // The Journal of Physical Chemistry C. 2013. V. 117. N 23. P. 12318–12324.
15. Baimuratov A.S., Rukhlenko I.D., Fedorov A.V. Engineering band structure in nanoscale quantum-dot supercrystals // Optics Letters. 2013. V. 38. N 13. P. 2259–2261.

Ушакова Елена Владимировна – кандидат физико-математических наук, младший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, el.ushakova@gmail.com

Голубков Валерий Викторович – доктор химических наук, старший научный сотрудник, заведующий лабораторией, Институт химии силикатов имени И.В. Гребенщикова Российской Академии Наук, Санкт-Петербург, 199034, Российская Федерация, golubkov@isc1.nw.ru

Осколков Евгений Олегович – лаборант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, zhekic@gmail.com

Литвин Александр Петрович – инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, litvin88@gmail.com

Парфенов Петр Сергеевич – кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, qrspeter@pochta.ru

Баранов Александр Васильевич – доктор физико-математических наук, профессор, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, a_v_baranov@yahoo.com

Elena V. Ushakova – PhD, junior research scientist, ITMO University, Saint Petersburg, 197101, Russian Federation, el.ushakova@gmail.com

Valery V. Golubkov – D.Sc., senior research scientist, Head of the laboratory, Institute of Silicate Chemistry of the Russian Academy of Sciences, 199034, Saint Petersburg, Russian Federation, golubkov@isc1.nw.ru

Evgeniy O. Oskolkov – assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, zhekic@gmail.com

Alexander P. Litvin – engineer, ITMO University, Saint Petersburg, 197101,Russian Federation, litvin88@gmail.com

Peter S. Parfenov – PhD, Associate professor, Associate professor, ITMO University, Saint Petersburg, 197101,Russian Federation, qrspeter@pochta.ru

Alexander V. Baranov – D.Sc., Professor, Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, a_v_baranov@yahoo.com

Принято к печати 10.09.14
Accepted 10.09.14

УДК 535.3, 519.85

ПРОСТРАНСТВЕННО-СЕЛЕКТИВНАЯ СПЕКЛ-КОРРЕЛОМЕТРИЯ СЛУЧАЙНО-НЕОДНОРОДНЫХ СРЕД: РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

А.А. Исаева[a], А.В. Неустроев[a]

[a]Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, isanna.1987@mail.ru

Аннотация. Представлены результаты численного моделирования методом Монте-Карло переноса излучения в средах со сложной структурой и динамикой с использованием оригинального подхода спекл-коррелометрии на основе использования локализованного источника излучения и приемника излучения с кольцевой апертурой. В качестве модельных сред рассматривались «динамические» протяженные объекты с различной геометрией и локализацией в «статическом» однородном слое, имитирующие биологические структуры с различными характеристиками микроциркуляции крови. Получены оценки коэффициента обратного рассеяния модельной среды, оцениваемого как отношение «динамических» парциальных составляющих обратно рассеянного поля к полному рассеянному полю. При этом и «динамические» парциальные составляющие обратно рассеянного поля, и полное рассеянное поле регистрируются детектором с заданным набором значений радиусов кольцевых апертур. В результате анализа зависимости коэффициентов обратного рассеяния от радиусов кольцевых детекторов были определены глубины залегания «динамического» протяженного объекта для различных случаев глубины локализации объекта. Также показано, что зависимости коэффициента обратного рассеяния от радиуса кольцевого приемника излучения для сред с различными оптическими свойствами и содержащими «динамический» объект с различными геометрическими размерами могут быть описаны δ-функцией, а наблюдаемый сдвиг пикового значения может быть обусловлен изменением показателя анизотропии рассеяния.

Ключевые слова: рассеяние, лазерное излучение, спеклы, модельные среды.

Благодарности. Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации.

SPATIALLY SELECTIVE SPECKLE-CORRELOMETRY OF RANDOM INHOMOGENEOUS MEDIA: SIMULATION RESULTS

A.A. Isaeva[a], A.V. Neustroev[a]

[a]ITMO University, Saint Petersburg, 197101, Russian Federation, isanna.1987@mail.ru

Abstract. The paper deals with the results of Monte Carlo simulation of light propagation in a media with complex structure and dynamics by an original speckle-correlometry approach based on ring-like apertures and localized source of probe light. The «dynamic» lengthy objects with different geometry and depth location in the «static» inhomogeneous layer imitating biotissues with different characteristics of blood microcirculation were chosen as simulated media. The backscattering coefficient of laser light for the simulated media evaluated as a ratio of the «dynamic» partial components of the backscattered field to the full backscattered field is obtained. At the same time the «dynamic» partial components of the backscattered field and the full backscattered field are detected by the ring detector with the set value of ring aperture radius. The depth location of «dynamic» lengthy objects was determined analyzing the results of the dependence of the backscattering coefficient on the ring detector radii. It was also shown that the dependences of the backscattering coefficient on the ring detector radius in the case of probed media with different optical properties and containing the «dynamic» lengthy object with different geometric sizes can be described by the δ-like function. But the displacement of the peak value of δ-like function can be caused by the change of the scattering anisotropy factor.

Keywords: scattering, laser light, speckles, simulated media.

Acknowledgements. The work has been carried out under financial support of the Ministry of Education and Science of the Russian Federation.

Введение

Неинвазивные методы диагностики с использованием лазерного излучения оптического диапазона широко используются в современной биомедицине. К подобным подходам следует отнести спекл-коррелометрические методы полного поля, в частности, метод LASCA (Laser Speckle Contrast Analysis), впервые предложенный Д. Брайерсом в 1996 г. [1–3]. Методы, основанные на статистическом анализе спекл-модулированных изображений поверхности объекта при его зондировании лазерным лучом, успешно применяются для исследования микрогемодинамики в поверхностных слоях внутренних органов [4–6] и процессов гемодинамики крови в капиллярах ногтевого ложа [7], процессов метаболического контроля местного кровотока в приповерхностных слоях кожи при гипертермии [8], процессов термической модификации фиброзных тканей при нагреве инфракрасным лазером [9], критических состояний границ раздела жидкой и газовой фаз в неупорядоченных пористых средах [10].

Для расширения возможностей применения спекл-коррелометрических методов, в частности, пространственно-селективной спекл-коррелометрии, в областях биомедицины необходима разработка и построение адекватных математических моделей, позволяющих проанализировать чувствительность подхода к динамическим характеристикам движущихся рассеивающих центров в анализируемой среде. В ходе работы была построена модель распределения излучения в «статической» однородной среде, содержащей протяженный «динамический» объект, с применением метода Монте-Карло и подхода спекл-коррелометрии с пространственной селекцией, и на основе построенной модели проведена оценка эффективности использования подхода для диагностики биологических тканей.

Метод спекл-коррелометрии с пространственной селекцией

Спекл-коррелометрические методы, несмотря на ряд имеющихся преимуществ, таких как неинвазивность, быстродействие, возможность диагностики в режиме реального времени, обладают низкой разрешающей способностью по глубине зондирования, что обусловлено особенностями формирования регистрируемого излучения в результате многократного рассеяния спекл-модулированного поля. Рассеянное поле, формирующее сигнал, представляет собой суперпозицию парциальных составляющих, которые распространяются в слое на различную глубину. Парциальные составляющие несут информацию о значениях подвижности рассеивающих центров, определяющих динамические процессы, протекающие в анализируемом слое. Таким образом, регистрируемый сигнал характеризует некоторое «интегральное» значение подвижности. Подобному интегральному значению соответствует значение подвижности, полученное в результате пространственного усреднения по глубине порядка транспортной длины для рассеивающей среды, которая соответствует длине волны зондирующего излучения. Транспортная длина – расстояние в рассеивающей среде, на котором теряется информация о первоначальном направлении распространения зондирующего излучения и происходит полная рандомизация волновых векторов парциальных составляющих излучения.

В работе предложен оригинальный подход на основе использования локализованного источника зондирующего излучения и селекции парциальных составляющих рассеянного поля [11–13]. Селекция на основе использования кольцевых пространственных фильтров с различными значениями радиусов внутреннего и внешнего колец позволяет осуществлять дискриминацию парциальных составляющих, проникающих в среду на различную глубину. Идея подхода проиллюстрирована на рис. 1, на котором показано проникновение зондирующего лазерного излучения в исследуемую среду. Каждый выбранный кольцевой фильтр селектирует набор фотонов, траектории которых лежат в пределах некоторого объема, соответствующего проникновению излучения на заданную глубину. Такой объем, ограничивающий траектории совокупности парциальных составляющих рассеянного излучения, имеет форму «banana shape» [12] (рис. 1). Регистрируемое излучение, проходящее через кольцевой пространственный фильтр с большими радиусами, соответствует большей глубине.

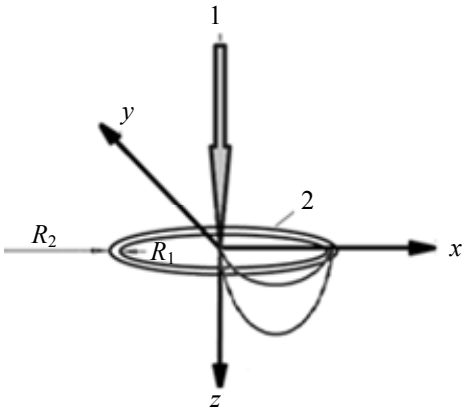


Рис. 1. Взаимная локализация источника зондирующего излучения (1) и кольцевого детектора (2): R_1 – радиус внутреннего кольца детектора; R_2 – радиус внешнего кольца детектора

Результаты моделирования

В рамках настоящей работы была проведена оценка эффективности описанного выше подхода пространственно-селективной спекл-коррелометрии с использованием метода статистического моделирования (метода Монте-Карло) [14–17].

Анализ динамических и структурных свойств биотканей, в частности, кожи человека и животного, представляет определенный интерес, так как они имеют сложную многослойную структуру с различной динамикой кровотока в каждом слое. В качестве модельной среды был выбран «статический» однородный слой, содержащий протяженный «динамический» объект с различной геометрией и глубиной залегания, имитирующий кровеносный сосуд. Модель однородной среды представляла собой плоскопараллельный слой бесконечной ширины и бесконечной толщины (в масштабе размеров объекта; для упрощения расчетов толщина полагалась равной 10 000 мкм) с коэффициентами рассеяния $\mu_s = 0,01$ мкм$^{-1}$ и поглощения $\mu_a = 10^{-11}$ мкм$^{-1}$ и параметром анизотропии ($g = 0,3$ или $g = 0,85$). Протяженный «динамический» объект представлял собой бесконечный круговой цилиндр, ось которого параллельна верхней границе однородной среды, с глубинами залегания $d_{cil} = 500$ мкм и $d_{cil} = 700$ мкм и радиусами оснований $R_{cil} = 10$ мкм и $R_{cil} = 100$ мкм (рис. 2). Радиусы «динамического» объекта были выбраны на основе анализа данных, опуб-

ликованных в литературе, согласно которым средний радиус кровеносного капилляра составляет 8–10 мкм, а артериол и посткапиллярных венул – примерно 50–150 мкм.

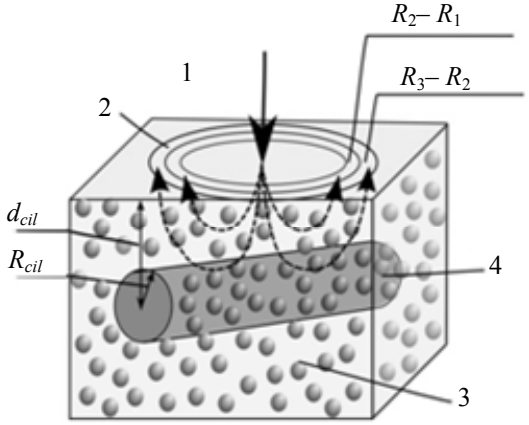


Рис. 2. Моделируемая среда и распространение излучения. Среда представлена большим количеством рассеивателей: 1 – локализованный источник зондирующего излучения (например, сфокусированный лазерный пучок); 2 – участок поверхности зондируемой среды, с которого с помощью кольцевой апертуры выделяется детектируемый оптический сигнал (R_1, R_2, R_3 – радиусы кольцевой апертуры детектора); 3 – слой, содержащий «статические» рассеиватели; 4 – протяженный объект, содержащий «динамические» рассеиватели (R_{cil} – радиус протяженного объекта, d_{cil} – глубина залегания протяженного объекта)

С учетом того, что при регистрации спекл-модулированного изображения время экспозиции превышает время корреляции, парциальные составляющие, достигшие заданного «динамического» объекта, состоящего на 100% из подвижных рассеивателей, и испытавшие даже однократное взаимодействие с рассеивателями, вносили вклад в формирование регистрируемых динамических спеклов. При этом детектирование сигнала осуществлялось с поверхности исследуемой среды с учетом радиуса кольцевого детектора. Как было отмечено выше, изменение радиуса кольцевого детектора позволяет осуществлять селекцию регистрируемого обратно рассеянного излучения, проходящего на различную глубину в среде. Кольцевые детекторы с большими радиусами регистрируют излучение, прошедшее на большую глубину. Коэффициент обратного рассеяния в зависимости от среднего радиуса кольцевого детектора оценивался как

$$R_{bs}(r_d) = \frac{I_{dyn}(r_d)}{I_{dyn+stat}(r_d)}, \qquad (1)$$

где $r_d = \frac{R_N + R_{N-1}}{2}$ – средний радиус кольцевого детектора; R_N – значения внешних радиусов детектора; R_{N-1} – значения внутренних радиусов детектора; $I_{dyn}(r_d)$ – интенсивность, пропорциональная числу фотонов, испытавших столкновения с «динамическими» рассеивателями и зарегистрированных детектором с радиусом r_d; $I_{dyn+stat}(r_d)$ – интенсивность, пропорциональная сумме фотонов, испытавших столкновения с «динамическими» и со «статическими» рассеивателями и зарегистрированных детектором с радиусом r_d. Зависимости коэффициента обратного рассеяния от расстояния между источником и детектором для различных параметров среды и геометрических размеров «динамического» объекта, рассчитанные по формуле (1) с использованием данных, полученных методом численного моделирования, представлены на рис. 3, 4.

Наблюдается немонотонная зависимость коэффициента обратного рассеяния от радиуса кольцевого детектора, при этом положение максимума кривой зависимости коэффициента обратного рассеяния, рассчитанного по результатам моделирования, определяется глубиной залегания протяженного «динамического» объекта в зондируемой среде [18] как

$$d_{cil} \approx \frac{r_d}{2\sqrt{2}}. \qquad (2)$$

Значения радиусов кольцевого детектора, рассчитанные по формуле (2) на основе результатов моделирования, полученных методом Монте-Карло, представлены в таблице.

Зависимости коэффициента обратного рассеяния от расстояния между источником и детектором для различных параметров среды и геометрических размеров «динамического» объекта (рис. 3, 4) могут быть интерпретированы как отклик среды на сигнал, описываемый δ-функцией, при этом наблюдаемый сдвиг пикового сигнала может быть обусловлен изменением показателя анизотропии рассеяния.

Показатель анизотропии среды g	Глубина залегания и радиус протяженного «динамического» объекта, d_{cil}, мкм и R_{cil}, мкм		Глубина залегания протяженного «динамического» объекта, $d_{cil}^{\text{модел}}$, мкм
$g = 0,3$	$R_{cil} = 10$	$d_{cil} = 500$	$d_{cil}^{\text{модел}} = 498$
		$d_{cil} = 700$	$d_{cil}^{\text{модел}} = 629$
	$R_{cil} = 100$	$d_{cil} = 500$	$d_{cil}^{\text{модел}} = 498$
		$d_{cil} = 700$	$d_{cil}^{\text{модел}} = 618$
$g = 0,85$	$R_{cil} = 10$	$d_{cil} = 500$	$d_{cil}^{\text{модел}} = 542$
		$d_{cil} = 700$	$d_{cil}^{\text{модел}} = 650$
	$R_{cil} = 100$	$d_{cil} = 500$	$d_{cil}^{\text{модел}} = 553$
		$d_{cil} = 700$	$d_{cil}^{\text{модел}} = 629$

Таблица. Значения глубин залегания протяженного «динамического» объекта, полученные по результатам моделирования, для различных параметров моделируемого протяженного «динамического» объекта

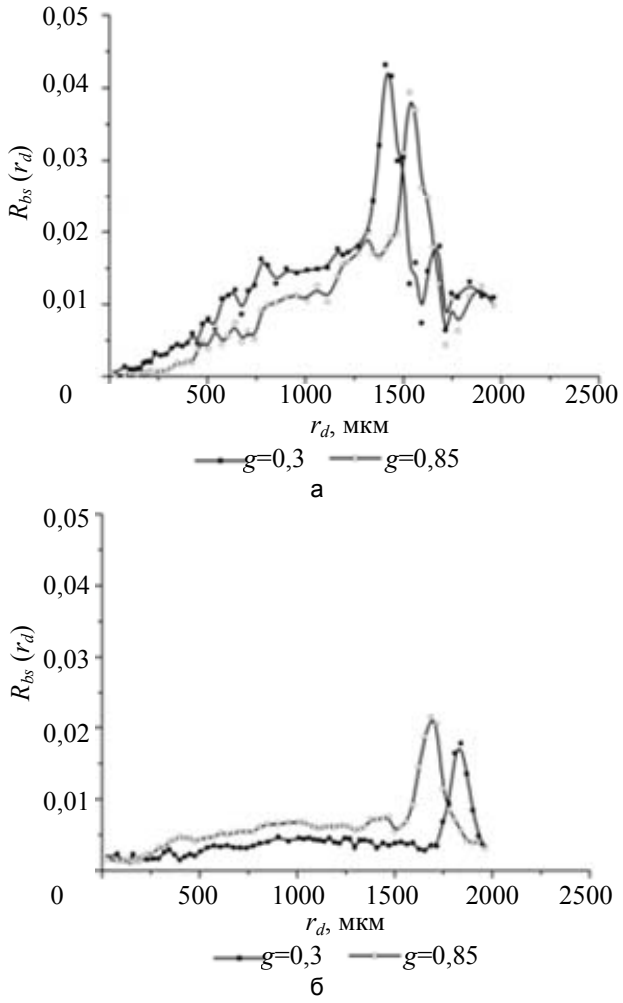
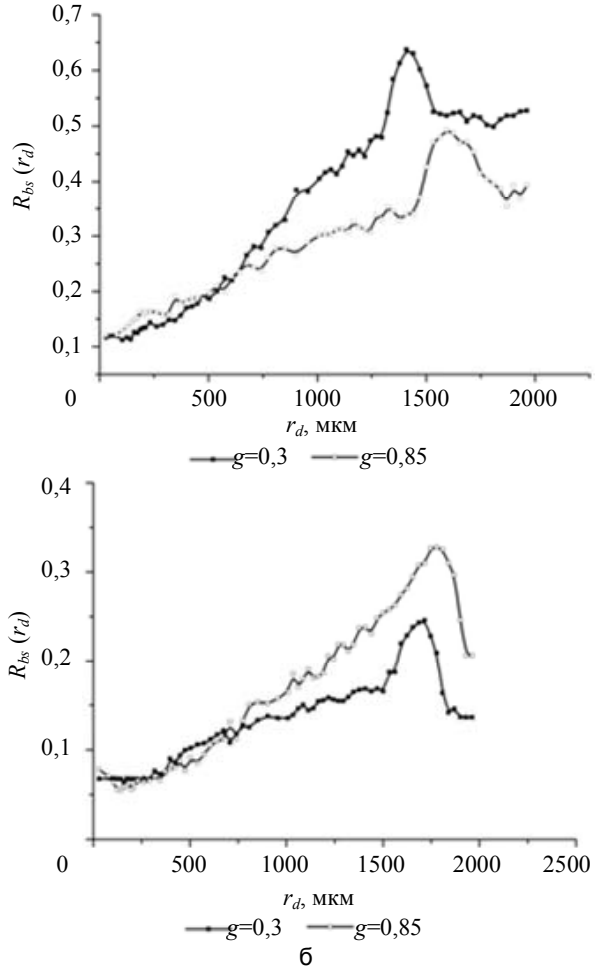


Рис. 3. Зависимость коэффициента обратного рассеяния от расстояния между детектором и источником для случая анизотропного рассеяния $g = 0,85$ и изотропного рассеяния $g = 0,3$ в однородной «статической» среде с глубинами залегания «динамического» объекта: $d_{cil} = 500$ мкм (а) и $d_{cil} = 700$ мкм (б) и радиусом $R_{cil} = 10$ мкм



б

Рис. 4. Зависимость коэффициента обратного рассеяния от расстояния между детектором и источником для случая анизотропного рассеяния $g = 0,85$ и изотропного рассеяния $g = 0,3$ в однородной «статической» среде с глубинами залегания «динамического» объекта: $d_{cil} = 500$ мкм (а) и $d_{cil} = 700$ мкм (б) и радиусом «динамического» объекта $R_{cil} = 100$ мкм

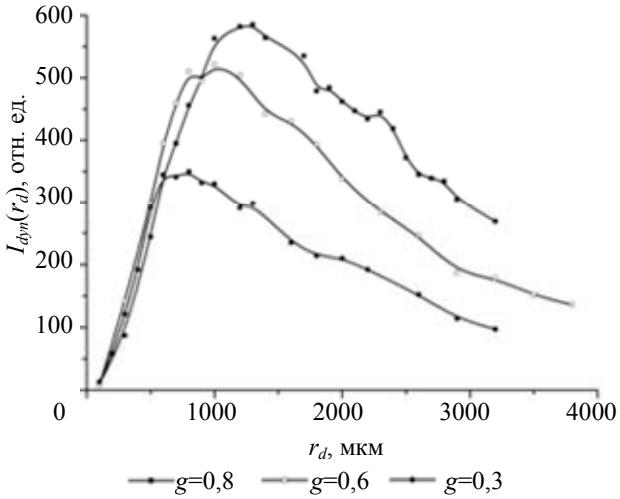


Рис. 5. Зависимость интенсивности составляющей рассеянного излучения, пропорциональной числу фотонов, испытавших столкновения с «динамическими» рассеивателями, от радиуса кольцевого детектора r_d для «статической» среды с различными показателями анизотропии, содержащей бесконечно протяженный плоский «динамический» объект [19]

Результаты моделирования, представленные в [19] для бесконечно протяженного плоского «динамического» объекта в «статическом» однородном слое, подобного сечению параллельно оси цилиндриче-

ского протяженного объекта, также демонстрируют зависимость парциальных компонент рассеянного излучения, испытавших столкновения с «динамическими» рассеивателями, от показателей анизотропии (рис. 5). Погрешность анизотропии обусловлена увеличением рандомизации направлений распространения фотонов в зондируемой среде, что вызывает смещение пика доли «динамических» фотонов (рис. 6).

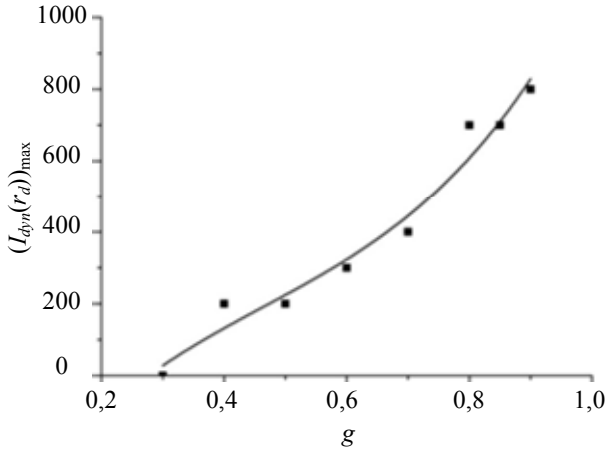


Рис. 6. Влияние показателя анизотропии среды на смещение максимума интенсивности составляющей рассеянного излучения, пропорциональной числу фотонов, испытавших столкновения с «динамическими» рассеивателями [19]

Заключение

Анализируя зависимости коэффициента обратного рассеяния от радиуса кольцевого детектора для случая изотропного и анизотропного рассеяния в однородной «статической» среде, содержащей «динамический» протяженный объект с различной глубиной залегания, следует отметить следующее.

1. Наблюдается уширение динамического отклика среды на падающий сигнал, описываемый δ-функцией, при условии увеличения радиуса цилиндрического «динамического» объекта, и возрастание доли парциальных составляющих обратно рассеянного излучения, испытавших динамическое рассеяние.
2. Наблюдается сдвиг пикового значения динамического отклика среды в условиях изотропного ($g = 0,3$) и анизотропного рассеяния ($g = 0,85$).

Исходя из зависимостей коэффициента обратного рассеяния излучения с учетом «динамических» и «статических» парциальных составляющих поля, полученных на основе подхода с селекцией многократно рассеянного излучения с использованием пространственных кольцевых фильтров с различными значениями внешнего и внутреннего радиусов, можно получить оценку глубины проникновения излучения и, соответственно, определить глубину залегания протяженного объекта, отличающегося от окружающей случайно-неоднородной среды своими характеристиками подвижности.

Таким образом, полученные результаты могут быть использованы при развитии метода калибровки спекл-коррелометрии полного поля с пространственной кольцевой фильтрацией регистрируемого излучения и демонстрируют возможность интерпретации результатов визуализации и диагностики биологических тканей.

Литература

1. Briers J.D. Webster S. Laser speckle contrast analysis (LASCA): a nonscanning, full-field technique for monitoring capillary blood flow // Journal of Biomedical Optics. 1996. V. 1. N 2. P. 174−179.
2. Briers J.D. Laser Doppler and time-varying speckle: a reconciliation // Journal of Optical Society of America A: Optics and Image Science, and Vision. 1996. V. 13. N 2. P. 345−350.
3. Richards G. and Briers J.D. Capillary blood flow monitoring using laser speckle contrast analysis (LASCA): improving the dynamic range // Proc. of SPIE. 1997. V. 2981. P. 160−171.
4. Зимняков Д.А., Хмара М.Б., Виленский М.А., Козлов В.В., Садовой А.В., Горфинкель И.В., Здражевский Р.А., Исаева А.А. Спекл-корреляционный мониторинг микрогемодинамики внутренних органов // Оптика и спектроскопия. 2009. Т. 107. № 6. С. 941−947.
5. Skipetrov S.E., Maynard R. Dynamic multiple scattering of light in multilayer turbid media // Physics Letters, Section A: General, Atomic and Solid State Physics. 1996. V. 217. N 2–3. P. 181−185.
6. Boas D.A., Dunn A.K. Laser speckle contrast imaging in biomedical optics // Journal of Biomedical Optics. 2010. V. 15. N 1. Art. 011109.
7. Виленский М.А., Агафонов Д.Н., Зимняков Д.А., Тучин В.В., Здражевский Р.А. Спекл-корреляционный анализ микрокапиллярного кровотока ногтевого ложа // Квантовая электроника. 2011. Т. 41. № 4. С. 324–328.

8. Roustit M., Millet C., Blaise S., Dufournet B., Cracowski J.L. Excellent reproducibility of laser speckle contrast imaging to assess skin microvascular reactivity // Microvascular Research. 2010. V. 80. N 3. P. 505–511.
9. Зимняков Д.А., Свиридов А.П., Кузнецова Л.В., Баранов С.А., Игнатьева Н.Ю., Лунин В.В. Анализ кинетики термической модификации биотканей методом спекл-коррелометрии // Журнал физической химии. 2007. Т. 81. № 4. С. 725–731.
10. Зимняков Д.А., Садовой А.В., Виленский М.А., Захаров П.В., Миллюля Р. Критическое поведение границ раздела фаз в пористых средах: анализ масштабных свойств с использованием некогерентного и когерентного света // Журнал экспериментальной и теоретической физики. 2009. Т. 135. № 2. С. 351–368.
11. Boas D.A., Yodh A.G. Spatially varying dynamical properties of turbid media probed with diffusing temporal light correlation // Journal of Optical Society of America A: Optics and Image Science, and Vision. 1997. V. 14. N 1. P. 192–215.
12. Lemieux P.-A., Vera M.U., Durian D.J. Diffusing-light spectroscopies beyond the diffusion limit: the role of ballistic transport and anisotropic scattering // Physical Review E. 1998. V. 57. N 4. P. 4498–4515.
13. Зимняков Д.А., Исаева А.А., Исаева Е.А., Ушакова О.В., Здражевский Р.А. О спекл-коррелометрическом методе оценки транспортного коэффициента рассеяния случайно-неоднородных сред // Письма в журнал технической физики. 2012. Т. 38. № 20. С. 43–49.
14. Wang L., Jacques S.L., Zheng L. MCML-Monte Carlo modeling of light transport in multi-layered tissues // Computer Methods and Programs in Biomedicine. 1995. V. 47. N 2. P. 131–146.
15. Воробьева Е.А., Гуров И.П. Модели распространения и рассеяния оптического излучения в случайно-неоднородных средах. В кн. Проблемы когерентной и нелинейной оптики / Под ред. И.П. Гурова, С.А. Козлова. СПб.: СПб ГИТМО (ТУ), 2006. С. 82–98.
16. Kuzmin V.L., Meglinski I.V. Coherent multiple scattering effects and Monte Carlo method // JETP Letters. 2004. V. 79. N 3. P. 109–112.
17. Berrocal E., Sedarsky D.L., Paciaroni M.E., Meglinski I.V., Linne M.A. Laser light scattering in turbid media Part I: Experimental and simulated results for the spatial intensity distribution // Optics Express. 2007. V. 15. N 17. P. 10649–10665.
18. Feng S., Zeng F.-A., Chance B. Photon migration in the presence of a single defect: a perturbation analysis // Applied Optics. 1995. V. 34. N 19. P. 3826–3837.
19. Isaeva A.A., Zimnyakov D.A. Full-field speckle analysis of spatially heterogeneous scatter dynamics with the improved depth resolution in stratified random media // Proc. of SPIE. 2011. V. 8338. Art. 83380Y1.

Исаева Анна Андреевна – кандидат физико-математических наук, ассистент кафедры, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, isanna.1987@mail.ru

Неустроев Артем Вячеславович – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, neustroev.artyom@yandex.ru

Anna A. Isaeva – PhD, assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, isanna.1987@mail.ru

Artem V. Neustroev – student, ITMO University, Saint Petersburg, 197101, Russian Federation, neustroev.artyom@yandex.ru

Принято к печати 06.10.14
Accepted 06.10.14

УДК 681.787:[519.245+519.6]

АНАЛИЗ ВЫЧИСЛИТЕЛЬНОЙ СЛОЖНОСТИ РЕКУРРЕНТНЫХ АЛГОРИТМОВ ОБРАБОТКИ ДАННЫХ В ОПТИЧЕСКОЙ КОГЕРЕНТНОЙ ТОМОГРАФИИ

М.А. Волынский[a], И.П. Гуров[a], П.А. Ермолаев[a], П.С. Скаков[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, maxim.volynsky@gmail.com

Аннотация. Рассмотрены основные принципы представления сигналов в оптической когерентной томографии с использованием формализма теории динамических систем; проведен сравнительный анализ вычислительной сложности алгоритмов динамического оценивания параметров сигналов в оптической когерентной томографии, таких как расширенный фильтр Калмана и последовательный метод Монте-Карло. Показано, что вычислительная сложность обработки одного отсчета сигнала при помощи расширенного фильтра Калмана полиномиально возрастает в зависимости от размера вектора параметров и вектора наблюдения, а сложность обработки отсчета сигнала последовательным методом Монте-Карло линейно зависит как от размеров вектора параметров и вектора наблюдения, так и от количества генерируемых случайных векторов. Приведены экспериментальные результаты оценивания времени обработки тестового сигнала при использовании каждого из алгоритмов. Показано, что время обработки сигнала, содержащего 500 дискретных отсчетов, при помощи расширенного фильтра Калмана в случае простейшей модели скалярного сигнала составляет примерно 0,1 с и возрастает при усложнении модели в несколько раз. Время обработки аналогичного сигнала при помощи последовательного метода Монте-Карло с использованием аналогичной простейшей модели и при фиксированном количестве генерируемых векторов составляет 0,7 с и при усложнении модели возрастает незначительно, примерно в 1,5 раза. Полученные результаты могут быть использованы при оценке ожидаемого времени обработки данных с помощью рекуррентных алгоритмов динамического оценивания параметров в системах оптической когерентной томографии.

Ключевые слова: оптическая когерентная томография, обработка интерферометрических сигналов, вычислительная сложность, расширенный фильтр Калмана, последовательный метод Монте-Карло.

Благодарности. Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации.

COMPUTATIONAL COMPLEXITY ANALYSIS OF RECURRENT DATA PROCESSING ALGORITHMS IN OPTICAL COHERENCE TOMOGRAPHY

M.A. Volynsky[a], I.P. Gurov[a], P.A. Ermolaev[a], P.S. Skakov[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, maxim.volynsky@gmail.com

Abstract. The paper deals with the basic principles of signals representation in optical coherence tomography with the usage of dynamic systems theory formalism. Computational complexity of algorithms for dynamic estimation of signals parameters is analyzed, such as extended Kalman filter and sequential Monte-Carlo method. It is shown that processing time of one discrete-time sample of the signal by extended Kalman filter increases polynomially with sizes of parameters vector and observation vector. Processing time of one discrete-time sample of the signal by sequential Monte-Carlo method depends linearly both on sizes of parameters vector and observation vector, and on the number of generating random vectors. Experimental results of processing time measurement by each algorithm are described. It is shown that processing time of the signal containing 500 discrete-time samples by extended Kalman filter in the case of the simplest model is approximately equal to 0.1 seconds and increases several times with complication of the model. Processing time of the same signal by sequential Monte-Carlo methods with fixed number of generated random vectors is equal to 0.7 seconds and slightly increases with complication of the model, approximately by 1.5 times. Obtained results may be used for estimation of expected data processing time by recurrent dynamic estimation algorithms in optical coherence tomography systems.

Keywords: optical coherence tomography, interferometric signals processing, computational complexity, extended Kalman filter, sequential Monte-Carlo method.

Acknowledgements. The work has been carried out with the financial support of the Ministry of Education and Science of the Russian Federation.

Введение

Оптическая когерентная томография (ОКТ) является одним из перспективных методов исследования микроструктуры объектов с высоким разрешением и широко применяется в различных областях медицины и биологии [1–6]. Системы ОКТ подразделяются по типу регистрируемых сигналов на корреляционные [2] и спектральные [3]. В первом случае на выходе системы регистрируется интерферометрический сигнал малой когерентности, а во втором случае – формируются полосы равного хроматического порядка. В этих сигналах содержится информация о внутренней микроструктуре объекта, для извлечения которой требуются специальные алгоритмы обработки сигналов [7].

Критическими требованиями к таким алгоритмам, помимо качества обработки, являются высокое быстродействие и малое количество требуемой памяти. Эти характеристики могут быть оценены в рамках теории вычислительной сложности [8, 9].

Традиционно для обработки сигналов в ОКТ используются алгоритмы на основе преобразования Фурье, однако их применение в ряде случаев сопряжено с недостатками, например, в системах, работающих в реальном времени, так как применение дискретного преобразования Фурье требует наличия полной реализации обрабатываемого сигнала.

Преимуществом рекуррентных алгоритмов динамического оценивания [7, 10], основанных на формализме стохастических дифференциальных уравнений [7], является использование априорной информации о процессе формирования сигнала и статистических характеристиках шума.

Наиболее распространенным алгоритмом динамического оценивания параметров является линейный фильтр Калмана [11, 12]. Этот алгоритм оптимален с точки зрения минимума средней квадратичной ошибки оценки. В системах ОКТ наблюдаемое значение сигнала зависит от параметров сигнала нелинейно, что ограничивает возможность использования алгоритма линейной фильтрации для оценивания параметров таких систем. Эта проблема может быть решена при помощи расширенного фильтра Калмана (РФК) [12, 13], в котором используется линеаризация уравнений динамической системы и (или) наблюдения путем разложения их в ряд Тейлора по параметрам.

Альтернативным подходом к оцениванию параметров нелинейных динамических систем является последовательный метод Монте-Карло (ПММК) [12, 14–17], работа которого основана на статистической аппроксимации плотности вероятности распределения параметров. ПММК является более перспективным методом обработки по сравнению с РФК, так как в условиях априорной неопределенности модели сигнала предоставляет больше возможностей по адаптации алгоритма путем изменения моделируемых плотностей вероятностей распределения параметров и более устойчив к неточности задания начальных условий [18].

В настоящей работе проведен сравнительный анализ вычислительной сложности РФК и ПММК на примере задачи обработки сигналов в корреляционной ОКТ. Даны рекомендации по использованию рассматриваемых алгоритмов.

Модель интерферометрического сигнала

Сигнал в корреляционной ОКТ может быть представлен параметрически в виде дискретной последовательности отсчетов [7]:

$$s(k) = B(k) + A(k)\cos(\Phi(k) + \delta\varphi(k)) + n(k), \quad (1)$$

где $k = 0..K–1$ – номер дискретного отсчета; $B(k)$ – фоновая составляющая; $A(k)$ – амплитуда; $n(k)$ – не коррелированный с сигналом белый шум, распределенный по нормальному закону с нулевым средним; $\delta\varphi(k)$ – случайные флуктуации фазы; $\Phi(k)$ – полная фаза сигнала, определяемая как

$$\Phi(k) = \sum_{k'=0}^{k} 2\pi f(k')\Delta z,$$

где $f(k')$ – частота; Δz – шаг дискретизации.

Набор параметров в уравнении (1) можно записать в виде вектора параметров

$$\boldsymbol{\theta} = (B, A, f, \Phi)^T. \quad (2)$$

Для применения рекуррентных алгоритмов динамического оценивания параметров требуется записать модель формирования интерферометрического сигнала (1), используя формализм теории динамических систем. Динамическая система задается в общем виде уравнениями системы и уравнением наблюдения:

$$\boldsymbol{\theta}(k) = \mathbf{f}\left(\boldsymbol{\theta}(k-1), \mathbf{w}(k)\right), \quad (3)$$

$$\mathbf{s}(k) = \mathbf{h}\left(\boldsymbol{\theta}(k), \mathbf{n}(k)\right), \quad (4)$$

где $\mathbf{f}(\boldsymbol{\theta})$ и $\mathbf{h}(\boldsymbol{\theta})$ – известные нелинейные векторные функции, $\mathbf{w}$ и $\mathbf{n}$ – случайные шумы системы и наблюдения, а вектор $\boldsymbol{\theta}$ имеет известную плотность вероятности распределения параметров на шаге $k = 0$. С учетом (1) и (2) для случая неизменных фоновой составляющей, амплитуды и частоты в пределах шага дискретизации векторные функции в уравнениях (3) и (4) можно записать как

$$\mathbf{h}(\boldsymbol{\theta}) = B + A\cos\Phi,$$

$$\mathbf{f}(\boldsymbol{\theta}) = \boldsymbol{\theta} + (0, 0, 0, 2\pi f \Delta z)^T.$$

Динамическая обработка сигналов в ОКТ заключается в получении оценки вектора параметров (2) для каждого отсчета сигнала (1).

В зависимости от решаемой задачи и способа формирования сигналов для их описания могут быть использованы другие модели, в том числе модели с векторным наблюдением (когда результат вычисления функции $\mathbf{h}(\boldsymbol{\theta})$ является вектором). Примером может служить оценивание параметров сигнала по нескольким предшествующим отсчетам по последовательности видеокадров, что позволяет повысить точность оценки частоты сигнала.

Краткое описание алгоритмов

РФК инициализируется заданием нелинейных функций $\mathbf{h}(\boldsymbol{\theta})$ и $\mathbf{f}(\boldsymbol{\theta})$, начальных значений параметров на шаге $k = 0$ и ковариационных матриц шума системы и шума наблюдения. Начальные значения

параметров и ковариационные матрицы могут быть найдены путем предварительного анализа интерферометрического сигнала.

Работа фильтра включает два этапа – предсказание и коррекцию по результатам наблюдения. На первом этапе происходит предсказание значений ковариационной матрицы ошибок и экстраполяция значений вектора параметров с учетом функции $\mathbf{f}(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}}(k) = \mathbf{f}(\boldsymbol{\theta}(k-1)).$$

На этапе коррекции осуществляется корректировка предсказанного значения $\hat{\boldsymbol{\theta}}(k)$ с учетом наблюдений на текущем шаге при помощи уравнения

$$\boldsymbol{\theta}(k) = \hat{\boldsymbol{\theta}}(k) + \mathbf{P}(k)[\mathbf{s}(k) - \mathbf{h}(\hat{\boldsymbol{\theta}}(k))],$$

где $\mathbf{s}(k)$ – вектор наблюдения на текущем шаге; $\mathbf{P}(k)$ – матричный коэффициент усиления фильтра, определяющий вклад невязки наблюдения и предсказания в оценку вектора параметров. Коэффициент усиления вычисляется как

$$\mathbf{P}(k) = \mathbf{R}_{pr}\mathbf{H}(k)^T(\mathbf{H}(k)\mathbf{R}_{pr}\mathbf{H}(k)^T + \mathbf{R}_n)^{-1},$$

где $\mathbf{R}_{pr}$ – предсказание ковариационной матрицы ошибок; $\mathbf{R}_n$ – ковариационная матрица шума наблюдения, а $\mathbf{H}(k)$ – матрица первых производных функции $\mathbf{h}(\boldsymbol{\theta})$ по компонентам вектора параметров. Подробнее особенности динамического оценивания параметров при помощи РФК и его модификаций рассмотрены в работах [7, 12, 19].

Инициализация алгоритма, реализующего ПММК, осуществляется не только заданием функций $\mathbf{h}(\boldsymbol{\theta})$, $\mathbf{f}(\boldsymbol{\theta})$ и начальных значений параметров на шаге k = 0, но и заданием количества генерируемых случайных векторов, правила отбора векторов (например, пороговой вероятности) и начальной функции плотности распределения параметров $p(\boldsymbol{\theta}(0))$. В простейшем случае задаются моменты нормального распределения для каждого параметра.

Обработка одного отсчета сигнала при помощи ПММК может быть разделена на четыре этапа. На первом этапе в соответствии с известной плотностью вероятности распределения параметров происходит генерация набора из N случайных векторов параметров. На втором этапе для каждого сгенерированного вектора осуществляется экстраполяция значений параметров в соответствии с функцией $\mathbf{f}(\boldsymbol{\theta})$. На третьем этапе осуществляется отбор векторов параметров, которые наилучшим образом удовлетворяют наблюдениям. Критерием отбора может служить, например, пороговое значение некоторой меры совпадения [10, 14, 18]. На четвертом этапе работы ПММК оценивается искомый вектор параметров (например, как среднее значений отобранных векторов) и корректируется функция плотности вероятности распределения параметров в соответствии с распределением отобранных векторов. Подробнее процесс обработки квазигармонических сигналов при помощи ПММК рассмотрен в работе [18].

Теоретическая оценка вычислительной сложности алгоритмов

Мерой вычислительной сложности считается количество элементарных операций, необходимых для выполнения алгоритма на ЭВМ. При оценке сложности пользуются асимптотическими оценками, определяющими порядок роста количества элементарных операций в зависимости от размера входных данных. Вычислительная сложность алгоритмов динамического оценивания зависит от количества отсчетов K в обрабатываемом сигнале, размеров вектора параметров и вектора наблюдения. Обозначим размеры вектора параметров и вектора наблюдения как $p \times 1$ и $q \times 1$ соответственно. Тогда ковариационная матрица шума системы имеет размер $p \times p$, шума наблюдения – $q \times q$, ошибок – $p \times p$. Размеры матриц производных функций $\mathbf{h}(\boldsymbol{\theta})$ и $\mathbf{f}(\boldsymbol{\theta})$ равны соответственно $p \times p$ и $q \times p$, а размер коэффициента усиления – $p \times q$ [12].

При обработке сигналов при помощи РФК наибольших временных затрат требуют расчеты, связанные с умножением и обращением матриц [12]. Простейший алгоритм умножения матриц требует $O(m^3)$ операций, где m – размер умножаемых матриц, следовательно, сложность уравнений РФК, в которых используется умножение матриц, может быть оценена как $O(p^3 + q^3)$. Существуют более эффективные алгоритмы, применяемые для умножения больших матриц, однако их использование для матриц малого размера нецелесообразно.

Известно, что обращение матрицы эквивалентно умножению двух матриц [8, 9] и также может быть выполнено в простейшем случае за $O(m^3)$ операций. Из этого следует, что вычислительная сложность расчета уравнений, содержащих обращение матриц, может быть оценена аналогично как $O(p^3 + q^3)$.

Так как функции $\mathbf{f}(\boldsymbol{\theta})$ и $\mathbf{h}(\boldsymbol{\theta})$ известны и не меняются в процессе динамического оценивания, уравнения для вычисления значений матриц их производных могут быть найдены априорно и заданы при инициализации фильтра. Сложность вычисления элементарных функций на ЭВМ оценивается как $O(1)$, так как зависит от разрядности используемой архитектуры и требуемой точности, а не от размера вход-

ных данных. Так, сложность вычисления функций $\mathbf{f}(\boldsymbol{\theta})$ и $\mathbf{h}(\boldsymbol{\theta})$ может быть оценена соответственно как $O(p^2)$ и $O(pq)$, а сложность расчета их производных – как $O(p^3)$ и $O(p^2q)$.

Общая вычислительная сложность РФК применительно к обработке сигнала из K отсчетов составляет $O(K(p^3 + q^3))$. Время обработки отсчета сигнала увеличивается полиномиально в зависимости от размеров векторов параметров и наблюдения. В работе [20] описаны методы оптимизации процесса динамического оценивания параметров при помощи РФК.

На вычислительную сложность ПММК влияют не только размеры векторов параметров и наблюдения, но и количество генерируемых случайных векторов N. Сложность генерации одного случайного числа составляет $O(1)$, следовательно, сложность первого этапа работы алгоритма – $O(pN)$. Так как вычисление функций $\mathbf{h}(\boldsymbol{\theta})$ и $\mathbf{f}(\boldsymbol{\theta})$ осуществляется для каждого из сгенерированных векторов, сложность второго этапа оценивается как $O(N(p^2 + pq))$.

Сложность третьего этапа зависит от метода отбора векторов. Когда критерием отбора является порог меры совпадения [10, 14, 18], сложность этапа равна $O(N)$. Альтернативным подходом является отбор фиксированного количества векторов, лучше всего удовлетворяющих наблюдению, что требует выполнения сортировки случайных векторов. В этом случае при использовании методов быстрой сортировки сложность третьего этапа может быть оценена как $O(N\log N)$.

Плотность распределения параметров корректируется на четвертом этапе независимо для каждого параметра и имеет сложность $O(p)$. Общая вычислительная сложность ПММК в случае отбора векторов по критерию порога меры совпадения может быть оценена как $O(N(p^2 + pq))$, а в случае отбора фиксированного количества векторов – как $O(N(p^2 + pq + \log N))$.

Экспериментальное сравнение времени работы алгоритмов

Экспериментальная оценка времени работы алгоритмов проводилась с использованием процессора Intel Core i7-4500U с номинальной тактовой частотой 1,8 ГГц. Тестовый сигнал содержал $K = 500$ отсчетов. В таблице представлено экспериментальное время работы алгоритмов в зависимости от размеров векторов параметров и наблюдения. Количество генерируемых векторов в ПММК $N = 150$.

Размер вектора параметров	Размер вектора наблюдения	Время работы РФК, с	Время работы ПММК, с
2	1	0,1	0,7
3	1	0,2	0,7
4	1	0,2	0,7
3	8	0,3	0,8
4	8	0,4	0,9
5	8	0,5	1,0
6	8	0,8	1,0

Таблица. Экспериментальные результаты оценки времени работы РФК и ПММК в зависимости от размеров векторов параметров и наблюдения в используемой модели

Видно, что время работы РФК растет быстрее, чем время работы ПММК с увеличением размеров векторов параметров и наблюдения. Так, при больших размерах этих векторов обработка сигнала при помощи ПММК требует незначительно больше времени, чем при помощи РФК. На рисунке представлен экспериментальный график зависимости времени работы ПММК от количества генерируемых случайных векторов. Видно, что эта зависимость близка к линейной.

Рисунок. Зависимость времени работы ПММК от количества генерируемых векторов параметров N

При создании систем ОКТ необходимо учитывать и объемы памяти, используемой алгоритмами обработки данных, однако этот вопрос требует отдельного рассмотрения, выходящего за рамки данной работы.

Заключение

В работе проведен анализ вычислительной сложности расширенного фильтра Калмана и последовательного метода Монте-Карло применительно к обработке данных в оптической когерентной томографии. Показано, что количество элементарных операций, требуемых для работы алгоритмов, растет линейно в зависимости от количества отсчетов в обрабатываемом сигнале.

Так как в процессе работы расширенного фильтра Калмана используются операции умножения и обращения матриц, вычислительная сложность обработки одного отсчета сигнала полиномиально зависит от размеров вектора параметров и вектора наблюдения. Время обработки отсчета сигнала последовательным методом Монте-Карло линейно зависит как от размеров вектора параметров и вектора наблюдения, так и от количества генерируемых случайных векторов, которое обычно значительно превышает размеры этих векторов и составляет несколько сотен, вследствие чего время работы последовательного метода Монте-Карло выше, чем время работы расширенного фильтра Калмана.

При малых размерах вектора параметров и вектора наблюдения и при наличии достаточно точной априорной информации о начальных значениях параметров и характеристиках шума рекомендуется использовать расширенный фильтр Калмана, так как он обеспечивает более высокое быстродействие. Последовательный метод Монте-Карло допускает существенную априорную неопределенность модели сигнала и обеспечивает большую устойчивость к шуму и неизвестным начальным значениям параметров, однако требует более длительного времени работы, чем расширенный фильтр Калмана при малых размерах вектора параметров и вектора наблюдения. При больших размерах вектора параметров системы и вектора наблюдения рекомендуется использовать последовательный метод Монте-Карло, так как это требует незначительно большего времени, чем использование расширенного фильтра Калмана.

Литература

1. Deck L., de Groot P. High-speed non-contact profiler based on scanning white light interferometry // Applied Optics. 1994. V. 33. P. 7334–7338.
2. Гуров И.П. Оптическая когерентная томография: принципы, проблемы и перспективы. В кн.: Проблемы когерентной и нелинейной оптики / Под ред. И.П. Гурова, С.А. Козлова. СПб.: СПбГУ ИТМО, 2004. С. 6–30.
3. Fercher A. Optical coherence tomography // Journal of Biomedical Optics. 1996. V. 1. N 2. P. 157–173.
4. Волынский М.А., Воробьева Е.А., Гуров И.П., Маргарянц Н.Б. Бесконтактный контроль микрообъектов методами интерферометрии малой когерентности и оптической когерентной томографии // Изв. вузов. Приборостроение. 2011. Т. 54. № 2. С. 75–82.
5. Wyant J.C. Interferometric optical metrology: basic principles and new systems // Laser Focus with Fiveroptic Technology. 1982. V. 18. N 5. P. 65–71.
6. Alarousu E., Krehut L., Prykari T., Myllyla R. Study on the use of optical coherence tomography in measurements of paper properties // Measurement Science and Technology. 2005. V. 16. N 5. P. 1131–1137.
7. Gurov I., Volynsky M. Interference fringe analysis based on recurrence computational algorithms // Optics and Lasers in Engineering. 2012. V. 50. N 4. P. 514–521.
8. Грин Д., Кнут Д. Математические методы анализа алгоритмов. М.: Мир, 1987. 120 с.
9. Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. Алгоритмы: построение и анализ. 2-е изд. М.: Вильямс, 2005. 1296 с.
10. Gurov I., Ermolaeva E., Zakharov A. Analysis of low-coherence interference fringes by the Kalman filtering method // Journal of the Optical Society of America A. 2004. V. 21. N 2. P. 242–251.
11. Kalman R.E. A new approach to linear filtering and prediction problems // Trans. ASME, J. Basic Eng. 1960. V. 82. P. 35–45.
12. Simon D. Optimal state estimation: Kalman, H∞, and nonlinear approaches. NY: John Wiley & Sons, Inc., 2006. 526 p.
13. Simon D. Using nonlinear Kalman filtering to estimate signals // Embedded Systems Design. 2006. V. 19. N 7. P. 38–53.
14. Doucet A., de Freitas N., Gordon N. Sequential Monte Carlo methods in practice. NY: Springer-Verlag, 2001. 583 p.
15. Gordon N.J., Salmond D.J., Smith A.F.M. Novel approach to nonlinear/non-Gaussian Bayesian state estimation // IEE Proceedings Part F: Radar and Signal Processing. 1993. V. 140. N 2. P. 107–113.
16. Ristic B., Arulampalam S., Gordon N. Beyond the Kalman filter: particle filters for tracking applications. Boston: Artech House, 2004. 318 p.

17. Gilks W., Berzuini C. Following a moving target – Monte Carlo inference for dynamic Bayesian models // Journal of the Royal Statistical Society. Series B: Statistical Maethodology. 2001. V. 63. N 1. P. 127–146.
18. Волынский М.А., Гуров И.П., Ермолаев П.А., Скаков П.С. Динамическое оценивание параметров интерферометрических сигналов на основе последовательного метода Монте-Карло // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 3 (91). С. 18–23.
19. Ермолаев П.А. Динамическое оценивание параметров интерферометрических сигналов методом расширенной фильтрации Калмана второго порядка // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 2 (90). С. 17–22.
20. Гуров И.П., Захаров А.С., Таратин М.А. Анализ и оптимизация вычислительного процесса нелинейной дискретной фильтрации Калмана // Изв. вузов. Приборостроение. 2004. Т. 47. № 8. С. 42–48.

Волынский Максим Александрович – кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, maxim.volynsky@gmail.com

Гуров Игорь Петрович – доктор технических наук, профессор, заведующий кафедрой, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, gurov@mail.ifmo.ru

Ермолаев Петр Андреевич – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Petr-ermolaev@hotmail.com

Скаков Павел Сергеевич – ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, pavelsx@gmail.com

Maxim A. Volynsky – PhD, Associate professor, ITMO University, Saint Petersburg, 197101, Russian Federation, maxim.volynsky@gmail.com

Igor P. Gurov – D.Sc., Professor, Department head, ITMO University, Saint Petersburg, 197101, Russian Federation, gurov@mail.ifmo.ru

Peter A. Ermolaev – student, ITMO University, Saint Petersburg, 197101, Russian Federation, Petr-ermolaev@hotmail.com

Pavel S. Skakov – assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, pavelsx@gmail.com

Принято к печати 10.10.14
Accepted 10.10.14

2 ОПТИЧЕСКИЕ СИСТЕМЫ И ТЕХНОЛОГИИ
OPTICAL ENGINEERING

УДК 004.932.2, 778.534.1

МЕТОДЫ ФОРМИРОВАНИЯ ИЗОБРАЖЕНИЙ СТЕРЕОПАРЫ С ЗАДАННЫМ ЗНАЧЕНИЕМ ПАРАЛЛАКСА

В.Г. Чафонова[a], И.В. Газеева[a]

[a] Санкт-Петербургский государственный университет кино и телевидения, Санкт-Петербург, 191119, Российская Федерация, vi777@nextmail.ru

Аннотация. Предложены два новых взаимодополняющих метода формирования изображений стереопары. Первый из них основан на нахождении максимального значения корреляционной функции между градиентными изображениями левого и правого кадров. Второй метод предполагает нахождение сдвига между двумя сопряженными ключевыми точками изображений стереопары, обнаруженными при помощи детектора точечных особенностей. Методы позволяют с высокой точностью задать выделенному на изображении объекту желаемую величину вертикального и горизонтального параллаксов. Их применение дает возможность измерить значения параллаксов у объектов на готовой стереопаре в пикселах и (или) процентах от общего размера изображения. Это позволяет заранее предсказать возможное превышение допустимых пределов величин параллаксов при печати или проекции стереопары. Предложенные методы легко автоматизируются после выделения на изображении объекта, для которого в дальнейшем будет выставлено заданное значение горизонтального параллакса. Совмещение изображений стереопары методом ключевых точек происходит менее чем за одну секунду. Метод с использованием корреляции требует чуть больше вычислительного времени, однако позволяет контролировать и совмещать неразделенное анаглифное изображение. Предложенные методы формирования стереопары могут найти применение в программах по монтажу и обработке изображений стереопары, в видеоконтрольных устройствах съемочных камер, в устройствах оценки качества видеопоследовательности.

Ключевые слова: вертикальный и горизонтальный параллакс, стереоизображение, изображения стереопары, корреляция, градиент.

METHODS OF STEREO PAIR IMAGES FORMATION WITH A GIVEN PARALLAX VALUE

V.G. Chafonova[a], I.V. Gazeeva[a]

[a] Saint Petersburg State University of Film and Television, Saint Petersburg, 191119, Russian Federation, vi777@nextmail.ru

Abstract. Two new complementary methods of stereo pair images formation are proposed. The first method is based on finding the maximum correlation between the gradient images of the left and right frames. The second one implies the finding of the shift between two corresponding key points of images for a stereo pair found by a detector of point features. These methods give the possibility to set desired values of vertical and horizontal parallaxes for the selected object in the image. Application of these methods makes it possible to measure the parallax values for the objects on the final stereo pair in pixels and / or the percentage of the total image size. It gives the possibility to predict the possible excesses in parallax values while stereo pair printing or projection. The proposed methods are easily automated after object selection, for which a predetermined value of the horizontal parallax will be exposed. Stereo pair images superposition using the key points takes less than one second. The method with correlation application requires a little bit more computing time, but makes it possible to control and superpose undivided anaglyph image. The proposed methods of stereo pair formation can find their application in programs for editing and processing images of a stereo pair, in the monitoring devices for shooting cameras and in the devices for video sequence quality assessment.

Keywords: vertical and horizontal parallax, stereo image, stereo pair images, correlation, gradient.

Введение

Изображения в стереоформате пользуются большой популярностью, однако встречаются случаи, когда их длительный просмотр вызывает дискомфорт и утомление. Во многом это связано с тем, что при создании стереоизображения не были выполнены необходимые требования, предъявляемые к параметрам стереосъемки и формируемым изображениям стереопары.

Обязательным этапом формирования стереоизображения является совмещение отдельно снятых кадров стереопары. В зависимости от значения горизонтального параллакса (рис. 1), заданного какому-либо объекту при совмещении одной и той же стереопары, будет меняться пространственное расположение всех объектов в формируемом стереоизображении [1]. Так, объект с заданным нулевым горизонтальным параллаксом воспринимается в плоскости рампы или экрана (рис. 2). Предметы, расположенные ближе данного объекта, имеют отрицательный горизонтальный параллакс и отображаются в предэкранном пространстве. Предметы, расположенные дальше данного объекта, имеют положительный горизонтальный параллакс и воспринимаются за экраном.

Для комфортного восприятия стереоизображения существует ряд требований, предъявляемых к величине параллаксов:

– вертикальный параллакс стереопары должен быть равен нулю;
– объекты на изображении, воспринимаемые в плоскости рампы (или экрана), должны иметь нулевой горизонтальный параллакс;
– горизонтальные параллаксы объектов, воспринимаемых в предэкранном и заэкранном пространствах, не должны превышать предельных значений [2, 3].

Для объектов, воспринимаемых за экраном, положительный горизонтальный параллакс на экране не должен превышать величину, равную базису зрения $B_{ЗР}$ (расстояние между оптическими центрами левого и правого глаз человека [4]). Для среднего наблюдателя $B_{ЗР} = 65$ мм. При величине параллакса, равном $B_{ЗР}$, объект, находящийся в заэкранном пространстве, воспринимается как бесконечно удаленный, а зрительные оси глаз человека, направленные на данный объект, пересекаются в бесконечности, т.е. параллельны друг другу. Рассматривание стереоизображений с параллаксом, превышающим $B_{ЗР}$, вызывает дивергенцию (расхождение) зрительных осей глаз, что является причиной утомляемости зрителей, а при увеличенных углах дивергенции – разрушения стереоскопического эффекта [5].

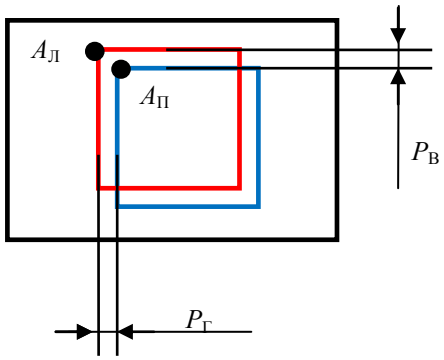


Рис. 1. Горизонтальный ($P_Г$) и вертикальный ($P_В$) параллаксы точки A изображаемого объекта; $A_Л$ и $A_П$ – изображение точки A на левом и правом кадрах соответственно (красной линией показан левый кадр стереопары, синей линией – правый)

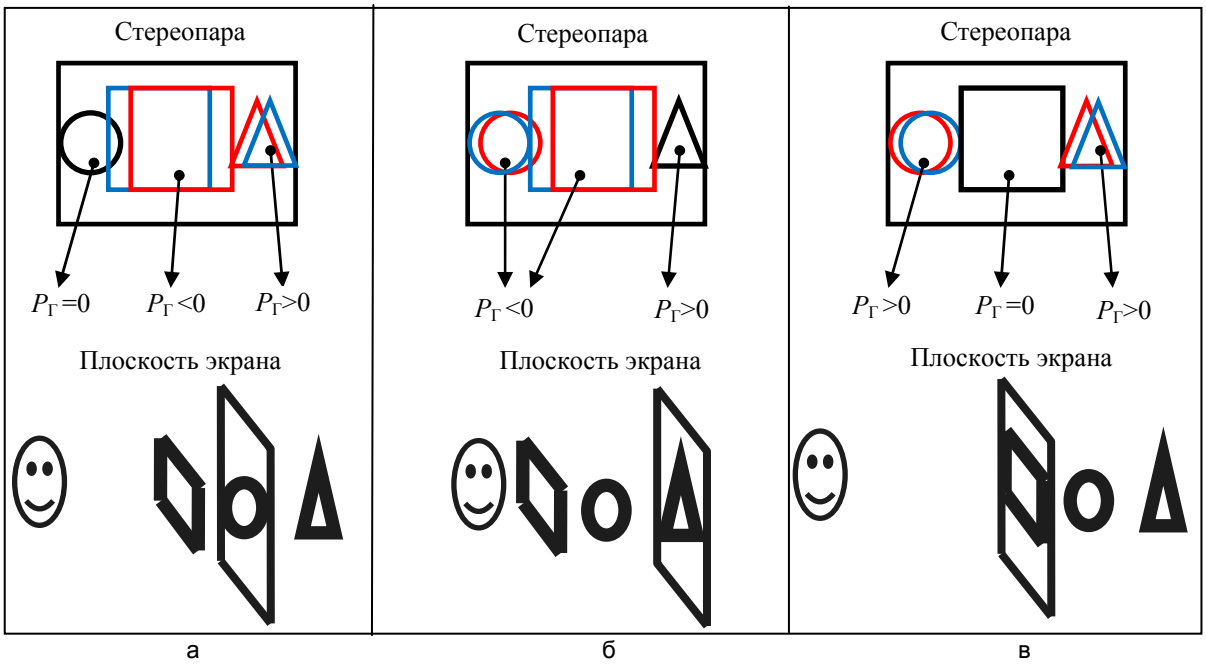


Рис. 2. Зависимость воспринимаемого пространственного расположения объектов в стереоизображении от значения величины горизонтальных параллаксов: параллакс стереопары имеет положительное, отрицательное и нулевое значения (а); параллакс стереопары имеет отрицательное и нулевое значения (б); параллакс стереопары имеет положительное и нулевое значения (в)

Величина максимально допустимого отрицательного горизонтального параллакса (P^-_{MAX}) на экране определяется из следующего выражения [6]:

$$\alpha = 2\,\mathrm{arctg}\frac{\left|P^-_{MAX}\right| + B_{ЗР}}{2L},$$

где α – угол конвергенции, предельное значение которого составляет 30º; $B_{ЗР}$ – базис зрения, мм; L – расстояние от экрана до глаз зрителя, мм. Так, например, при значении L = 1250 мм и P^{-}_{MAX} = 600 мм, а также при значении L = 1500 мм и P^{-}_{MAX} = 740 мм угол α = 30º. Величина α уменьшается с увеличением L и увеличивается при возрастании P^{-}_{MAX}. При значении L = 1500 мм и P^{-}_{MAX} = 600 мм α = 25º.

При совмещении изображений стереопары возникает проблема точного задания необходимого значения параллаксов. Во многих имеющихся на сегодняшний день программах по обработке и совмещению кадров стереопары задание величины параллаксов осуществляется весьма приблизительно (вручную, на глаз) путем перемещения ползунков курсором. При этом точность совмещения изображений невелика, поскольку во многом зависит от зрительного восприятия человека, формирующего стереопару. При неточном совмещении левого и правого кадров могут возникнуть нежелательные параллаксы, величина которых увеличивается пропорционально увеличению размеров печатной фотографии или размеров экрана, на который проецируется полученное изображение.

Известны компьютерные программы, например, «StereoPhoto Maker» (URL: http://stereo.jpn.org/eng/stphmkr), в которых существует возможность совмещения изображений стереопары в автоматическом режиме. Однако при этом довольно редко удается достичь желаемого результата, так как автоматическое совмещение изображений в данных программах происходит чаще всего по объекту, расположенному в центре кадра на переднем плане снимаемой сцены. А значит, все изображаемые объекты на сформированной стереопаре будут восприниматься только в заэкранном пространстве, что ухудшит зрелищность рассматриваемого стереоизображения.

Таким образом, существующие программные продукты не позволяют в автоматическом режиме и с высокой точностью выполнить совмещение изображений стереопары по желаемому объекту, а также задать определенное значение величины параллаксов. Кроме того, большинство подобных программ не дает возможности измерить величины параллаксов и определить, превысят ли они свои предельные значения при последующем просмотре стереоизображения.

В настоящей работе предлагается описание разработанных авторами методов формирования изображений стереопары, которые позволяют выделить на изображении интересующий объект, выставить для него нулевой вертикальный параллакс и необходимое значение горизонтального параллакса и, таким образом, совместить стереопару, задав данному объекту определенное пространственное расположение в воспринимаемом стереоизображении. Кроме того, разработанные методы позволяют осуществить цифровую конвергенцию изображений стереопары [7], измерить горизонтальный и вертикальный параллаксы стереопары в пикселах и (или) процентах от общего размера изображения, а также определить, не превысят ли они свои допустимые пределы при печати или проекции.

Далее изложены два предлагаемых метода формирования стереопары. Первый основан на нахождении максимальной корреляции между градиентными изображениями левого и правого кадров, второй – на нахождении сдвига между двумя сопряженными ключевыми точками изображений стереопары, обнаруженными при помощи детектора точечных особенностей.

Метод формирования стереопары, основанный на корреляции изображений

Рассмотрим принцип действия метода формирования стереопары, основанного на корреляции изображений, на примере совмещения левого и правого кадров стереопары (рис. 3) по одному из изображенных объектов (статуэтки ангела).

а б

Рис. 3. Исходные кадры стереопары: левый (а); правый (б)

Первым действием задаем требуемое значение горизонтального параллакса для нужного нам объекта. Для данного примера зададим горизонтальный параллакс, равный нулю. Затем красная составляющая левого кадра стереопары, а также зеленая и синяя составляющие правого кадра стереопары объединяются, и на экран выводится анаглифное изображение. Дважды кликнув курсором, задаем координаты двух точек (рис. 4), выделяя тем самым интересующий объект. Далее образуются градиенты левого и правого изображений выделенного объекта (рис. 5).

Рис. 4. Выделение объекта, для которого в дальнейшем будут выставлены заданные значения параллаксов

Градиентом двумерной функции $f(x,y)$, где x, y – пространственные координаты, а амплитуда f – уровень яркости изображения в этой точке, называется вектор, модуль (длина) которого равен

$$\left|\nabla f\right| = mag(\nabla f) = \left|G_x^2 + G_y^2\right|^{1/2} = \left[(\partial f / \partial x)^2 + (\partial f / \partial y)^2\right]^{1/2}.$$

Данную величину аппроксимируют с помощью суммы абсолютных величин

$$\left|\nabla f\right| \approx \left|G_x\right| + \left|G_y\right|.$$

Такое приближение равно нулю на областях с постоянной яркостью (цветом) пикселов, и его величина пропорциональна степени изменения яркости на неоднородных областях [8].

После этого корреляционным способом решается задача нахождения позиции на градиенте правого изображения выделенного объекта (рис. 5, б), которая максимально соответствует градиенту левого изображения выделенного объекта (рис. 5, а). Эта позиция является точкой максимума результирующей матрицы корреляции (рис. 6).

а б

Рис. 5. Градиенты изображений выделенного объекта: левого изображения (а); правого изображения (б)

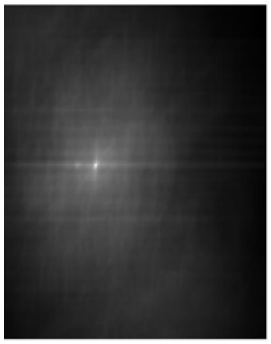
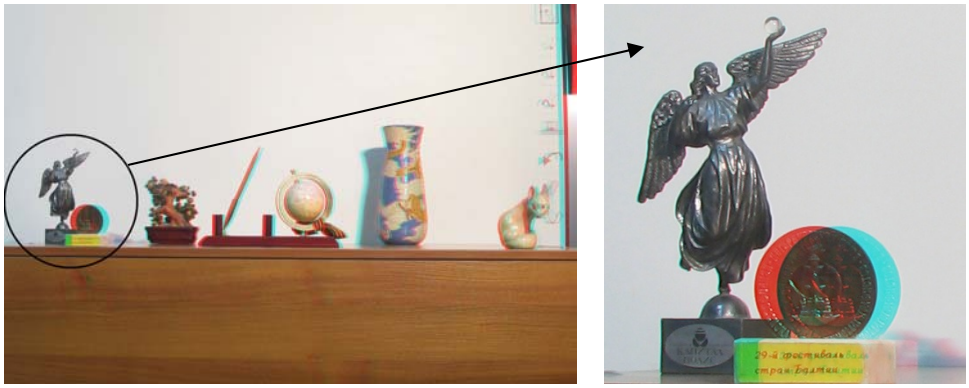


Рис. 6. Изображение матрицы, полученной в результате корреляции

С учетом заданного ранее значения горизонтального параллакса, нулевого вертикального параллакса, а также на основе значений координат точки максимума в матрице корреляции определяется величина сдвига одного градиентного изображения относительно другого, а также размер итоговых кадров стереопары.

Красная цветоделенная матрица итогового левого кадра стереопары, а также зеленая и синяя цветоделенные матрицы итогового правого кадра стереопары стыкуются вместе, и, таким образом, формируется изображение, предназначенное для просмотра в анаглифных красно-голубых очках (рис. 7). На данном рисунке показано, как точно совмещены левый и правый кадры стереопары. Вертикальный и горизонтальный параллаксы выбранного объекта (статуэтка ангела), измеренные в пикселах, равны нулю.



а б

Рис. 7. Анаглифное изображение, сформированное методом, который основан на корреляции изображений (а), и его фрагмент (б)

В работе использован программный пакет MATLAB R2013b, процессор Intel(R) Core™ i3-2330M с тактовой частотой 2,20 ГГц и ОЗУ 3,00 ГБ, 64-разрядная операционная система Windows 7.

Метод формирования стереопары, основанный на применении детектора точечных особенностей

Рассмотрим принцип метода формирования стереопары, основанного на применении детектора точечных особенностей, на примере совмещения тех же левого и правого кадров стереопары (рис. 3) по выбранному объекту – статуэтке ангела. Как и в предыдущем методе, вначале задаем требуемую величину горизонтального параллакса для нужного объекта. На одном из изображений, например, левом, при помощи курсора выделяем интересующий объект (рис. 8).

Рис. 8. Выделение объекта, для которого в дальнейшем будут выставлены заданные значения параллаксов

После того как был выделен объект, формируется бинарное изображение с единицами внутри интересующей области и с нулями – вне нее. Данное бинарное изображение перемножается поэлементно с матрицей левого полутонового изображения, и, таким образом, образуется полутоновое изображение интересующего нас объекта.

При помощи детектора точечных особенностей на полученном изображении и на правом полутоновом изображении определяются соответствующие друг другу ключевые точки (рис. 9).

Для обнаружения сопряженных ключевых точек существуют различные детекторы. Наиболее удобными представляются детекторы FAST (Features from Accelerated Segment Test [10, 11]) и SURF (Speeded Up Robust Features [12, 13]). Однако при использовании алгоритма SURF зачастую обнаружи-

ваются нежелательные посторонние точки, и чтобы их исключить, применяется метод оценки параметров модели на основе случайных выборок RANSAC (RANdom SAmple Consensus [14, 15]).

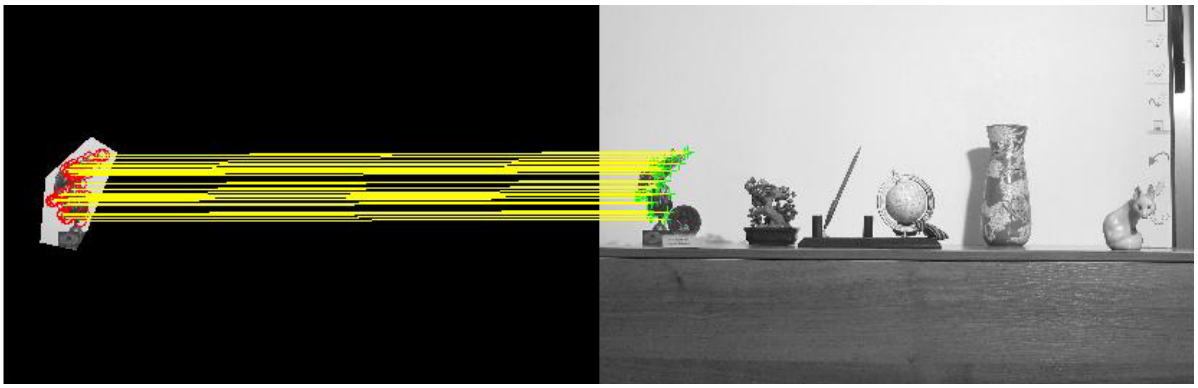
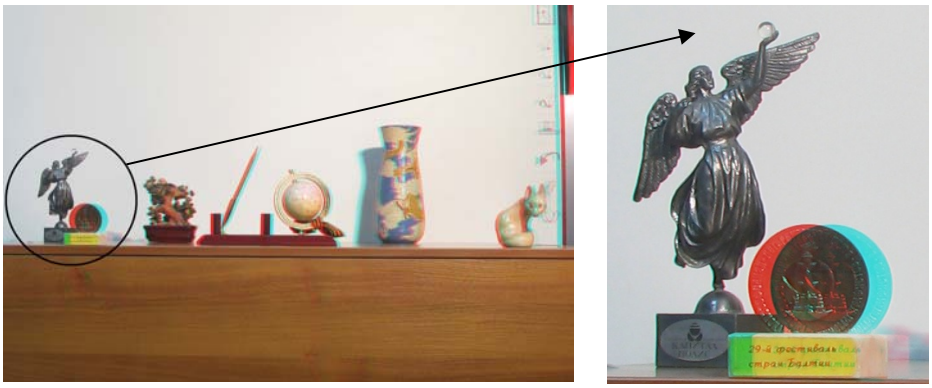


Рис. 9. Ключевые точки полутонового изображения выделенного объекта на левом кадре, сопряженные с ключевыми точками полутонового изображения правого кадра стереопары

С использованием значений координат пары сопряженных точек, отобранных по значению медианы параллаксов, и на основании заданной вначале величины горизонтального параллакса вычисляются величина сдвига одного изображения относительно другого, а затем и размер итоговых кадров стереопары.

Подобно действиям в предыдущем методе, происходит формирование анаглифного изображения, предназначенного для просмотра в красно-голубых очках (рис. 10). На данном рисунке видно, что изображения совмещены с высокой точностью, обусловленной тем, что вертикальный и горизонтальный параллаксы выбранного объекта (статуэтка ангела), выраженные в пикселах, равны нулю.



а б

Рис. 10. Анаглифное изображение, сформированное методом, который основан на применении детектора точечных особенностей (а), и его фрагмент (б)

Заключение

Предложенные в работе методы позволяют совмещать изображения стереопары по выбранному объекту с высокой точностью и формировать стереоизображения с любым пространственным расположением. Разработанные методы позволяют измерять горизонтальный и вертикальный параллаксы уже сформированной стереопары и определять, не превысят ли они свои допустимые пределы при печати или проекции. Таким образом, применением на практике представленных методов можно обеспечить необходимые требования к величине параллаксов с целью формирования стереоизображения, комфортного для восприятия.

Методы реализуются практически полностью автоматически, необходимо только выделить объект, для которого в дальнейшем будет выставлено заданное значение горизонтального параллакса. Совмещение изображений стереопары методом ключевых точек происходит менее чем за одну секунду. Метод с использованием корреляции требует чуть больше времени, однако позволяет контролировать и совмещать неразделенное анаглифное изображение.

К возможным областям применения предложенных методов формирования стереопары относится съемка и обработка стереофотографий, стереокинематограф. Представленные методы можно использовать в программах по монтажу и обработке изображений стереопары, в видеоконтрольных устройствах съемочных камер или в устройствах оценки качества видеопоследовательности (применительно к выборочным кадрам), исключая субъективное оценивание.

Литература

1. Мелкумов А.С. Инструментарий для малобюджетной стереосъемки // Мир техники кино. 2011. № 22. С. 25–32.
2. Комар В.Г., Рожков С.Н., Чекалин Д.А. Необходимость нормирования параметров стереопары и стереопроекции с целью снижения зрительного дискомфорта в условиях кинозала // Мир техники кино. 2012. № 24. С. 31–44.
3. Елхов В.А., Кондратьев Н.В., Овечкис Ю.Н., Паутова Л.В. Анализ параметров многообъективной съёмки в системе безочкового кинопоказа многоракурсных стереоизображений // Мир техники кино. 2010. № 17. С. 2–7.
4. Рожков С.Н., Овсянникова Н.А. Стереоскопия в кино-, фото-, видеотехнике. Терминологический словарь. М.: Парадиз, 2003. 138 с.
5. Мелкумов А.С. Основы стереографии // Мир техники кино. 2010. № 18. С. 30–38.
6. Рожков С.Н. Особенности восприятия стереоизображения в кинозале // Мир техники кино. 2008. № 10. С. 10–15.
7. Газеева И.В., Тихомирова Г.В., Чафонова В.Г. Алгоритмы цифровой конвергенции изображений стереопары // Мир техники кино. 2014. № 1 (31). С. 10–17.
8. Гонсалес Р., Вудс Р., Эддинс С. Цифровая обработка изображений в среде MATLAB. М.: Техносфера, 2006. 616 с.
9. Гонсалес Р., Вудс Р. Цифровая обработка изображений. М.: Техносфера, 2005. 1072 с.
10. Rosten E., Drummond T. Machine learning for high-speed corner detection // Proc. 9th European Conference on Computer Vision ECCV 2006. Graz, Austria, 2006. V. 3951 LNCS. P. 430–443.
11. Rosten E., Porter R., Drummond T. Faster and better: a machine learning approach to corner detection // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010. V. 32. N 1. P. 105–119.
12. Bay H., Ess A., Tuytelaars T., Van Gool L. Speeded up robust features (SURF) // Computer Vision and Image Understanding. 2008. V. 110. N 3. P. 346–359.
13. Rosten E., Drummond T. Fusing points and lines for high performance tracking // Proc. 10th IEEE International Conference on Computer Vision. 2005. V. 2. P. 1508–1515.
14. Волкович А.Н., Жук Д.В., Тузиков А.В. Восстановление трехмерных моделей объектов по стереоизображениям с учетом распараллеливания // Научно-технический вестник СПбГУ ИТМО. 2008. № 58. С. 3–10.
15. Гошин Е., Фурсов В.А. Метод согласованной идентификации в задаче определения соответственных точек на изображениях // Компьютерная оптика. 2012. Т. 36. №1. С. 131–135.

Чафонова Виктория Германовна – аспирант, Санкт-Петербургский государственный университет кино и телевидения, Санкт-Петербург, 191119, Российская Федерация, vi777@nextmail.ru

Газеева Ирина Варисовна – кандидат технических наук, доцент, доцент кафедры, Санкт-Петербургский государственный университет кино и телевидения, Санкт-Петербург, 191119, Российская Федерация, igazeeva@mail.ru

Viktoriya G. Chafonova – postgraduate, Saint Petersburg State University of Film and Television, Saint Petersburg, 191119, Russian Federation, vi777@nextmail.ru

Irina V. Gazeeva – PhD, Associate professor, Associate professor, Saint Petersburg State University of Film and Television, Saint Petersburg, 191119, Russian Federation, igazeeva@mail.ru

Принято к печати 02.06.14
Accepted 02.06.14

УДК 681.786

МЕТОД ОПРЕДЕЛЕНИЯ ПРОСТРАНСТВЕННЫХ КООРДИНАТ В АКТИВНОЙ СТЕРЕОСКОПИЧЕСКОЙ СИСТЕМЕ

В.В. Коротаев[a], Т.С. Джамийков[b], Х.В. Нгуен[a], С.Н. Ярышев[a]

[a]Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, korotaev@grv.ifmo.ru

[b]Технический университет - София, София, 1000, Болгария

Аннотация. Предложена структурная схема активной стереоскопической системы и алгоритм ее работы, обеспечивающий быстрое вычисление пространственных координат. Система включает в себя две одинаковые камеры, образующие стереопару, и лазерный сканер, который осуществляет вертикальное сканирование лазерным лучом пространства перед системой. Синхронная работа двух камер обеспечивается с помощью отдельного блока-синхронизатора. Разработанный алгоритм работы системы реализован в среде MATLAB. В предложенном алгоритме влияние фоновой засветки устраняется за счет межкадровой обработки. Алгоритм базируется на предварительном вычислении координат эпиполярных линий и сопряженных точек стереоскопического изображения. Эти данные используются для быстрого вычисления трехмерных координат точек, составляющих трехмерные изображения объектов. Приведено описание эксперимента на физической модели. Результаты эксперимента подтверждают работоспособность предложенной активной стереоскопической системы и алгоритма ее работы. Предложенная авторами схема активной стереоскопической системы и метод вычисления пространственных координат могут быть рекомендованы для создания стереоскопических систем, работающих в реальном времени и требующих повышенного быстродействия: устройства распознавания лиц, системы контроля положения железнодорожного пути, активные системы безопасности автомобиля.

Ключевые слова: стереоскопическая система, обработка изображения, трехмерное изображение, лазерный сканер, сопряженные точки, эпиполярные линии.

Благодарности. Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

METHOD FOR DETERMINING THE SPATIAL COORDINATES IN THE ACTIVE STEREOSCOPIC SYSTEM

V.V. Korotaev[a],T.S. Djamiykov, H.V. Nguyen[a], S.N. Yaryshev[a]

[a]ITMO University, Saint Petersburg, 197101, Russian Federation, korotaev@grv.ifmo.ru

[b] Technical University of Sofia, Sofia, 1000, Bulgaria

Abstract. The paper deals with the structural scheme of active stereoscopic system and algorithm of its operation, providing the fast calculation of the spatial coordinates. The system includes two identical cameras, forming a stereo pair, and a laser scanner, which provides vertical scanning of the space before the system by the laser beam. A separate synchronizer provides synchronous operation of the two cameras. The developed algorithm of the system operation is implemented in MATLAB. In the proposed algorithm, the influence of background light is eliminated by interframe processing. The algorithm is based on precomputation of coordinates for epipolar lines and corresponding points in stereoscopic image. These data are used to quick calculation of the three-dimensional coordinates of points that form the three-dimensional images of objects. Experiment description on a physical model is given. Experimental results confirm the efficiency of the proposed active stereoscopic system and its operation algorithm. The proposed scheme of active stereoscopic system and calculating method for the spatial coordinates can be recommended for creation of stereoscopic systems, operating in real time and at high processing speed: devices for face recognition, systems for the position control of railway track, automobile active safety systems.

Keywords: stereoscopic system, image processing, three-dimensional image, laser scanner, corresponding points, epipolar lines.

Acknowledgements. The work was partially financially supported by the Government of the Russian Federation (grant 074-U01).

Введение

В настоящее время 3D-сканирование активно используется в различных отраслях науки и техники. Существует множество моделей 3D-сканеров [1–4], каждая из которых обладает своими возможностями и особенностями и находит применение в той или иной области. Стереоскопическая система реализует один из простых методов получения трехмерных моделей объектов. Для определения трехмерных координат на каждом изображении стереопары необходимо определить сопряженные точки. Для решения этой задачи предложено много методов [5]. Их можно подразделить на глобальные [6–8] и локальные [9–12] методы. Все эти методы требуют больших вычислительных ресурсов и их нельзя применить в быстродействующей системе. Исходя из этого, авторы предлагают новый вид стереоскопической системы – активную стереоскопическую систему, которая свободна от вышеуказанного недостатка.

Теоретические положения

Стереоскопическая система представляет собой оптическую систему, состоящую из двух произвольно ориентированных камер. Каждая камера описывается с помощью матриц $\mathbf{A_1}$ и $\mathbf{A_2}$, известных под названием матриц внутренних параметров камер. Они содержат только параметры оптических систем и фотоприемников камер. Параметры стереоскопической системы описываются вектором переноса **t** и мат-

рицей поворота **R**. Если известны координаты изображений в плоскостях изображений камер v' и v'', координаты объекта в пространстве $M' = (X',Y',Z')$ относительно первой камеры и $M'' = (X'',Y'',Z'')$ относительно второй камеры можно найти с помощью следующих формул:

$$\begin{bmatrix} Z' \\ Z'' \end{bmatrix} = \begin{bmatrix} v'^T A_1^{-T} A_1^{-1} v' & -v'^T A_1^{-T} R^T A_2^{-1} v'' \\ -v'^T A_1^{-T} R^T A_2^{-1} v'' & v''^T A_2^{-T} A_2^{-1} v'' \end{bmatrix}^{-1} \begin{bmatrix} v'^T A_1^{-T} R^T \\ v''^T A_2^{-T} \end{bmatrix} t, \quad (1)$$

$$M' = Z' A_1^{-1} v', M'' = Z'' A_2^{-1} v''.$$

Как указано выше, для получения координат объекта в пространстве необходимо найти сопряженные точки в стереопаре. В стереоскопической системе есть одна особенность. Если задавать координаты точки на одном изображении стереопары, то ее сопряженная точка на другом изображении стереопары должна находиться на одной линии, которая называется эпиполярной линией [13]. Функция эпиполярной линии имеет следующий вид:

$$a''^T v'' = 0, \quad (2)$$

где вектор коэффициентов $a'' = Fv'$, а $F = A_2^{-T} [t]_x R A_1^{-1}$. Принимая во внимание эту особенность, авторы добавили к стереоскопической системе лазерный сканер. Лазерный сканер формирует вертикальную лазерную линию в пространстве и поворачивается вокруг своей оси, обеспечивая при этом сканирование пространства перед камерами. На первом изображении стереопары по строкам определяется энергетический центр тяжести лазерной линии, и с помощью формулы (2) на втором изображении стереопары определяется эпиполярная линия. Пересечение эпиполярной линии с лазерной линией на втором изображении является сопряженной точкой, которая соответствует точке на первом изображении. С помощью формулы (1) можно определить координаты данной точки в пространстве.

Структурная схема активной стереоскопической системы

На основе рассмотренных теоретических положений предлагается структурная схема активной стереоскопической системы (рис. 1).

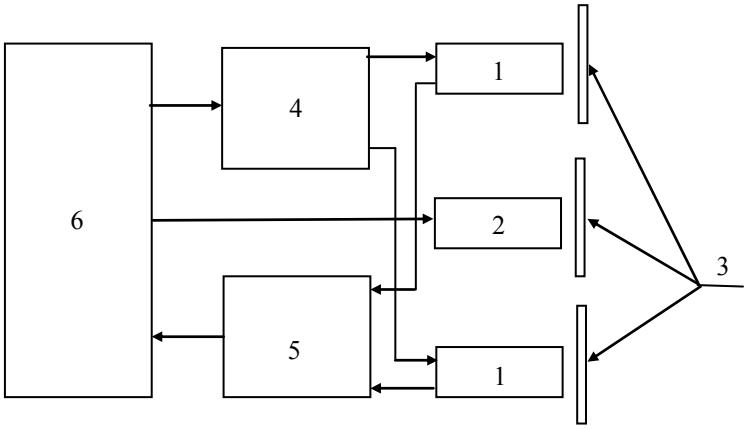


Рис. 1. Структурная схема активной стереоскопической системы: 1 – видеокамера; 2 – лазерный сканер; 3 – узкополосный оптический фильтр; 4 – синхронизатор; 5 – блок предварительной обработки изображения; 6 – ЭВМ

Система включает в себя две одинаковые камеры 1 и лазерный сканер 2. Перед камерами и лазером установлены узкополосные оптические фильтры 3. Средняя длина волны пропускания этих фильтров соответствует длине волны лазерного излучения сканера. Поэтому в камеру поступает только оптический сигнал, полученный в результате отражения лазерного луча от объекта. В результате спектральной селекции упрощается процесс распознавания лазерного луча на изображениях стереопары. Лазерный сканер имеет шаговый двигатель, которым управляет ЭВМ 6. Этот компьютер также управляет синхронизатором 4, который обеспечивает синхронную работу видеокамер. Сигнал с камер поступает в блок предварительной обработки 5. В нем происходит процесс распознавания лазерного луча на каждом изображении. При получении каждой строки кадра с обеих камер блок предварительной обработки сравнивает ее с такой же строкой предыдущего кадра. Разница между ними является зоной, где находится лазерный луч. По информации о распределении энергии отраженного луча в данной зоне можно построить график распределения этой энергии и найти точку максимального значения энергии. Эта точка является энергетическим центром тяжести отраженного лазерного луча в данной строке. Метод дает возможность определения энергетического центра тяжести лазерного луча с точностью до 0,1 пикселя. После считывания информации со всех строк данного кадра и получения координат энергетического центра тяжести отраженного излучения для всех строк блок предварительной обработки передает координаты в компьютер. Во время получения следующей пары кадров с камер в компьютере происходит определение трехмерных

координат всех точек предыдущего кадра. В памяти компьютера уже сохранены все эпиполярные линии для всех точек на первом изображении стереопары и все пространственные координаты всех точек на этих эпиполярных линиях. После получения информации с блока предварительной обработки компьютер сканирует все энергетические центры тяжести на первом изображении стереопары. Для каждого энергетического центра тяжести компьютер выбирает из памяти соответствующую эпиполярную линию и находит пересечение этой линии с линией, образуемой энергетическими центрами тяжести второго изображения. Получив точку пересечения двух линий, ЭВМ выбирает из памяти соответствующие пространственные координаты этой точки. Представленный алгоритм обработки дает возможность получения трехмерных координат объектов в реальном времени.

Экспериментальное исследование

Для проверки работоспособности системы авторы построили и исследовали физическую модель (рис. 2), а также разработали соответствующее программное обеспечение в среде MATLAB. Модель включает в себя две камеры 1 марки Microsoft Lifecam HD-3000 с разрешением 1280×720 пикселей. Перед каждой камерой установлен узкополосный фильтр 3, настроенный на длину волны лазера 650 нм. Между камерами установлен лазерный блок, включающий в себя лазерный источник 2 с длиной волны 650 нм, управляемый шаговым двигателем 4. Шаговый двигатель подключен к блоку управления 5. Камеры и лазерный блок подключаются к персональной ЭВМ.

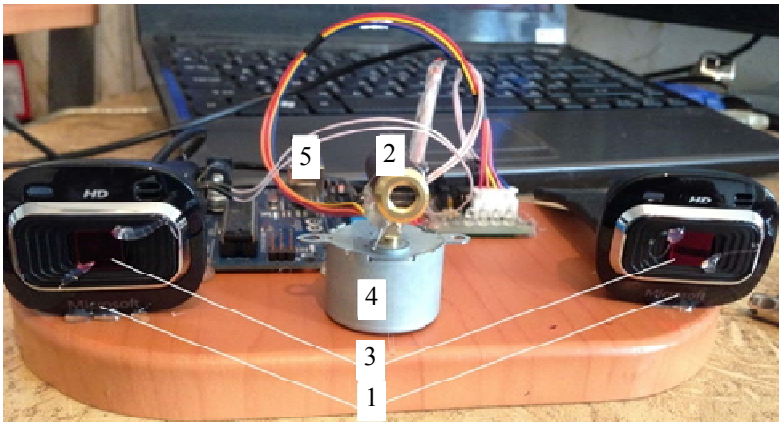


Рис. 2. Физическая модель активной стереоскопической системы: 1 – видеокамера; 2 – источник лазерного излучения; 3 – узкополосный фильтр; 4 – шаговый двигатель; 5 – блок управления шагового двигателя

На этапе предварительной обработки изображения (рис. 3) выделяется изображение отраженного лазерного излучения от объекта в условиях фоновой засветки. В настоящей работе авторами применена межкадровая фильтрация. Вследствие того, что скорость получения информации в оптических каналах значительно больше, чем скорость движения, в двух соседних кадрах перемещается только лазерный луч. Чтобы отделить изображение отраженного лазерного луча от фона, значение яркости каждого пикселя в текущем кадре вычитается из значения яркости того же пикселя предыдущего кадра. Если полученная разность меньше нуля, конечный результат обнуляется. На рис. 3, а, показано исходное изображение, полученное с левой камеры.

Результат межкадровой фильтрации показан на рис. 3, б. После получения области, где есть изображение отраженного от объекта лазерного луча, на каждой строке находится центр лазерного луча. Для решения этой задачи на каждой строке отыскивается пиксель, яркость которого имеет максимальное значение в данной строке. После получение координат такой точки происходит возврат к исходному изображению и отыскивается ее реальный максимум по яркости, который находится в зоне первого максимума и одновременно в области изображения лазерного луча. Далее находятся все пиксели, которые располагаются вокруг реального максимума и имеют яркость, большую или равную 75% яркости реального максимума. На рис. 3, в, приведены диаграмма яркости строки номер 544, положение реального максимума (красная точка) и зона, где яркость больше или равна 75% яркости реального максимума (зона между пунктирными красными линиями). Используя значения яркости этих пикселей, построим полигон и получим центр лазерного луча на данной строке с точностью до одной десятой размера пикселя. На рис. 3, г, показан вычисленный полигон (зеленая кривая) и центр тяжести лазерного луча на строке номер 544. На этом завершается этап предварительной обработки изображения.

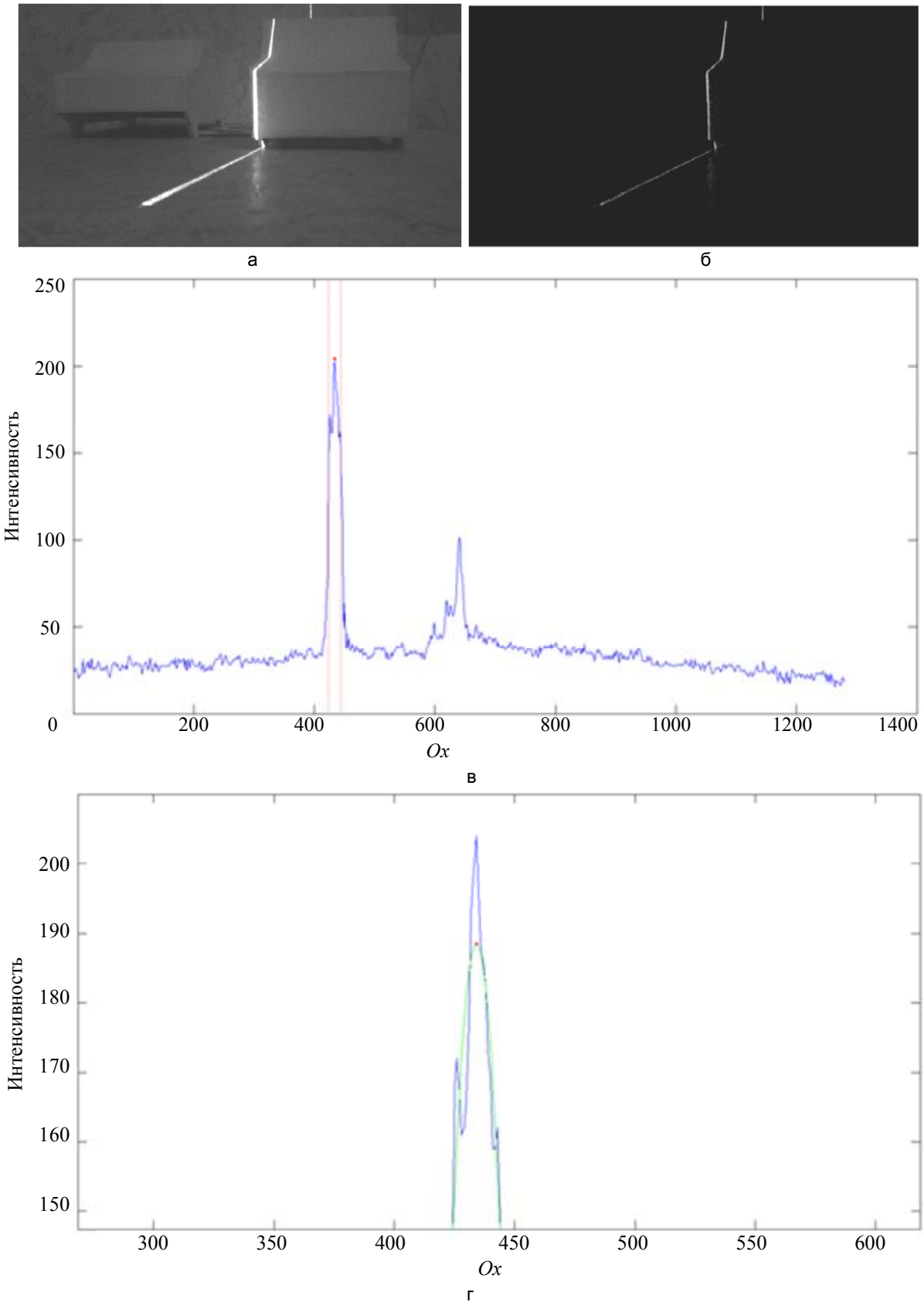


Рис. 3. Результат обработки изображения в предварительном блоке: исходное изображение (а); результат межкадровой обработки (б); диаграмма яркости на строке 544 (в); результат определения центра тяжести на строке 544 (г); *Ox* – номер столбцов

Следующий этап – вычисление трехмерных координат точек. В компьютер поступают координаты всех центров тяжести изображений отраженного лазерного луча для левой и правой камер. В левом изображении по строкам выбираются координаты центра тяжести. Затем из памяти ЭВМ выводятся координаты эпиполярной линии для этого центра тяжести, и определяется пересечение этой линией совокупности центров тяжести правого изображения. Это пересечение является сопряженной точкой центра тяжести левого изображения. Используя координаты этих двух точек, выводим из памяти трехмерные координаты точек объекта. На рис. 4, а, показано левое изображение с набором центров тяжести по строкам. Точка, находящаяся в середине горизонтальной линии, соответствует сопряженной точке, и ее необходимо найти в правом изображении. На рис. 4, б, показаны эпиполярная линия и ее пересечение с совокупностью центров тяжести правого изображения. На рис. 4, в, показан конечный результат построения трехмерного изображения сцены активной стереоскопической системой.

Отметим, что предложенный метод дает не только изображение сцены, но позволяет получать информацию о размерах объектов практически в реальном масштабе времени, т.е. появляется возможность установки предложенной стереоскопической системы, к примеру, на подвижных объектах.

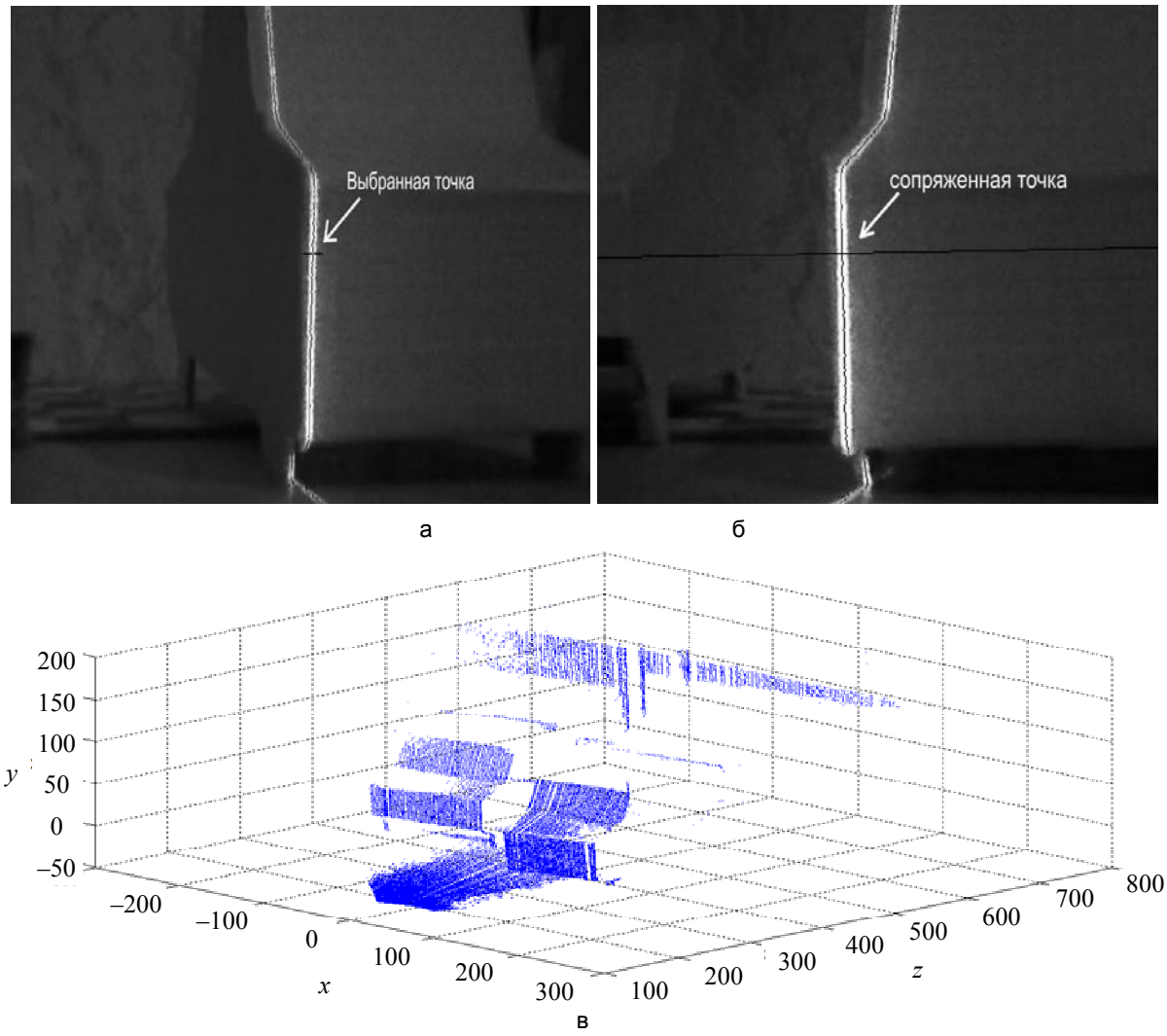


Рис. 4. Результат восстановления трехмерного изображения сцены, левое изображение с выбранной точкой (а); правое изображение с эпиполярной линией и сопряженной точкой (б); трехмерное изображение сцены (в)

Заключение

В работе представлена активная стереоскопическая система с лазерным сканирующим устройством, алгоритм обработки изображения и получения трехмерного изображения сцены. Благодаря лазерному сканирующему устройству активная стереоскопическая система решает основную задачу автоматической стереоскопической системы – нахождение сопряженных точек в стереопаре, у которой нет общего алгоритма решения. Для проверки работоспособности метода авторы реализовали его алгоритм в среде MATLAB. Благодаря возможности получения информации о реальных размерах объектов и их координа-

тах удается построить системы, которые решают большое количество прикладных задач, таких как распознавание лиц [5], контроль положения железнодорожного пути [14], реализовать активную систему безопасности автомобиля.

Литература

1. Мачихин А.С., Колючкин В.Я., Тимашова Л.Н. Однокамерная сканирующая стереоскопическая система для реконструкции трехмерной структуры объектов // Научно-технический вестник СПбГУ ИТМО. 2007. № 38. С. 142–146.
2. Зуев В.Е. Дистанционное оптическое зондирование атмосферы. СПб.: Гидрометеоиздат, 1992. 232 с.
3. Каталог 3D-сканеров с подробным описанием и ценой [Электронный ресурс]. Режим доступа: http://www.foto-business.ru/3D-skaneryi, свободный. Яз. рус. (дата обращения 01.07.2014).
4. Климанов М.М. Лазерная триангуляционная измерительная система // Материалы XI научной конференции МГТУ «Станкин» по математическому моделированию и информатике. М.: МГТУ «Станкин», 2008. С. 230–232.
5. Пономарев С.В. Методика сравнения алгоритмов стереозрения при восстановлении трехмерной модели лица человека // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 6 (88). С. 40–45.
6. Gutierrez S., Marroquin J.L. Robust approach for disparity estimation in stereo vision // Image and Vision Computing. 2004. V. 22. N 3. P. 183–195.
7. Bleyer M., Gelautz M. A layered stereo matching algorithm using image segmentation and global visibility constraints // ISPRS Journal of Photogrammetry and Remote Sensing. 2005. V. 59. N 3. P. 128–150.
8. Kim H., Sohn K. 3D reconstruction from stereo images for interactions between real and virtual objects // Signal Processing: Image Communication. 2005. V. 20. N 1. P. 61–75.
9. Stefano L.D., Marchionni M., Mattocia S. A fast area-based stereo matching algorithm // Image and Vision Computing. 2004. V. 22. N 12. P. 983–1005.
10. Binaghi E., Gallo I., Marino G., Raspanti M. Neural adaptive stereo matching // Pattern Regonition Letters. 2004. V. 25. N 15. P. 1743–1758.
11. Ogale A.S., Aloimonos Y. Shape and the stereo correspondence problem // International Journal of Computer Vision. 2005. V. 65. N 3. P. 147–162.
12. Yoon S., Park S.-K., Kang S., Kwak Y.K. Fast correlation-based stereo matching with the reduction of systematic errors // Pattern Recognition Letters. 2005. V. 26. N 14. P. 2221–2231.
13. Грузман И.С., Киричук В.С., Косых В.П., Перетягин Г.И., Спектор А.А. Цифровая обработка изображений в информационных системах: Учебное пособие. Новосибирск: НГТУ, 2000. 168 с.
14. Араканцев К.Г., Горбачёв А.А., Серикова М.Г. Стереоскопическая система контроля фактического положения железнодорожного пути // Изв. вузов. Приборостроение. 2013. Т. 56. № 5. Р. 34–39.

Коротаев Валерий Викторович – доктор технических наук, профессор, декан факультета, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, korotaev@grv.ifmo.ru

Джамийков Тодор Стоянов – доцент, Технический Университет - София, София, 1000, Болгария, tsd@tu-sofia.bg

Нгуен Хоанг Вьет – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ngvietvn@gmail.com

Ярышев Сергей Николаевич – кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ysn63@mail.ru

Valery V. Korotaev – D.Sc., Professor, Department head, ITMO University, Saint Petersburg, 197101, Russian Federation, korotaev@grv.ifmo.ru

Todor S. Djamiykov – PhD, Associate professor, Technical University of Sofia, Sofia, 1000, Bulgaria, tsd@tu-sofia.bg

Hoang Viet Nguyen – postgraduate, ITMO University, Saint Petersburg, 197101,Russian Federation, ngvietvn@gmail.com

Sergei N. Yaryshev – PhD, Associate professor, Associate professor, ITMO University, Saint Petersburg, 197101, Russian Federation, ysn63@mail.ru

Принято к печати 16.10.14
Accepted 16.10.14

3 АВТОМАТИЧЕСКОЕ УПРАВЛЕНИЕ И РОБОТОТЕХНИКА AUTOMATIC CONTROL AND ROBOTICS

УДК 681.51

СИСТЕМА УПРАВЛЕНИЯ БЕСПИЛОТНЫМ ЛЕТАТЕЛЬНЫМ АППАРАТОМ, ОСНАЩЕННЫМ РОБОТОТЕХНИЧЕСКИМ МАНИПУЛЯТОРОМ

А.А. Маргун[a], К.А. Зименко[a], Д.Н. Базылев[a], А.А. Бобцов[a], А.С. Кремлев[a], Д.Д. Ибраев[a], М. Чех[b]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, alexeimargun@gmail.ru

[b] Университет Западной Богемии, Пльзень, 306 14, Чехия

Аннотация. Рассматривается задача синтеза системы управления для мультиротационного беспилотного летательного аппарата, оснащенного робототехническим манипулятором. Предложен алгоритм управления, основанный на методе линеаризации обратной связью и синтезе пропорционально-дифференциального регулятора с учетом изменений тензора инерции, положения центра масс и компенсацией реактивного момента сил, порождаемого динамикой манипулятора. В качестве модели рассматриваемого объекта управления выбран квадрокоптер с плоским двухзвенным манипулятором. На основании законов механики Ньютона и уравнений Эйлера–Лагранжа получены системы уравнений, описывающие поведение рассматриваемой динамической системы. Предложены выражения, определяющие тензор инерции и положение центра масс системы в зависимости от текущего положения манипулятора, а также реактивный момент сил, действующих на квадрокоптер со стороны манипулятора. Для полученной нелинейной системы с перекрестными связями применена линеаризация обратной связью с компенсацией влияния манипулятора на квадрокоптер, в результате чего уравнения динамики робота были преобразованы к линейной стационарной системе. Управление преобразованной системой осуществлено с помощью пропорционально-дифференциального регулятора. Проведено моделирование рассматриваемой системы с описанным в работе методом управления и классическим методом на базе пропорционально-дифференциального регулятора. Результаты моделирования показали, что предложенный подход позволяет достигнуть более высоких показателей точности и эффективности при движении по заданной траектории, чем управление с помощью пропорционально-дифференциального регулятора.

Ключевые слова: квадрокоптер, манипулятор, БПЛА, система управления.

Благодарности. Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01) и при финансовой поддержке Минобрнауки Российской Федерации, Договор 14.Z50.31.003.

CONTROL SYSTEM FOR UNMANNED AIRCRAFT EQUIPPED WITH ROBOTICS ARM

A.A. Margun[a], K.A. Zimenko[a], D.N. Bazylev[a], A.A. Bobtsov[a], A.S. Kremlev[a], D.D. Ibraev[a], M. Cech[b]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, alexeimargun@gmail.ru

[b] University of West Bohemia, Pilzen, 306 14, Czech Republic, cechyn@gmail.com

Abstract. The paper deals with the problem of control system synthesis for multi rotational UAV equipped with robotics arm. Control algorithm is proposed based on the method of feedback linearization and synthesis of proportional-differential controller with the real time computation of the inertia tensor and center of mass changes and compensation of the reactive torque generated by the dynamics of the manipulator. Quadrocopter with attached articulated manipulator is selected as a model of the control object. Systems of equations describing the behavior of considered dynamical system are obtained according to the Newton and Euler-Lagrange laws. Expressions are offered, defining the inertia tensor and the position of the system center of mass depending on the current position of the manipulator, and the torque acting on the quadrocopter from the manipulator. Feedback linearization with arm influence compensation on quadrocopter is applied for the resulting nonlinear coupled system. As a result, robot dynamics equations have been converted to a linear stationary system. Converted system control is achieved by a proportional-differential controller. Examined system simulation is done with control method described in the paper and the classical method based on a proportional-differential controller. Simulation results confirm the effectiveness of the proposed approach and demonstrate that the proposed approach provides higher accuracy of the tracking error, than control method by means of proportional-differential regulator.

Keywords: quadrocopter, manipulator, UAV, control system.

Acknowledgements. The work is partially financially supported by the Government of the Russian Federation (grant 074-U01) and the RF Ministry of Education and Science, agreement 14.Z50.31.003

Введение

На сегодняшний день в разных странах по всему миру мультиротационные летательные аппараты активно используются для решения широкого круга задач, таких как проведение спасательных операций во время чрезвычайных ситуаций, мониторинг трубопроводов в газовой и нефтяной промышленности, военная разведка и т.д. Многие всемирно известные университеты, научно-исследовательские институты и коммерческие компании занимаются разработкой беспилотных летательных аппаратов (БПЛА) данного типа, где наиболее распространенным типом мультиротационных летательных аппаратов является квадрокоптер.

Так, в последние несколько лет были опубликованы работы, которые посвящены различным задачам, связанным с эксплуатацией квадрокоптеров. В работах [1–4] обсуждались вопросы моделирования динамики. В работе [3] рассмотрена проблема планирования оптимальной по времени траектории, где авторы предложили простой прямой численный метод, основанный на параметризации траектории квадрокоптера и использующий нелинейную оптимизацию с учетом основных ограничений системы и окружающей среды. Управление бэкстеппингом, скользящими режимами, визуальное управление и различные законы нелинейного управления также представлены в работах [2, 5–9].

Помимо этого, так как мультиротационные летательные аппараты обладают повышенной мобильностью, в последние годы большой интерес представляет их внедрение в манипуляцию небольшими грузами. Можно разделить исследования в области решения задач по транспортировке небольших грузов на несколько типов. Во-первых, захват манипуляторов может быть установлен в нижней части воздушного транспортного средства, чтобы переносить полезную нагрузку. В этом случае квадрокоптер, оснащенный захватом, может перевозить блоки и строить конструкции, как в [10, 11]. В работе [12] было реализовано совместное использование нескольких квадрокоптеров со схватами для переноса крупных объектов. Анализ устойчивости данных систем рассмотрен в [13]. Второй тип исследований связан с протяжкой кабеля в линиях электропередач [14–16]. Проблема увеличения функциональности квадрокоптеров со схватом рассмотрена в [17]. Существует также ряд работ, посвященных проблеме стабилизации квадрокоптера с полезной нагрузкой под внешним возмущением, таким как сильный ветер и т.д.

Тем не менее, все перечисленные подходы имеют ряд недостатков, которые включают в себя ограничение на траектории полета квадрокоптера, конструктивные особенности, ограничение на круг выполняемых задач и условия функционирования. Таким образом, определенный тип БПЛА способен выполнять только узкоспециализированные задачи и не имеет гибкости в применении.

Одним из способов, позволяющих сделать БПЛА более универсальным в применении, является комбинированный робот, состоящий из летательного аппарата и роботизированной руки, прикрепленной к нему. Например, в [17] описывается позиционирование робота с картезианским манипулятором. Модель, состоящая из крана и манипулятора с четырьмя степенями свободы, представлена в [18], где кран имитирует траекторию движения летательного аппарата. В [19] формулы Эйлера–Лагранжа применяются для получения динамических уравнений квадрокоптера с несколькими степенями свободы у манипулятора. Можно сделать вывод, что, несмотря на актуальность и привлекательность объекта исследований, очень мало работ посвящено планированию и стабилизации движения БПЛА с манипулятором в виде интегрированной системы.

В настоящей работе предложен способ управления квадрокоптером, оснащенным манипулятором, на основе линеаризации обратной связью и синтеза пропорционально-дифференциального (ПД) регулятора. Также регулятор обеспечивает компенсацию реактивных моментов сил, действующих на квадрокоптер со стороны манипулятора. Измерение смещенного центра масс и вычисление тензора инерции выполняются в реальном времени.

Математическая модель

Уравнения динамики квадрокоптера (рис. 1) как механической системы получены с использованием законов Ньютона и уравнений Эйлера–Лагранжа. Движение квадрокоптера описывается системой из шести нелинейных дифференциальных уравнений:

$$\begin{cases} \ddot{x} = (\sin\psi\sin\varphi + \cos\psi\cos\theta\cos\varphi)\dfrac{U_1}{m} \\ \ddot{y} = (-\cos\psi\sin\varphi + \sin\psi\sin\theta\cos\varphi)\dfrac{U_1}{m} \\ \ddot{z} = -g + (\cos\theta\cos\varphi)\dfrac{U_1}{m} \\ \ddot{\varphi} = \dfrac{I_y - I_z}{I_x}\dot{\theta}\dot{\psi} - \dfrac{U_2}{I_x} \\ \ddot{\theta} = \dfrac{I_z - I_x}{I_y}\dot{\varphi}\dot{\psi} - \dfrac{U_3}{I_y} \\ \ddot{\psi} = \dfrac{I_x - I_y}{I_z}\dot{\varphi}\dot{\theta} - \dfrac{U_4}{I_z} \end{cases},$$

где x, y, z – декартовы координаты квадрокоптера; φ, θ, ψ – углы Эйлера (φ – угол рыскания, θ – угол таганжа, ψ – угол крена); I_x, I_y, I_z – диагональные элементы тензора инерции квадрокоптера; m – масса квадрокоптера; g – ускорение свободного падения; $U = (U_1, U_2, U_3, U_4)$ – виртуальные силы управления, связанные с управляющими силами двигателя уравнениями

$$U_1 = b(\Omega_1^2 + \Omega_2^2 + \Omega_3^2 + \Omega_4^2),$$
$$U_2 = b(-c_1\Omega_2^2 + c_2\Omega_4^2),$$
$$U_3 = b(-c_3\Omega_1^2 + c_4\Omega_3^2),$$
$$U_4 = d(-c_5\Omega_1^2 + c_6\Omega_2^2 - c_7\Omega_3^2 + c_8\Omega_4^2),$$

где $\Omega = (\Omega_1, \Omega_2, \Omega_3, \Omega_4)$ – скорости вращения двигателей (тяга винта пропорциональна квадрату его угловой скорости); b, d – некоторые физические константы, которые можно получить экспериментально, c_i – плечо силы, $i = \overline{1,8}$.

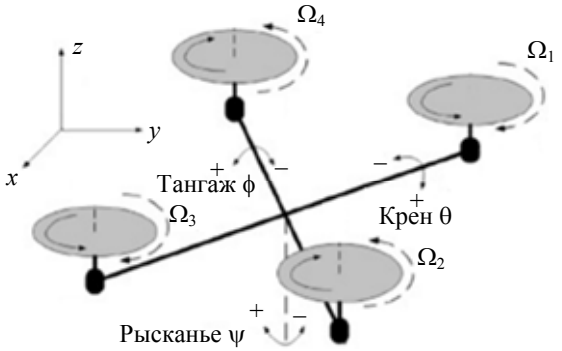


Рис. 1. Модель квадрокоптера

Рассмотрим плоский манипулятор, изображенный на рис. 2. Будем считать, что двухзвенный робот находится в инерциальной системе отчета.

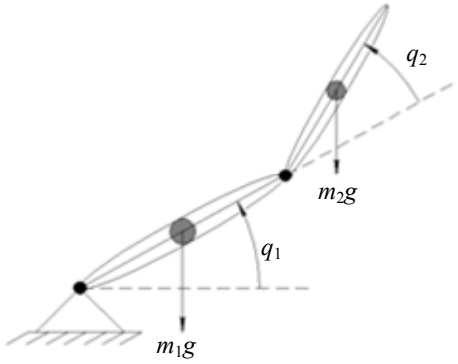


Рис. 2. Плоский двухзвенный манипулятор

Тогда математическая модель манипулятора, полученная с использованием уравнений Эйлера–Лагранжа, может быть представлена в следующем виде:

$$M(q)\ddot{q} + C(q,\dot{q}) + g(q) = \tau, \qquad (1)$$

где $M(q)$ – матрица инерции; $C(q,\dot{q})$ – матрица кориолисовых и центробежных сил; $g(q)$ – матрица потенциальных и гравитационных сил; q – обобщенные координаты в инерциальной системе отсчета; τ – моменты сил двигателей манипулятора.

Предположим, что звенья манипулятора имеют вид тонких однородных стержней с длинами l_1 и l_2 соответственно. Центр масс каждого звена находится в геометрическом центре стержня. Обозначим моменты инерции звеньев I_1 и I_2. Матрицы в уравнении (1) примут вид

$$q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}, \tau = \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix}, g(q) = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix},$$
$$M(q) = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix},$$
$$C(q,\dot{q}) = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$
$$g(q) = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix},$$

где

$$M_{11} = m_1\left(\frac{l_1}{2}\right)^2 + m_2\left(\frac{l_2}{2}\right)^2 + l_1 l_2(1+\cos q_2) + I_1 + I_2,$$

$$M_{12} = M_{21} = m_2\left(\frac{l_2}{2}\right)^2 + \frac{l_1 l_2}{2}\cos q_2 + I_2,$$

$$M_{22} = m_2\left(\frac{l_2}{2}\right)^2 + I_2,$$

$$C_{11} = h\dot{q}_2, C_{12} = h(\dot{q}_1 + \dot{q}_2), C_{21} = -h\dot{q}_1, C_{22} = 0,$$

$$h = -m_2 l_1 l_2 \sin q_2,$$

$$g_1 = \left(m_1\frac{l_1}{2} + m_2 l_1\right) g\cos q_1 + m_2\frac{l_2}{2}\cos(q_1 + q_2),$$

$$g_2 = g m_2\frac{l_2}{2}\cos(q_1 + q_2).$$

Когда манипулятор прикреплен к квадрокоптеру, значения элементов тензора инерции изменяются из-за дополнительной инерции звеньев. Тензор инерции квадрокоптера имеет вид

$$I = \begin{pmatrix} I_x & 0 & 0 \\ 0 & I_y & 0 \\ 0 & 0 & I_z \end{pmatrix},$$

Момент инерции, создаваемый каждой материальной точкой тонкого стержня, равен

$$dm = \frac{m dr}{l},\ dI = r^2 dm = \frac{m r^2 dr}{l},$$

где m – масса стержня; l – его длина; r – расстояние от материальной точки до оси вращения; dm и dl – масса и длина материальной точки соответственно. Используя теорему Штейнера, определим момент инерции каждого звена манипулятора в системе координат квадрокоптера. Для первого звена манипулятора

$$I_{xm1} = \int_0^{l_1\sin q_1} dI = \frac{m_1 l_1^2 \sin^3 q_1}{3},\ I_{ym1} = \int_0^{l_1} dI = \frac{m_1 l_1^2}{3},\ I_{zm1} = \int_0^{l_1\cos q_1} dI = \frac{m_1 l_1^2 \cos^3 q_1}{3};$$

для второго звена

$$I_{xm2} = \int_0^{l_2\sin q_1} dI + m_2(l_1\sin q_1)^2 = \frac{m_2 l_2^2 \sin^3 q_2}{3} + m_2 l_1^2 \sin^2 q_1,$$

$$I_{ym2} = \int_0^{l_2} dI + m_2 l_1^2 = \frac{m_2 l_2^2}{3} + m_2 l_1^2,$$

$$I_{zm2} = \int_0^{l_2\cos q_1} dI + m_2(l_1\cos q_1)^2 = \frac{m_2 l_2^2 \cos^3 q_2}{3} + m_2 l_1^2 \cos^2 q_2.$$

Таким образом, тензор инерции комбинированной системы имеет вид

$$I = \begin{pmatrix} I_{xx} & 0 & 0 \\ 0 & I_{yy} & 0 \\ 0 & 0 & I_{zz} \end{pmatrix},$$

где $I_{xx} = I_x + I_{xm_1} + I_{xm_2}$, $I_{yy} = I_y + I_{ym_1} + I_{ym_2}$, $I_{zz} = I_z + I_{zm_1} + I_{zm_2}$.

Также вычислим изменение положения центра масс комбинированной системы:

$$\Delta x_c = \frac{l_1\cos q_1\left(\frac{m_1}{2} + m_2\right) + m_2\frac{l_2}{2}\cos(q_1 - q_2)}{m_0},$$

$$\Delta y_c = 0,$$

$$\Delta z_c = \frac{l_1\sin q_1\left(\frac{m_1}{2} + m_2\right) + m_2\frac{l_2}{2}\sin(q_1 - q_2)}{m_0},$$

где $m_0 = m + m_1 + m_2$.

Чтобы определить взаимное влияние между квадрокоптером и роботизированным манипулятором, необходимо вычислить момент сил, действующий на квадрокоптер со стороны манипулятора, а также

взаимодействие между всеми механическими частями системы. С учетом моментов, созданных силой тяжести, получим τ_r – реактивный момент сил, действующий со стороны манипулятора на квадрокоптер:

$$\tau_r = \tau_1 + m_2 g l_{1c} \cos q_1 + \tau_{r21},$$

$$\tau_{r21} = \tau_{r12} = \tau_2 + m_2 g l_{2c} \cos(q_1 - q_2),$$

$$\tau_r = \tau_1 + m_2 g l_{1c} \cos q_1 + m_2 g l_{2c} \cos(q_1 - q_2) + \tau_2.$$

Исходя из вышеперечисленного, модель, состоящая из квадрокоптера, оснащенного манипулятором (рис. 3), описывается следующей системой уравнений:

$$\begin{cases} \ddot{x} = (\sin\psi \sin\varphi + \cos\psi \cos\theta \cos\varphi)\dfrac{U_1}{m_0}, \\ \ddot{y} = (-\cos\psi \sin\varphi + \sin\psi \sin\theta \cos\varphi)\dfrac{U_1}{m_0}, \\ \ddot{z} = -g + (\cos\theta \cos\varphi)\dfrac{U_1}{m_0}, \\ \ddot{\varphi} = \dfrac{I_{yy} - I_{zz}}{I_{xx}}\dot{\theta}\dot{\psi} - \dfrac{U_2 - \tau_r}{I_{xx}}, \\ \ddot{\theta} = \dfrac{I_{zz} - I_{xx}}{I_{yy}}\dot{\varphi}\dot{\psi} - \dfrac{U_3 - (m_1 + m_2) g \Delta y \sin\theta}{I_{yy}}, \\ \ddot{\psi} = \dfrac{I_{xx} - I_{yy}}{I_{zz}}\dot{\varphi}\dot{\theta} - \dfrac{U_4}{I_{zz}}, \\ M(\overline{q})\ddot{\overline{q}} + C(\overline{q}, \dot{\overline{q}})\dot{\overline{q}} + g(\overline{q}) = \overline{\tau}, \end{cases} \quad (2)$$

где $\overline{q}$ – обобщенные координаты в инерциальной системе отсчета.

Синтез закона управления

Система управления состоит из двух частей – регулятора для манипулятора и регулятора для квадрокоптера. В силу того, что квадрокоптеры оснащены высокопроизводительными вычислительными устройствами, гироскопами, акселерометрами, а манипуляторы оснащаются энкодерами, предположим, что обобщенные координаты и их первые производные в (2) измеримы. Уравнения динамики манипулятора являются нелинейными, поэтому используем метод линеаризации обратной связью [20]. Выберем нелинейный закон управления вида

$$\tau = M(\overline{q}) a_{\overline{q}} + C(\overline{q}, \dot{\overline{q}})\dot{\overline{q}} + g(\overline{q}),$$

где $a_{\overline{q}}$ является новым входным воздействием манипулятора.

Так как $\det M(\overline{q}) \neq 0$, то матрица инерции обратима. Новая система линейна и представляет систему независимых двойных интеграторов без перекрестных связей:

$$\ddot{\overline{q}} = a_{\overline{q}}.$$

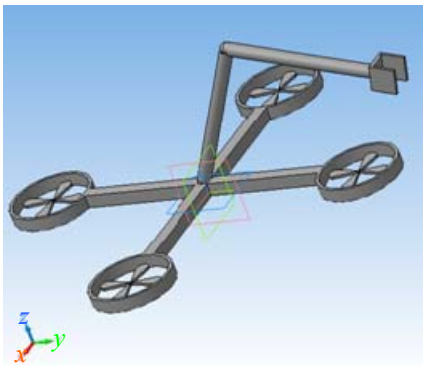


Рис. 3. Комбинированная система

Для управления полученной системой воспользуемся ПД-регулятором:

$$a_{\overline{q}} = \ddot{\overline{q}}^* + K_p(\overline{q} - \overline{q}^*) + K_d(\dot{\overline{q}} - \dot{\overline{q}}^*),$$

где K_p и K_d – пропорциональные и дифференциальные коэффициенты регулятора, $\overline{q}^*$ – желаемые траектории.

Цель управления состоит в обеспечении желаемых траекторий рабочего органа манипулятора. Для достижения этой цели необходимо реализовать закон управления квадрокоптером, обеспечивающий его позиционирование и компенсацию реактивного момента, создаваемого за счет движения манипулятора.

Мультикоптер имеет две неголономные связи, и, следовательно, для управления необходимо регулировать только четыре координаты. Выберем углы рысканья, тангажа, крена и высоту как опорные координаты для синтеза регулятора.

$$\begin{pmatrix} \ddot{\varphi} \\ \ddot{\theta} \\ \ddot{\psi} \\ \ddot{z} \end{pmatrix} = -\begin{pmatrix} \frac{U_2 - \tau_r}{I_{xx}} \\ \frac{U_3 - (m_1 + m_2) g \Delta y \sin\theta}{I_{yy}} \\ \frac{U_4}{I_{zz}} \\ -(\cos\theta\cos\varphi)\frac{U_1}{m_0} \end{pmatrix} - \begin{pmatrix} \frac{I_{yy} - I_{zz}}{I_{xx}} \dot{\theta}\dot{\psi} \\ \frac{I_{zz} - I_{xx}}{I_{yy}} \dot{\varphi}\dot{\psi} \\ \frac{I_{xx} - I_{yy}}{I_{zz}} \dot{\varphi}\dot{\theta} \\ g \end{pmatrix}. \quad (3)$$

Полученная система (3) является нелинейной и нестационарной. Для ее упрощения и синтеза регулятора используем метод линеаризации обратной связью и ПД-регулятор:

$$U_1 = -\frac{m_0}{\cos\varphi\cos\theta}(K_p e_z + K_d \dot{e}_z + g),$$

$$U_2 = -I_{xx}\left(K_d \dot{e}_\varphi + K_p e_\varphi + \tau_r + \frac{I_{yy} - I_{zz}}{I_{xx}} \dot{\theta}\dot{\psi}\right),$$

$$U_3 = -I_{yy}\left(K_d \dot{e}_\theta + K_p e_\theta + \frac{I_{zz} - I_{xx}}{I_{yy}} \dot{\varphi}\dot{\psi}\right),$$

$$U_4 = -I_{zz}\left(K_d \dot{e}_\psi + K_p e_\psi + \frac{I_{xx} - I_{yy}}{I_{zz}} \dot{\varphi}\dot{\theta} + (m_1 + m_2) g \Delta y \sin\theta\right),$$

где $e_\varphi = \varphi^* - \varphi, e_\theta = \theta^* - \theta, e_\psi = \psi^* - \psi, e_z = z^* - z$ – ошибки слежения (разница между желаемыми и текущими координатами). Тогда замкнутая система сводится к модели ошибки, которая может быть представлена в виде

$$\ddot{e} + \dot{e}K_d + eK_p = 0, \quad (4)$$

где $e = (e_\varphi, e_\theta, e_\psi, e_z, e_{\overline{q}_1}, e_{\overline{q}_2})^T$. Полученная модель ошибки является набором из шести систем второго порядка без перекрестных связей. При выборе положительных коэффициентов регулятора система (4) является асимптотически устойчивой.

Результаты моделирования

В ходе моделирования обобщенные координаты манипулятора изменялись в соответствии с рис. 4. При этом задачей управления квадрокоптером являлось движение по заранее заданной траектории. На рис. 5–8 показаны выходные значения координат квадрокоптера с предложенной компенсацией динамики манипулятора ($\varphi_1, \theta_1, \psi_1$ и z_1) и без нее ($\varphi_2, \theta_2, \psi_2$ и z_2).

На рис. 5 представлено изменение угла рысканья квадрокоптера при управлении предложенным методом (φ_1) и при управлении классическим методом на базе ПД-регулятора (φ_2). Исходя из результатов моделирования, можно заключить, что отклонение траектории φ_2 от желаемой φ_d значительно выше, чем отклонение φ_1 от φ_d. Аналогичные выводы можно сделать для изменения углов крена (рис. 6) и тангажа (рис. 7). Было выявлено, что динамика манипулятора не влияет на изменение высоты полета квадрокоптера (рис. 8).

Из результатов моделирования следует, что предложенный алгоритм управления обеспечивает устойчивость системы. Использование регулятора, компенсирующего динамику манипулятора, позволяет достигнуть более высоких показателей точности и эффективности, чем использование только пропорционально-дифференциального управления.

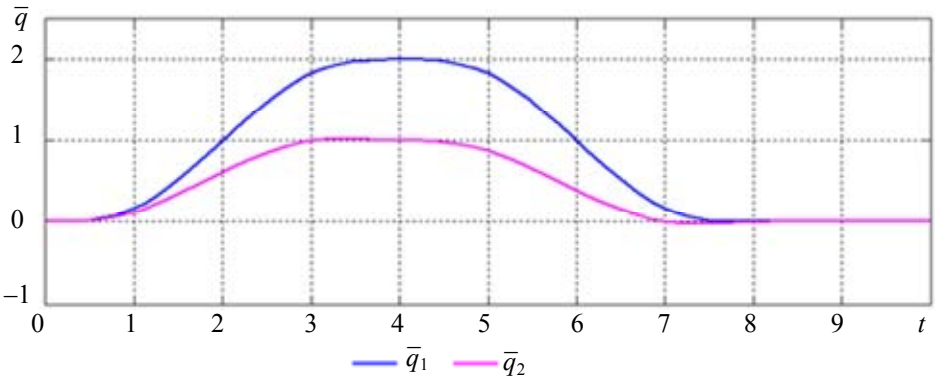


Рис. 4. Изменение обобщенных координат манипулятора

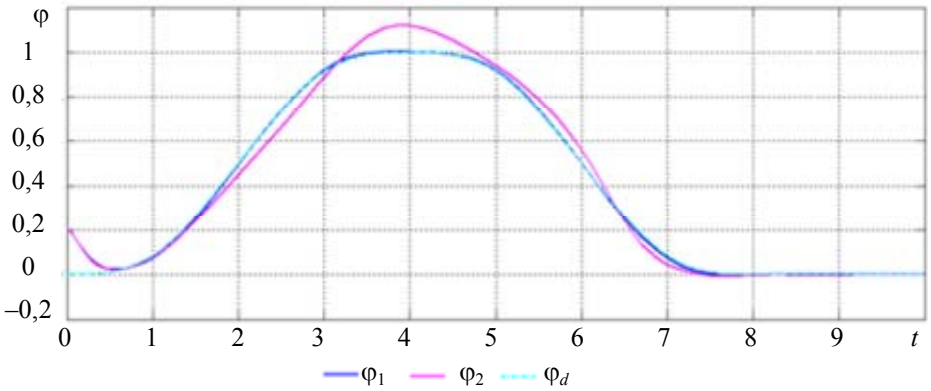


Рис. 5. Моделирование угла рысканья

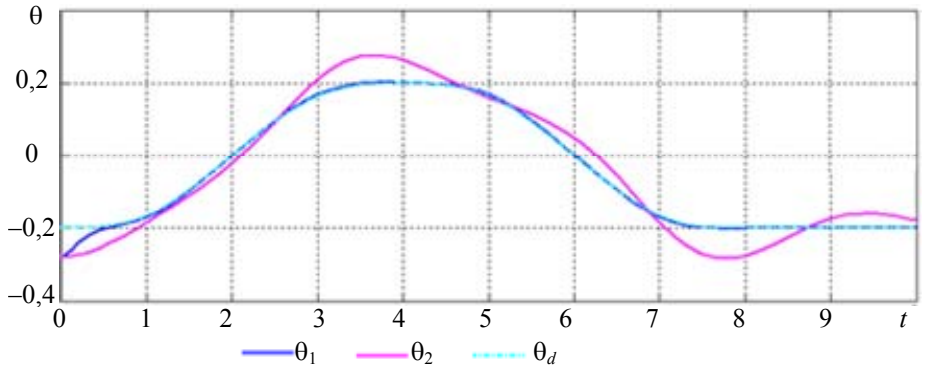


Рис. 6. Моделирование угла тангажа

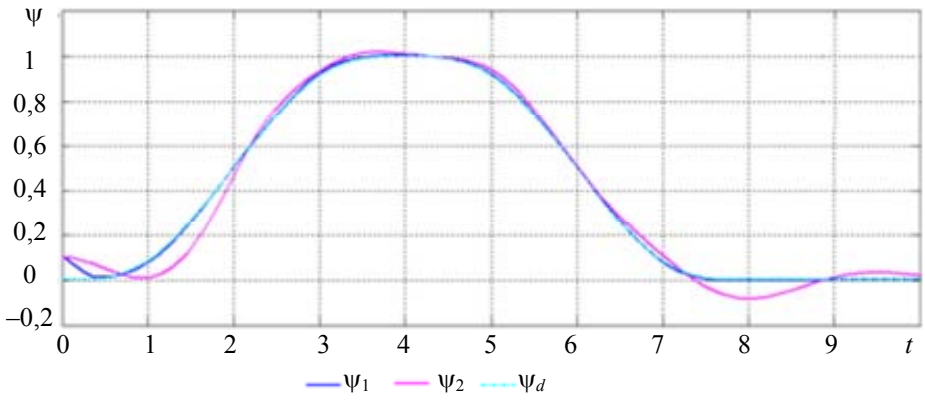


Рис. 7. Моделирование угла крена

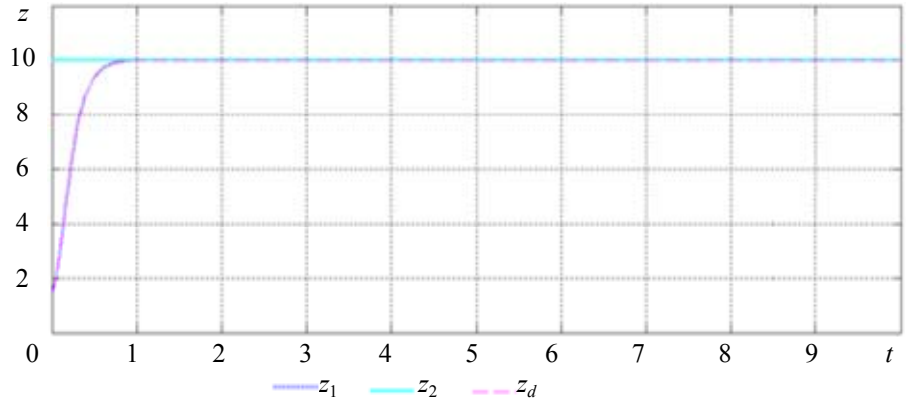


Рис. 8. Моделирование высоты

Заключение

В работе описаны кинематические и динамические модели квадрокоптера с присоединенным к центру плоским манипулятором с двумя степенями свободы. Для этой комбинированной системы разработан алгоритм управления на базе линеаризации обратной связью и пропорционально-дифференциального регулятора. Предложенный алгоритм обеспечивает компенсацию реактивного момента сил, действующих на квадрокоптер со стороны манипулятора, смещения тензора инерции и положения центра масс комбинированной системы. Результаты компьютерного моделирования подтверждают работоспособность и эффективность предложенного метода. В дальнейшем предполагается продолжить совершенствование синтезируемого алгоритма управления для функционирования квадрокоптера в условиях неопределенности, неизвестных параметров, внешних возмущающих воздействий, неучтенной динамики и проведения практических экспериментов.

Литература

1. Chettibi T., Haddad M. Dynamic modelling of a quadrotor aerial robot // Journees D'etudes Nationales de Mecanique. Batna, Algerie, 2007. P. 22–27.
2. Mokhtari A., Benallegue A. Dynamic feedback controller of Euler angles and wind parameters estimation for a quadrotor unmanned aerial vehicle // Proceedings – IEEE International Conference on Robotics and Automation. 2004. V. 2004. N 3. P. 2359–2366.
3. Гриценко П.А., Кремлев А.С., Шмигельский Г.М. Управление движением квадрокоптера по заранее заданной траектории // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 4 (86). С. 22–25.
4. Derafa L., Madani T., Benallegue A. Dynamic modelling and experimental identification of four rotors helicopter parameters // Proceedings of the IEEE International Conference on Industrial Technology. 2006. Art. 4237837. P. 1834–1839.
5. Altug E., Ostrowski J.P., Mahony R. Control of a quadrotor helicopter using visual feedback // Proceedings – IEEE International Conference on Robotics and Automation. 2002. V. 1. P. 72–77.
6. Waslander S.L., Hoffmann G.M., Jang J.S., Tomlin C.J. Multi-agent quadrotor testbed control design: integral sliding mode vs. reinforcement learning // IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS. 2005. Art. 1545025. P. 468–473.
7. Yang C.C., Lai L.C., Wu C.J. Time optimal control of a hovering quadrotor helicopter // IEEE ICSS International Conference on Systems and Signals. 2005. P. 295–300.
8. Catillo P., Loranzo R., Dzul A. Stabilization of a mini-rotorcraft having four rotors // 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2004. V. 3. P. 2693–2698.
9. Фуртат И.Б. Робастное субоптимальное управление боковым движением летательного аппарата в режиме захода на посадку // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 3 (85). С. 51–55.
10. Lindsey Q., Mellinger D., Kumar V. Construction of cubic structures with quadrotor teams // Robotics: Science and Systems VII. 2012. P. 177–184.
11. Willmann J., Augugliaro F., Cadalbert T., D'Andrea R., Gramazio F., Kohler M. Aerial robotic construction towards a new field of architectural research // International Journal of Architectural Computing. 2012. V. 10. N 3. P. 439–460.
12. Фуртат И.Б. Субоптимальное управление нелинейными мультиагентными системами // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 1 (83). С. 19–23.
13. Pounds P., Bersak D., Dollar A. Grasp from the air: hovering capture and load stability // IEEE/RSJ International Conference on Intelligent Robots and Systems. San Francisco, 2011. P. 2491–2498.

14. Bisgaard M., la Cour-Harbo A., Bendtsen J. Adaptive control system for autonomous helicopter slung load operations // Control Engineering Practice. 2010. V. 18. N 7. P. 800–811.
15. Palunko I., Fierro R., Cruz P. Trajectory generation for swing-free maneuvers of a quadrotor with suspended payload: a dynamic programming approach // Proc. IEEE International Conference on Robotics and Automation. 2012. Art. 6225213. P. 2691–2697.
16. Michael N., Fink J., Kumar V. Cooperative manipulation and transportation with aerial robots // Autonomous Robots. 2011. V. 30. N 1. P. 73–86.
17. Lippiello V., Ruggiero F. Exploiting redundancy in Cartesian impedance control of UAVs equipped with a robotic arm // IEEE International Conference on Intelligent Robots and Systems. 2012. Art. 6386021. P. 3768–3773.
18. Korpela C.M., Danko T.W., Oh P.Y. MM-UAV: mobile manipulating unmanned aerial vehicle // Journal of Intelligent and Robotic Systems: Theory and Applications. 2012. V. 65. N 1–4. P. 93–101.
19. Lippiello V., Ruggiero F. Cartesian impedance control of a UAV with a robotic arm // IFAC Proceedings Volumes. 2012. V. 10. N Part 1. P. 704–709.
20. Spong M.W., Hutchinson S., Vidyasagar M. Robot Modeling and Control. Wiley, 2005. 496 p.

Маргун Алексей Анатольевич – студент, лаборант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, alexeimargun@gmail.ru

Зименко Константин Александрович – студент, инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, kostyazimenko@gmail.com

Базылев Дмитрий Николаевич – студент, лаборант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, bazylevd@mail.ru

Бобцов Алексей Алексеевич – доктор технических наук, профессор, заведующий кафедрой, декан, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, bobtsov@mail.ru

Кремлев Артем Сергеевич – кандидат технических наук, зам. декана, доцент кафедры, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, kremlev_artem@mail.ru

Ибраев Денис Дамирович – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Ibray1522@gmail.ru

Чех Мартин – кандидат технических наук, научный сотрудник, Университет Западной Богемии, Пльзень, 306 14, Чехия, cechyn@gmail.com

Alexei A. Margun – laboratory assistant, student, ITMO University, Saint Petersburg, 197101, Russian Federation, alexeimargun@gmail.ru

Konstantin A. Zimenko – engineer, student, ITMO University, Saint Petersburg, 197101, Russian Federation, kostyazimenko@gmail.com

Dmitry N. Bazylev – laboratory assistant, student, ITMO University, Saint Petersburg, 197101, Russian Federation, bazylevd@mail.ru

Alexei A. Bobtsov – D.Sc., Professor, Dean, Department head, ITMO University, Saint Petersburg, 197101, Russian Federation, bobtsov@mail.ru

Artem S. Kremlev – PhD, Deputy Dean, Associate professor, ITMO University, Saint Petersburg, 197101, Russian Federation, kremlev_artem@mail.ru

Denis D. Ibraev – student, ITMO University, Saint Petersburg, 197101, Russian Federation, Ibray1522@gmail.ru

Martin Cech – PhD, scientific researcher, University of West Bohemia, Pilzen, 306 14, Czech Republic, cechyn@gmail.com

Принято к печати 31.03.14
Accepted 31.03.14

УДК 681.51

АДАПТИВНОЕ УПРАВЛЕНИЕ ПО ВЫХОДУ МНОГОКАНАЛЬНЫМИ ЛИНЕЙНЫМИ СТАЦИОНАРНЫМИ ПАРАМЕТРИЧЕСКИ НЕОПРЕДЕЛЕННЫМИ СИСТЕМАМИ

А.А. Бобцов[a], М.В. Фаронов[a], И.Б. Фуртат[a,b], А.А. Пыркин[a], В. Цзянь[c]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, cainenash@mail.ru
[b] Институт проблем машиноведения РАН, Санкт-Петербург, 199178, Российская Федерация
[c] Университет Ханчжоу Дяньцзы, Ханчжоу, 310018, Китай

Аннотация. Рассмотрена задача адаптивного управления параметрически неопределенными многоканальными линейными стационарными объектами с произвольной относительной степенью каждой локальной подсистемы. Синтезирован регулятор, обеспечивающий стабилизацию объекта управления при условии, что для каждой подсистемы: измеряются только выходные переменные; точно известны относительные степени, но не порядок линейных дифференциальных уравнений; выполнены условия минимальной фазовости. Для упрощения общей методики синтеза управления рассматривается процедура стабилизации двухканальной системы. В качестве базового подхода управления выбирается метод «последовательный компенсатор», основанный на использовании теоремы о пассификации А.Л. Фрадкова, с дополнительными фильтрами, содержащими в своей структуре большие коэффициенты усиления. Анализируется устойчивость замкнутой системы в классе указанных типов регуляторов, а также рассматриваются необходимые и достаточные условия, обеспечивающие экспоненциальные свойства сходимости. В качестве практических рекомендаций использования рассматриваемого подхода предлагается адаптивная версия метода «последовательный компенсатор», основанная на настройке коэффициента усиления, на базе алгоритма интегрального типа. Для иллюстрации работоспособности предлагаемого в работе подхода приведены результаты компьютерного моделирования для подсистем третьего и второго порядка соответственно, функционирующих в условиях полной параметрической неопределенности. Показано, что применение данной методики синтеза позволяет синтезировать алгоритмы управления многоканальными параметрически неопределенными системами, обладающими минимальным динамическим порядком в сравнении с известными зарубежными и отечественными аналогами.

Ключевые слова: адаптивное управление, многоканальные системы, параметрическая неопределенность.

Благодарности. Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01, Проект 14.Z50.31.0031, Госзадание 2014/190 (проект 2118)).

ADAPTIVE OUTPUT CONTROL OF MULTICHANNEL LINEAR STATIONARY SYSTEMS UNDER PARAMETRIC UNCERTAINTY

A.A. Bobtsov[a], M.V. Faronov[a], I.B. Furtat[a,b], A.A. Pyrkin[a], W. Jian[c]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, cainenash@mail.ru
[b] Institute of Problems of Mechanical Engineering Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation
[c] Hangzhou Dianzi University, Hangzhou 310018, China

Abstract. The paper deals with the problem of adaptive control for multi-channel linear stationary plants under parametric uncertainty with arbitrary relative degree of each local subsystem. The synthesized regulator provides stabilization of control plant on condition that for each local subsystem only output variables are measured with known relative degrees, but the order of linear differential equations is unknown. We consider the synthesis of control system for two-channel system for simplification of the synthesis method. The "serial compensator" algorithm is chosen as basic approach with A.L. Fradkov's passification theorem and additional filters containing high gain constants in their structure. Durability of the closed system in the group of pointed types of regulators is analyzed and the necessary and sufficient conditions for exponential convergence properties are considered. We suggest adaptive version of the "serial compensator" method from the practical point of view, where customization of the gain constant is based on the integral type algorithm. We show the results of computer simulation for the third and second order subsystems under parametric uncertainty to illustrate the proposed approach workability. It is shown that the proposed technique makes it possible to synthesize control algorithms for multichannel systems under parametric uncertainty with minimal dynamical order as compared to known foreign and domestic counterparts.

Keywords: adaptive control, multichannel systems, parametric uncertainty.

Acknowledgements. The work is partially financially supported by the Government of the Russian Federation (grant 074-U01), Project 14.Z50.31.0031, Government order 2014/190 (project 2118)).

Введение

Задача управления объектами в условии параметрической неопределенности со скалярным входом–выходом стала одной из классических задач современной теории управления. Для ее решения предложены и исследованы различные методы синтеза алгоритмов регулирования, например [1–4]. В последнее время наблюдается рост интереса к проблемам управления многоканальными объектами. Это связано с появлением новых задач в биологии, физике, робототехнике, энергетических и телекоммуникационных сетях и т.п. При этом возникает новая проблема, связанная с управлением группой взаимосвязанных объектов. На сегодняшний день предложено достаточно методов и подходов для управления многоканальными объектами. Так, в [2] предложен новый метод вложения систем, позволяющий синтезировать статические регуляторы, обеспечивающие инвариантность по отношению к возмущениям. В [5] предложен метод вспомогательного контура для компенсации возмущений, обобщенный затем для управления ли-

нейными [6, 7], нелинейными [8] и неминимально-фазовыми [9] многоканальными объектами. В [10] предложен алгоритм, близкий по идее к настоящей работе и позволяющий синтезировать простые алгоритмы компенсации возмущений. В [11] рассмотрено адаптивное управление многосвязными объектами с неминимальной реализацией эталонной модели. В [12, 13] представлена адаптивная схема управления по выходу нелинейными многоканальными системами с неизвестным нестационарным запаздыванием. В [14, 15] на базе метода бэкстеппинга предложено адаптивное управление для класса нелинейных многоканальных систем с несимметричными входными ограничениями. В [16] рассмотрена задача робастного слежения за эталонным сигналом на скользящих режимах. В [17] предложена адаптивная схема стабилизации многоканальной системы с использованием наблюдателя с большим коэффициентом усиления. Задача адаптивного дискретного управления нестационарной многоканальной системой на базе нечеткой логики рассмотрена в [18]. В [4] предложены алгоритмы управления многоканальными объектами с использованием метода скоростного градиента. Задачи адаптивного управления с оптимизацией, оптимального управления непрерывными и дискретными линейными, нелинейными, стохастическими объектами рассмотрены в [19, 20].

Несмотря на большое количество результатов в области управления многоканальными системами, по-прежнему актуальной задачей остается поиск простых алгоритмов управления, что особенно важно при управлении большой группой взаимосвязанных объектов. Одним из таких алгоритмов является «последовательный компенсатор», впервые предложенный в [12] для управления параметрически неопределенными объектами. Достоинство данного алгоритма состоит в простоте реализации и подборе настраиваемых параметров. Настоящая работа посвящена обобщению метода «последовательного компенсатора» на управление многоканальными системами управления.

В работе на примере двухканальной системы управления рассматривается применение метода [21] для управления многоканальными объектами, связанными через каналы выходов. Предполагается, что параметры каждой подсистемы неизвестны и доступны измерению только скалярные входы и выходы. Получены децентрализованные регуляторы и условия на расчет их параметров, которые обеспечивают экспоненциальную устойчивость выходов каждой локальной подсистемы. Приведены результаты моделирования, иллюстрирующие эффективность предложенного алгоритма.

В отличие от [10], где предложены результаты, схожие с результатами настоящей работы, здесь будет применен алгоритм «последовательный компенсатор», который не содержит дополнительного динамического звена (вспомогательного контура) для выделения возмущений. Исходя из этого, предложенный здесь алгоритм будет иметь меньший динамический порядок.

Постановка задачи

Ради простоты обобщения метода [21] на многоканальные системы рассмотрим двухканальный объект управления, математическая модель которого представлена следующими выражениями:

$$y_1(t) = \frac{b(p)}{a(p)} u_1(t) + \frac{c(p)}{a(p)} y_2(t), \tag{1}$$

$$y_2(t) = \frac{d(p)}{e(p)} u_2(t) + \frac{f(p)}{e(p)} y_1(t), \tag{2}$$

где $p = d/dt$ – оператор дифференцирования; выходные переменные $y_1(t)$, $y_2(t)$ измеряются, но их производные не поддаются измерению; $a(p) = p^n + a_{n-1}p^{n-1} + ... + a_1 p + a_0$, $b(p) = b_m p^m + ... + b_1 p + b_0$, $c(p) = c_r p^r + c_{r-1}p^{r-1} + ... + c_1 p + c_0$, $e(p) = p^z + e_{z-1}e^{z-1} + ... + e_1 p + e_0$, $d(p) = d_g p^g + ... + d_1 p + d_0$, $f(p) = f_i p^i + f_{i-1}p^{i-1} + ... + f_1 p + c_0$ – операторы с неизвестными коэффициентами; $m \le n-1$; $g \le z-1$; передаточные функции $W_1(s) = \frac{b(s)}{a(s)}$ и $W_2(s) = \frac{d(s)}{e(s)}$ имеют относительные степени $\rho_1 = n - m$ и $\rho_2 = z - g$ соответственно, s – комплексная переменная; полиномы $b(s)$ и $d(s)$ – гурвицевы, а коэффициенты $b_m > 0$, $d_g > 0$. Цель управления состоит в том, чтобы обеспечить экспоненциальную устойчивость положений равновесия $y_1 = 0$ и $y_2 = 0$.

Синтез алгоритма управления

В соответствии с [21] выберем закон управления для каждого канала следующим образом:

$$u_1(t) = -\tilde{k}_1 \alpha_1(p) \xi_{11}(t), \tag{3}$$

$$\begin{cases} \dot{\xi}_{11}(t) = \sigma_1 \xi_{12}(t), \\ \dot{\xi}_{12}(t) = \sigma_1 \xi_{13}(t), \\ \ldots \\ \dot{\xi}_{1;\rho_1-1}(t) = \sigma_1 \left(-k_{11}\xi_{11}(t) - k_{12}\xi_{12}(t) - \ldots - k_{1;\rho_1-1}\xi_{1;\rho_1-1}(t) + k_{11} y_1(t) \right), \end{cases} \tag{4}$$

$$u_2 = -\tilde{k}_2 \alpha_2(p) \xi_{21}(t), \tag{5}$$

$$\begin{cases} \dot{\xi}_{21}(t) = \sigma_2 \xi_{22}(t), \\ \dot{\xi}_{22}(t) = \sigma_2 \xi_{23}(t), \\ \ldots \\ \dot{\xi}_{2;\rho_2-1}(t) = \sigma_2 \left(-k_{21}\xi_{21}(t) - k_{22}\xi_{22}(t) - \ldots - k_{2;\rho_2-1}\xi_{2;\rho_2-1}(t) + k_{21} y_2(t) \right), \end{cases} \tag{6}$$

$$\tilde{k}_1 = \kappa_1 + \mu_1, \ \tilde{k}_2 = \kappa_2 + \mu_2, \tag{7}$$

где числа $\mu_1 > 0$, $\mu_2 > 0$ и оператор $\alpha_1(p)$ степени $\rho_1 - 1$ и $\alpha_2(p)$ степени $\rho_2 - 1$ выбираются таким образом, чтобы передаточные функции $H_1(s) = \dfrac{\alpha_1(s) b(s)}{a(s) + \mu_1 \alpha_1(s) b(s)}$ и $H_2(s) = \dfrac{\alpha_2(s) d(s)}{e(s) + \mu_2 \alpha_2(s) b(s)}$ были строго вещественно положительными, числа $\sigma_1 > \tilde{k}_1$, $\sigma_2 > \tilde{k}_2$, а коэффициенты k_{1i}, k_{2i} рассчитываются из требований экспоненциальной устойчивости систем (4), (6) при нулевых входах $y_1(t)$, $y_2(t)$.

Закон управления (3)–(7) является технически реализуемым, так как содержит только известные или измеряемые сигналы. Однако требуется найти аналитические условия его применимости для стабилизации объекта (1), (2) или, иными словами, найти ограничения на числа $\tilde{k}_1$, $\tilde{k}_2$ и σ_1, σ_2, при которых система (1)–(7) является экспоненциально устойчивой.

Основной результат

Проведем ряд преобразований. Подставляя (3) в уравнение (1), получим

$$y_1(t) = \frac{b(p)}{a(p)}[-k_1 \alpha_1(p) \hat{y}_1(t)] + \frac{c(p)}{a(p)} y_2(t) = \frac{b(p)}{a(p)}[-k_1 \alpha_1(p) y_1(t) + k_1 \alpha_1(p) \varepsilon_1(t)] + \frac{c(p)}{a(p)} y_2(t), \tag{8}$$

где функция $\hat{y}_1(t) = \xi_{11}(t)$ – оценка первой выходной переменной, ошибка $\varepsilon_1(t) = y_1(t) - \hat{y}_1(t)$. Аналогично для второго канала, подставляя (5) в (2), имеем

$$y_2(t) = \frac{d(p)}{e(p)}[-k_2 \alpha_2(p) \hat{y}_2(t)] + \frac{f(p)}{e(p)} y_1(t) = \frac{d(p)}{e(p)}[-k_2 \alpha_2(p) y_2(t) + k_2 \alpha_2(p) \varepsilon_2(t)] + \frac{f(p)}{e(p)} y_1(t), \tag{9}$$

где функция $\hat{y}_2(t) = \xi_{12}(t)$ – оценка второй выходной переменной, $\varepsilon_2(t) = y_2(t) - \hat{y}_2(t)$.

После простых преобразований модели (8), (9) можно представить как

$$y_1(t) = \frac{b(p)\alpha_1(p)}{a(p) + \mu_1 b(p) \alpha_1(p)}[-\kappa_1 y_1(t) + (\mu_1 + \kappa_1)\varepsilon_1(t)] + \frac{c(p)}{a(p) + \mu_1 b(p)\alpha_1(p)} y_2(t), \tag{10}$$

$$y_2(t) = \frac{d(p)\alpha_2(p)}{e(p) + \mu_2 d(p) \alpha_2(p)}[-\kappa_2 y_2(t) + (\mu_2 + \kappa_2)\varepsilon_2(t)] + \frac{f(p)}{e(p) + \mu_2 d(p)\alpha_2(p)} y_1(t), \tag{11}$$

где передаточные функции

$$H_1(s) = \frac{b(s)\alpha_1(s)}{a(s) + \mu_1 b(s)\alpha_1(s)}; \ H_2(s) = \frac{d(s)\alpha_2(s)}{e(s) + \mu_2 d(s)\alpha_2(s)} \tag{12}$$

строго вещественно положительны. Представим (10), (11) в форме вход–состояние–выход:

$$\dot{\mathbf{x}}_1(t) = \mathbf{A}_1 \mathbf{x}_1(t) + \mathbf{b}_1 \left(-\kappa_1 y_1(t) + (\mu_1 + \kappa_1)\varepsilon_1(t) \right) + \mathbf{q}_1 y_2(t), \tag{13}$$

$$y_1(t) = \mathbf{c}_1^T \mathbf{x}_1(t), \tag{14}$$

$$\dot{\mathbf{x}}_2(t) = \mathbf{A}_2 \mathbf{x}_2(t) + \mathbf{b}_2 \left(-\kappa_2 y_2(t) + (\mu_2 + \kappa_2)\varepsilon_2(t) \right) + \mathbf{q}_2 y_1(t), \tag{15}$$

$$y_2(t) = \mathbf{c}_2^T \mathbf{x}_2(t), \tag{16}$$

где $\mathbf{x}_1(t) \in R^n, \mathbf{x}_2(t) \in R^z$ – векторы переменных состояния (10), (11); $\mathbf{A}_1$, $\mathbf{b}_1$, $\mathbf{q}_1$ $\mathbf{c}_1$, $\mathbf{A}_2$, $\mathbf{b}_2$, $\mathbf{q}_2$ и $\mathbf{c}_2$ – соответствующие матрицы перехода от (10), (11) к (13), (14) и (15), (16).

Так как передаточные функции (12) удовлетворяет условиям строгой вещественной положительности, то в силу леммы Якубовича–Калмана [3] можно указать симметрические положительно определенные матрицы $\mathbf{P}_1$, $\mathbf{P}_2$, удовлетворяющие следующим матричным уравнениям:

$$\mathbf{A}_1^T\mathbf{P}_1 + \mathbf{P}_1\mathbf{A}_1 = -\mathbf{Q}_1, \quad \mathbf{P}_1\mathbf{b}_1 = \mathbf{c}_1, \tag{17}$$

$$\mathbf{A}_2^T\mathbf{P}_2 + \mathbf{P}_2\mathbf{A}_2 = -\mathbf{Q}_2, \quad \mathbf{P}_2\mathbf{b}_2 = \mathbf{c}_2, \tag{18}$$

где $\mathbf{Q}_1 = \mathbf{Q}_1^T$, $\mathbf{Q}_2 = \mathbf{Q}_2^T$ – некоторые положительно определенные матрицы.

Представим модели (4), (6) в векторно-матричной форме:

$$\dot{\xi}_1(t) = \sigma_1\left(\mathbf{\Gamma}_1\xi_1(t) + \mathbf{d}_1k_{11}y_1(t)\right), \tag{19}$$

$$\hat{y}_1(t) = \mathbf{h}_1^T\xi_1(t), \tag{20}$$

$$\dot{\xi}_2(t) = \sigma_2\left(\mathbf{\Gamma}_2\xi_2(t) + \mathbf{d}_2k_{21}y_2(t)\right), \tag{21}$$

$$\hat{y}_2(t) = \mathbf{h}_2^T\xi_2(t), \tag{22}$$

где $\mathbf{\Gamma}_1 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -k_{11} & -k_{12} & -k_{13} & \dots & -k_{1;\rho_1-1} \end{bmatrix}$, $\mathbf{\Gamma}_2 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -k_{21} & -k_{22} & -k_{23} & \dots & -k_{2;\rho_2-1} \end{bmatrix}$, $\mathbf{d}_1^T = \begin{bmatrix} 0 & 0 & 0 & \dots & 1 \end{bmatrix}_{(\rho_1-1)\times 1}$,

$\mathbf{d}_2^T = \begin{bmatrix} 0 & 0 & 0 & \dots & 1 \end{bmatrix}_{(\rho_2-1)\times 1}$ и $\mathbf{h}_1^T = [1 \quad 0 \quad 0 \quad \dots \quad 0]_{(\rho_1-1)\times 1}$, $\mathbf{h}_2^T = [1 \quad 0 \quad 0 \quad \dots \quad 0]_{(\rho_2-1)\times 1}$.

Введем в рассмотрение векторы отклонений:

$$\mathbf{\eta}_1(t) = \mathbf{h}_1y_1(t) - \xi_1(t), \tag{23}$$

$$\mathbf{\eta}_2(t) = \mathbf{h}_2y_2(t) - \xi_2(t), \tag{24}$$

тогда в силу структуры $\mathbf{h}_1$, $\mathbf{h}_2$ ошибки $\varepsilon_1(t)$ и $\varepsilon_2(t)$ можно представить как

$$\varepsilon_1(t) = y_1(t) - \hat{y}_1(t) = \mathbf{h}_1^T\mathbf{h}_1y_1(t) - \mathbf{h}_1^T\xi_1(t) = \mathbf{h}_1^T\mathbf{\eta}_1(t), \tag{25}$$

$$\varepsilon_2(t) = y_2(t) - \hat{y}_2(t) = \mathbf{h}_2^T\mathbf{h}_2y_2(t) - \mathbf{h}_2^T\xi_2(t) = \mathbf{h}_2^T\mathbf{\eta}_2(t). \tag{26}$$

Дифференцируя векторы (23), (24), получаем:

$$\dot{\mathbf{\eta}}_1(t) = \mathbf{h}_1\dot{y}_1(t) - \sigma_1\left(\mathbf{\Gamma}_1(\mathbf{h}_1y_1(t) - \mathbf{\eta}_1(t)) + \mathbf{d}_1k_{11}y_1(t)\right) = \mathbf{h}_1\dot{y}_1(t) + \sigma_1\mathbf{\Gamma}_1\mathbf{\eta}_1(t) - \sigma_1(\mathbf{d}_1k_{11} + \mathbf{\Gamma}_1\mathbf{h}_1)y_1(t), \tag{27}$$

$$\dot{\mathbf{\eta}}_2(t) = \mathbf{h}_2\dot{y}_2(t) - \sigma_2\left(\mathbf{\Gamma}_2(\mathbf{h}_2y_2(t) - \mathbf{\eta}_2(t)) + \mathbf{d}_2k_{21}y_2(t)\right) =$$

$$\mathbf{h}_2\dot{y}_2(t) + \sigma_2\mathbf{\Gamma}_2\mathbf{\eta}_2(t) - \sigma_2(\mathbf{d}_2k_{21} + \mathbf{\Gamma}_2\mathbf{h}_2)y_2(t). \tag{28}$$

Так как $\mathbf{d}_1k_{11} = -\mathbf{\Gamma}_1\mathbf{h}_1$ и $\mathbf{d}_2k_{21} = -\mathbf{\Gamma}_2\mathbf{h}_2$ (может быть проверено подстановкой), то

$$\dot{\mathbf{\eta}}_1(t) = \mathbf{h}_1\dot{y}_1(t) + \sigma_1\mathbf{\Gamma}_1\mathbf{\eta}_1(t), \varepsilon_1(t) = \mathbf{h}_1^T\mathbf{\eta}_1(t), \tag{29}$$

$$\dot{\mathbf{\eta}}_2(t) = \mathbf{h}_2\dot{y}_2(t) + \sigma_2\mathbf{\Gamma}_2\mathbf{\eta}_2(t), \ \varepsilon_2(t) = \mathbf{h}_2^T\mathbf{\eta}_2(t), \tag{30}$$

где матрицы $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$ – гурвицевы в силу расчета параметров k_i систем (4), (6) и, следовательно,

$$\mathbf{\Gamma}_1^T\mathbf{N}_1 + \mathbf{N}_1\mathbf{\Gamma}_1 = -\mathbf{Q}_3, \quad \mathbf{\Gamma}_2^T\mathbf{N}_2 + \mathbf{N}_2\mathbf{\Gamma}_2 = -\mathbf{Q}_4 \tag{31}$$

где $\mathbf{N}_1 = \mathbf{N}_1^T > 0$, $\mathbf{N}_2 = \mathbf{N}_2^T > 0$ и $\mathbf{Q}_3 = \mathbf{Q}_3^T > 0$, $\mathbf{Q}_4 = \mathbf{Q}_4^T > 0$.

Условия работоспособности закона управления (3)–(7) для стабилизации системы (13)–(16), (19)–(22), (29), (30) приведены в следующей теореме.

Утверждение. Рассмотрим двухканальную систему (1), (2). Пусть числа $\rho_1 = n - m > 0$ и $\rho_2 = z - g > 0$, а полиномы $b(s)$ и $d(s)$ – гурвицевы.

Пусть числа $0 < \delta < 0{,}5$, κ_1 и κ_2 такие, что выполняются неравенства:

$$-\mathbf{Q}_1 + \delta(\mu_1 + \kappa_1)\mathbf{c}_1\mathbf{c}_1^T + \kappa_2^{-1}\mathbf{P}_1\mathbf{q}_1\mathbf{q}_1^T\mathbf{P}_1^T + \kappa_2^{-1}\mathbf{c}_1^T\mathbf{q}_2^T\mathbf{b}_2^T\mathbf{b}_2\mathbf{q}_2\mathbf{c}_1 + \delta\mathbf{A}_1^T\mathbf{A}_1 \le -\mathbf{Q} < 0,$$

$$-\mathbf{Q}_2 + \delta(\mu_2 + \kappa_2)\mathbf{c}_2\mathbf{c}_2^T + \kappa_1^{-1}\mathbf{P}_2\mathbf{q}_2\mathbf{q}_2^T\mathbf{P}_2^T + \kappa_1^{-1}\mathbf{c}_2^T\mathbf{q}_1^T\mathbf{b}_1^T\mathbf{b}_1\mathbf{q}_1\mathbf{c}_2 + \delta\mathbf{A}_2^T\mathbf{A}_2 \le -\mathbf{Q} < 0.$$

Пусть числа σ_1 и σ_2 такие, что выполняются неравенства:

$$-\sigma_1\mathbf{Q}_3 + \delta^{-1}(\mu_1 + \kappa_1)\mathbf{h}\mathbf{h}^T + \delta^{-1}\mathbf{N}_1\mathbf{h}\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}^T\mathbf{N}_1^T + \kappa_1\mathbf{N}_1\mathbf{h}\mathbf{c}_1^T\mathbf{b}_1\mathbf{b}_1^T\mathbf{c}_1\mathbf{h}^T\mathbf{N}_1^T + (\mu_1 + \kappa_1)\mathbf{N}_1\mathbf{h}\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}^T$$
$$+(\mu_1 + \kappa_1)\mathbf{h}\mathbf{b}_1^T\mathbf{b}_1\mathbf{h}^T + \kappa_1\mathbf{N}_1\mathbf{h}\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}^T\mathbf{N}_1^T \le -\mathbf{Q} < 0,$$

$$-\sigma_2\mathbf{Q}_4 + \delta^{-1}(\mu_2 + \kappa_2)\mathbf{h}\mathbf{h}^T + \delta^{-1}\mathbf{N}_2\mathbf{h}\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}^T\mathbf{N}_2^T + \kappa_2\mathbf{N}_2\mathbf{h}\mathbf{c}_2^T\mathbf{b}_2\mathbf{b}_2^T\mathbf{c}_2\mathbf{h}^T\mathbf{N}_2^T + (\mu_2 + \kappa_2)\mathbf{N}_2\mathbf{h}\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}^T$$
$$+(\mu_2 + \kappa_2)\mathbf{h}\mathbf{b}_2^T\mathbf{b}_2\mathbf{h}^T + \kappa_2\mathbf{N}_2\mathbf{h}\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}^T\mathbf{N}_2^T \le -\mathbf{Q} < 0.$$

Тогда система (13)–(16), (29), (30) экспоненциально устойчива.

Доказательство. Рассмотрим следующую функцию Ляпунова:

$$V(t) = \mathbf{x}_1^T(t)\mathbf{P}_1\mathbf{x}_1(t) + \mathbf{x}_2^T(t)\mathbf{P}_2\mathbf{x}_2(t) + \boldsymbol{\eta}_1^T(t)\mathbf{N}_1\boldsymbol{\eta}_1(t) + \boldsymbol{\eta}_2^T(t)\mathbf{N}_2\boldsymbol{\eta}_2(t). \quad (32)$$

Дифференцируя (32) с учетом уравнений (13), (15), (25)–(30), получаем

$$\dot V(t) = \mathbf{x}_1^T(t)\left(\mathbf{A}_1^T\mathbf{P}_1 + \mathbf{P}_1\mathbf{A}_1\right)\mathbf{x}_1(t) - 2\kappa_1\mathbf{x}_1^T(t)\mathbf{P}_1\mathbf{b}_1y_1(t) + 2(\mu_1+\kappa_1)x_1^T(t)\mathbf{P}_1\mathbf{b}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t) + 2\mathbf{x}_1^T(t)\mathbf{P}_1\mathbf{q}_1y_2(t)$$
$$+\mathbf{x}_2^T(t)\left(\mathbf{A}_2^T\mathbf{P}_2 + \mathbf{P}_2\mathbf{A}_2\right)\mathbf{x}_2(t) - 2\kappa_2\mathbf{x}_2^T(t)\mathbf{P}_2\mathbf{b}_2y_2(t) + 2(\mu_2+\kappa_2)\mathbf{x}_2^T(t)\mathbf{P}_2\mathbf{b}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t) + 2\mathbf{x}_2^T(t)\mathbf{P}_2\mathbf{q}_2y_1(t)$$
$$+\sigma_1\boldsymbol{\eta}_1^T(t)\left(\boldsymbol{\Gamma}_1^T\mathbf{N}_1 + \mathbf{N}_1\boldsymbol{\Gamma}_1\right)\boldsymbol{\eta}_1(t) + 2\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{A}_1\mathbf{x}_1(t) - 2\kappa_1\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{b}_1y_1(t) + 2(\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{b}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t)$$
$$+2\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{b}_1\mathbf{q}_1y_2(t) + \sigma_2\boldsymbol{\eta}_2^T(t)\left(\boldsymbol{\Gamma}_2^T\mathbf{N}_2 + \mathbf{N}_2\boldsymbol{\Gamma}_2\right)\boldsymbol{\eta}_2(t) + 2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{A}_2\mathbf{x}_2(t) - 2\kappa_2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{b}_2y_2(t)$$
$$+2(\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{b}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t) + 2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{b}_2\mathbf{q}_2y_1(t). \quad (33)$$

Воспользуемся следующими оценками:

$$2(\mu_1+\kappa_1)\mathbf{x}_1^T(t)\mathbf{c}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t) \le \delta(\mu_1+\kappa_1)\mathbf{x}_1^T(t)\mathbf{c}_1\mathbf{c}_1^T\mathbf{x}_1(t) + \delta^{-1}(\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{h}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t),$$
$$2\mathbf{x}_1^T(t)\mathbf{P}_1\mathbf{q}_1y_2(t) \le \kappa_2^{-1}\mathbf{x}_1^T(t)\mathbf{P}_1\mathbf{q}_1\mathbf{q}_1^T\mathbf{P}_1^T\mathbf{x}_1(t) + \kappa_2y_2^2(t),$$
$$2(\mu_2+\kappa_2)\mathbf{x}_2^T(t)\mathbf{c}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t) \le \delta(\mu_2+\kappa_2)\mathbf{x}_2^T(t)\mathbf{c}_2\mathbf{c}_2^T\mathbf{x}_2(t) + \delta^{-1}(\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{h}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t),$$
$$2\mathbf{x}_2^T(t)\mathbf{P}_2\mathbf{q}_2y_1(t) \le \kappa_1^{-1}\mathbf{x}_2^T(t)\mathbf{P}_2\mathbf{q}_2\mathbf{q}_2^T\mathbf{P}_2^T\mathbf{x}_2(t) + \kappa_1y_1^2(t),$$
$$2\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{A}_1\mathbf{x}_1(t) \le \delta^{-1}\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}_1^T\mathbf{N}_1^T\boldsymbol{\eta}_1(t) + \delta\mathbf{x}_1^T(t)\mathbf{A}_1^T\mathbf{A}_1\mathbf{x}_1(t),$$
$$-2\kappa_1\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{b}_1y_1(t) \le \kappa_1\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{b}_1\mathbf{b}_1^T\mathbf{c}_1\mathbf{h}_1^T\mathbf{N}_1^T\boldsymbol{\eta}_1(t) + \kappa_1y_1^2(t),$$
$$2(\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{b}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t) \le (\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t) + (\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{h}_1\mathbf{b}_1^T\mathbf{b}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t),$$
$$2\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{b}_1\mathbf{q}_1y_2(t) \le \kappa_1\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}_1^T\mathbf{N}_1^T\boldsymbol{\eta}_1(t) + \kappa_1^{-1}\mathbf{x}_2^T(t)\mathbf{c}_2^T\mathbf{q}_1^T\mathbf{b}_1^T\mathbf{b}_1\mathbf{q}_1\mathbf{c}_2\mathbf{x}_2(t),$$
$$2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{A}_2\mathbf{x}_2(t) \le \delta^{-1}\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}\mathbf{N}_2^T\boldsymbol{\eta}_2(t) + \delta\mathbf{x}_2^T(t)\mathbf{A}_2^T\mathbf{A}_2\mathbf{x}_2(t),$$
$$-2\kappa_2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{b}_2y_2(t) \le \kappa_2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{b}_2\mathbf{b}_2^T\mathbf{c}_2\mathbf{h}_2^T\mathbf{N}_2^T\boldsymbol{\eta}_2(t) + \kappa_2y_2^2(t),$$
$$2(\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{b}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t) \le (\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t) + (\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{h}_2\mathbf{b}_2^T\mathbf{b}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t),$$
$$2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{b}_2\mathbf{q}_2y_1(t) \le \kappa_2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}_2^T\mathbf{N}_2^T\boldsymbol{\eta}_2(t) + \kappa_2^{-1}\mathbf{x}_1^T(t)\mathbf{c}_1^T\mathbf{q}_2^T\mathbf{b}_2^T\mathbf{b}_2\mathbf{q}_2\mathbf{c}_1\mathbf{x}_1(t).$$

Принимая во внимание полученные оценки и подставляя в (33) уравнения (17), (18), (31), можно получить следующее соотношение для производной (33):

$$\dot V(t) \le -\mathbf{x}_1^T(t)\mathbf{Q}_1\mathbf{x}_1(t) - \mathbf{x}_2^T(t)\mathbf{Q}_2\mathbf{x}_2(t) - \sigma_1\boldsymbol{\eta}_1^T(t)\mathbf{Q}_3\boldsymbol{\eta}_1(t) - \sigma_2\boldsymbol{\eta}_2^T(t)\mathbf{Q}_4\boldsymbol{\eta}_2(t)$$
$$+\delta(\mu_1+\kappa_1)\mathbf{x}_1^T(t)\mathbf{c}_1\mathbf{c}_1^T\mathbf{x}_1(t) + \kappa_2^{-1}\mathbf{x}_1^T(t)\mathbf{P}_1\mathbf{q}_1\mathbf{q}_1^T\mathbf{P}_1^T\mathbf{x}_1(t) + \kappa_2^{-1}\mathbf{x}_1^T(t)\mathbf{c}_1^T\mathbf{q}_2^T\mathbf{b}_2^T\mathbf{b}_2\mathbf{q}_2\mathbf{c}_1\mathbf{x}_1(t) + \delta\mathbf{x}_1^T(t)\mathbf{A}_1^T\mathbf{A}_1\mathbf{x}_1(t)$$
$$+\delta(\mu_2+\kappa_2)\mathbf{x}_2^T(t)\mathbf{c}_2\mathbf{c}_2^T\mathbf{x}_2(t) + \kappa_1^{-1}\mathbf{x}_2^T(t)\mathbf{P}_2\mathbf{q}_2\mathbf{q}_2^T\mathbf{P}_2^T\mathbf{x}_2(t) + \kappa_1^{-1}\mathbf{x}_2^T(t)\mathbf{c}_2^T\mathbf{q}_1^T\mathbf{b}_1^T\mathbf{b}_1\mathbf{q}_1\mathbf{c}_2\mathbf{x}_2(t) + \delta\mathbf{x}_2^T(t)\mathbf{A}_2^T\mathbf{A}_2\mathbf{x}_2(t)$$
$$+\delta^{-1}(\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{h}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t) + \delta^{-1}\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}_1^T\mathbf{N}_1^T\boldsymbol{\eta}_1(t) + \boldsymbol{\kappa}_1\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{b}_1\mathbf{b}_1^T\mathbf{c}_1\mathbf{h}_1^T\mathbf{N}_1^T\boldsymbol{\eta}_1(t)$$
$$+(\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t) + (\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{h}_1\mathbf{b}_1^T\mathbf{b}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t) + \kappa_1\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}\mathbf{N}_1^T\boldsymbol{\eta}_1(t)$$
$$+\delta^{-1}(\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{h}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t) + \delta^{-1}\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}_2^T\mathbf{N}_2^T\boldsymbol{\eta}_2(t) + \kappa_2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{b}_2\mathbf{b}_2^T\mathbf{c}_2\mathbf{h}_2^T\mathbf{N}_2^T\boldsymbol{\eta}_2(t)$$
$$+(\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t) + (\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{h}_2\mathbf{b}_2^T\mathbf{b}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t) + \kappa_2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}_2^T\mathbf{N}_2^T\boldsymbol{\eta}_2(t). \quad (34)$$

Пусть числа $0 < \delta < 0{,}5$, κ_1 и κ_2 такие, что выполняются неравенства:

$$-\mathbf{Q}_1 + \delta(\mu_1+\kappa_1)\mathbf{c}_1\mathbf{c}_1^T + \kappa_2^{-1}\mathbf{P}_1\mathbf{q}_1\mathbf{q}_1^T\mathbf{P}_1^T + \kappa_2^{-1}\mathbf{c}_1^T\mathbf{q}_2^T\mathbf{b}_2^T\mathbf{b}_2\mathbf{q}_2\mathbf{c}_1 + \delta\mathbf{A}_1^T\mathbf{A}_1 \le -\mathbf{Q} < 0, \quad (35)$$
$$-\mathbf{Q}_2 + \delta(\mu_2+\kappa_2)\mathbf{c}_2\mathbf{c}_2^T + \kappa_1^{-1}\mathbf{P}_2\mathbf{q}_2\mathbf{q}_2^T\mathbf{P}_2^T + \kappa_1^{-1}\mathbf{c}_2^T\mathbf{q}_1^T\mathbf{b}_1^T\mathbf{b}_1\mathbf{q}_1\mathbf{c}_2 + \delta\mathbf{A}_2^T\mathbf{A}_2 \le -\mathbf{Q} < 0. \quad (36)$$

В этом случае из выражения (34) следует:

$$\dot V(t) \le -\mathbf{x}_1^T(t)\mathbf{Q}_1\mathbf{x}_1(t) - \mathbf{x}_2^T(t)\mathbf{Q}_2\mathbf{x}_2(t) - \sigma_1\boldsymbol{\eta}_1^T(t)\mathbf{Q}_3\boldsymbol{\eta}_1(t) - \sigma_2\boldsymbol{\eta}_2^T(t)\mathbf{Q}_4\boldsymbol{\eta}_2(t) + \delta^{-1}(\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{h}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t)$$
$$+\delta^{-1}\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}\mathbf{N}_1^T\boldsymbol{\eta}_1(t) + \kappa_1\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{b}_1\mathbf{b}_1^T\mathbf{c}_1\mathbf{h}_1^T\mathbf{N}_1^T\boldsymbol{\eta}_1(t) + (\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t)$$
$$+(\mu_1+\kappa_1)\boldsymbol{\eta}_1^T(t)\mathbf{h}_1\mathbf{b}_1^T\mathbf{b}_1\mathbf{h}_1^T\boldsymbol{\eta}_1(t) + \kappa_1\boldsymbol{\eta}_1^T(t)\mathbf{N}_1\mathbf{h}_1\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}_1^T\mathbf{N}_1^T\boldsymbol{\eta}_1(t) + \delta^{-1}(\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{h}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t)$$
$$+\delta^{-1}\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}_2^T\mathbf{N}_2^T\boldsymbol{\eta}_2(t) + \kappa_2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{b}_2\mathbf{b}_2^T\mathbf{c}_2\mathbf{h}_2^T\mathbf{N}_2^T\boldsymbol{\eta}_2(t) + (\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t)$$
$$+(\mu_2+\kappa_2)\boldsymbol{\eta}_2^T(t)\mathbf{h}_2\mathbf{b}_2^T\mathbf{b}_2\mathbf{h}_2^T\boldsymbol{\eta}_2(t) + \kappa_2\boldsymbol{\eta}_2^T(t)\mathbf{N}_2\mathbf{h}_2\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}^T\mathbf{N}_2^T\boldsymbol{\eta}_2(t). \quad (37)$$

Пусть числа σ_1 и σ_2 такие, что выполняются неравенства

$$-\sigma_1\mathbf{Q}_3 + \delta^{-1}(\mu_1+\kappa_1)\mathbf{h}\mathbf{h}^T + \delta^{-1}\mathbf{N}_1\mathbf{h}\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}^T\mathbf{N}_1^T + \kappa_1\mathbf{N}_1\mathbf{h}\mathbf{c}_1^T\mathbf{b}_1\mathbf{b}_1^T\mathbf{c}_1\mathbf{h}^T\mathbf{N}_1^T + (\mu_1+\kappa_1)\mathbf{N}_1\mathbf{h}\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}^T$$
$$+(\mu_1+\kappa_1)\mathbf{h}\mathbf{b}_1^T\mathbf{b}_1\mathbf{h}^T + \kappa_1\mathbf{N}_1\mathbf{h}\mathbf{c}_1^T\mathbf{c}_1\mathbf{h}^T\mathbf{N}_1^T \le -\mathbf{Q} < 0, \quad (38)$$

$$-\sigma_2\mathbf{Q}_4 + \delta^{-1}(\mu_2+\kappa_2)\mathbf{h}\mathbf{h}^T + \delta^{-1}\mathbf{N}_2\mathbf{h}\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}^T\mathbf{N}_2^T + \kappa_2\mathbf{N}_2\mathbf{h}\mathbf{c}_2^T\mathbf{b}_2\mathbf{b}_2^T\mathbf{c}_2\mathbf{h}^T\mathbf{N}_2^T + (\mu_2+\kappa_2)\mathbf{N}_2\mathbf{h}\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}^T$$
$$+(\mu_2+\kappa_2)\mathbf{h}\mathbf{b}_2^T\mathbf{b}_2\mathbf{h}^T + \kappa_2\mathbf{N}_2\mathbf{h}\mathbf{c}_2^T\mathbf{c}_2\mathbf{h}^T\mathbf{N}_2^T \le -\mathbf{Q} < 0. \quad (39)$$

Тогда из выражения (37) следует, что

$$\dot{V}(t) \le -\mathbf{x}_1^T(t)\mathbf{Q}_1\mathbf{x}_1(t) - \mathbf{x}_2^T(t)\mathbf{Q}_2\mathbf{x}_2(t) - \sigma_1\boldsymbol{\eta}_1^T(t)\mathbf{Q}_3\boldsymbol{\eta}_1(t) - \sigma_2\boldsymbol{\eta}_2^T(t)\mathbf{Q}_4\boldsymbol{\eta}_2(t). \quad (40)$$

Из выражения (40) следует экспоненциальная устойчивость системы (13)–(16), (19)–(22), (29), (30), что и требовалось доказать.

Замечание. При внимательном рассмотрении можно отметить, что неравенства (35), (36), (38), (39) не являются противоречивыми. Очевидно, что при некотором малом δ, больших κ_1, κ_2 и еще больших $\sigma_1 > \kappa_1$, $\sigma_1 > \delta^{-1}$ и $\sigma_2 > \kappa_2$, $\sigma_2 > \delta^{-1}$ неравенства будут выполнены. Таким образом, в условиях неопределенности объекта в качестве возможного варианта настройки параметров κ_1, κ_2, σ_1 и σ_2 можно увеличивать их значения до тех пор, пока не будут выполнены следующие условия:

$$|y_1(t)| < \delta_0 \text{ для } \forall t \ge t_1, \quad (41)$$

$$|y_2(t)| < \delta_0 \text{ для } \forall t \ge t_2, \quad (42)$$

где положительное число δ_0 задается разработчиком системы управления.

Для настройки параметров $\tilde{k}_1 = \kappa_1 + \mu_1$ и $\tilde{k}_2 = \kappa_2 + \mu_2$ воспользуемся алгоритмом

$$\tilde{k}_1(t) = \int_{t_0}^{t} \lambda_1(\tau)d\tau, \ \tilde{k}_2(t) = \int_{t_0}^{t} \lambda_2(\tau)d\tau, \quad (43)$$

где функции $\lambda_1(t)$, $\lambda_2(t)$ выбираются как

$$\lambda_1(t) = \begin{cases} \lambda_{01} & при \quad |y_1(t)| > \delta_0 \\ 0 & при \quad |y_1(t)| \le \delta_0, \end{cases} \quad \lambda_2(t) = \begin{cases} \lambda_{02} & при \quad |y_2(t)| > \delta_0 \\ 0 & при \quad |y_2(t)| \le \delta_0, \end{cases} \quad (44)$$

а числа $\lambda_{01} > 0$, $\lambda_{02} > 0$. Для настройки σ_1 и σ_2 будем использовать алгоритм

$$\sigma_1(t) = \sigma_{01}\left(\tilde{k}_1(t)\right)^2, \ \sigma_2 = \sigma_{02}\left(\tilde{k}_2(t)\right)^2, \quad (45)$$

где числа $\sigma_{01} > 0$, $\sigma_{02} > 0$. Очевидно, что в этом случае найдутся такие моменты времени t_1, t_2, начиная с которых (41) и (42) соответственно будут выполнены.

Результаты моделирования

Рассмотрим в качестве примера следующую двухканальную систему управления с неизвестными параметрами:

$$y_1(t) = \frac{1}{s^3 + 2{,}5s^2 + 0{,}5s - 1}u_1(t) - \frac{6}{s^3 + 2{,}5s^2 + 0{,}5s - 1}y_2(t), \quad (46)$$

$$y_2(t) = \frac{6}{s^2 - s + 1}u_2(t) - \frac{6}{s^2 - s + 1}y_1(t). \quad (47)$$

Передаточные функции в первом и втором каналах имеют относительные степени $\rho_1 = 3$ и $\rho_2 = 2$ соответственно.

Алгоритм управления выбирается согласно (3)–(6):

$$u_1(t) = -\tilde{k}_1(t)\left(p^2 + p + 1\right)\xi_{11}(t), \quad (48)$$

$$\begin{cases} \dot{\xi}_{11}(t) = \sigma_1(t)\xi_{12}(t), \\ \dot{\xi}_{12}(t) = \sigma_1(t)\left(-0{,}5\xi_{11}(t) - 4\xi_{12}(t) + 0{,}5y_1(t)\right), \end{cases} \quad (49)$$

$$u_2 = -\tilde{k}_2(t)\left(p + 1\right)\xi_{21}(t), \quad (50)$$

$$\dot{\xi}_{21}(t) = \sigma_2(t)\left(-\xi_{21}(t) + y_2(t)\right), \quad (51)$$

Алгоритм адаптации выбирается в соответствии с (43)–(45) с параметрами:

$$\lambda_{01} = 25, \ \lambda_{02} = 100, \ \delta_0 = 0{,}1, \ \sigma_{01} = 0{,}015, \ \sigma_{02} = 0{,}04. \quad (52)$$

Результаты моделирования при $y_1(0) = y_2(0) = 2$ представлены на рис. 1 и рис. 2.

Таким образом, результаты моделирования подтверждают, что замкнутая система является устойчивой.

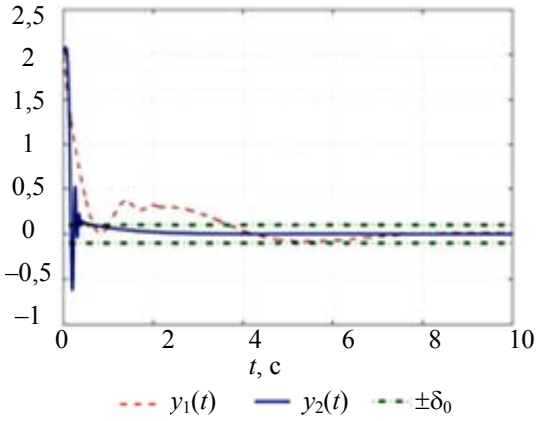
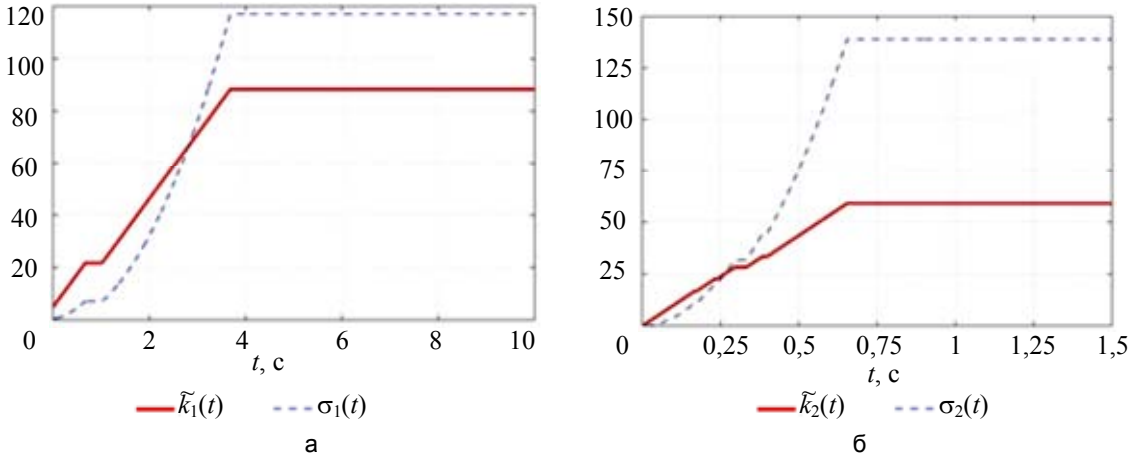


Рис. 1. Выходные переменные системы (46)–(52)



а б

Рис. 2. Результаты моделирования функций настройки параметров (43), (45) для функций: $\widetilde{k}_1(t)$, $\sigma_1(t)$ (а); $\widetilde{k}_2(t)$, $\sigma_2(t)$ (б)

Заключение

В работе рассмотрена проблема синтеза закона управления для стабилизации по выходу линейных стационарных параметрически неопределенных многоканальных систем. В предположении, что известны относительные степени каждой из подсистем, на примере объекта второго порядка было показано, что алгоритм управления «последовательный компенсатор», впервые предложенный в [21], может быть успешно применен и в случае многоканальных систем. Эффективность алгоритма управления проиллюстрирована моделированием.

Литература

1. Поляк Б.Т., Щербаков П.С. Робастная устойчивость и управление. М.: Наука, 2002. 303 с.
2. Буков В.Н. Вложение систем. Аналитический подход к анализу и синтезу матричных систем. Калуга: Изд-во научной литературы Н.Ф. Бочкаревой, 2006. 720 с.
3. Мирошник И.В., Никифоров В.О., Фрадков А.Л. Нелинейное адаптивное управление сложными динамическими системами. СПб.: Наука, 2000. 549 с.
4. Фрадков А.Л. Управление в сложных системах. М.: Наука, 1990. 296 с.
5. Цыкунов А.М. Алгоритмы робастного управления с компенсацией ограниченных возмущений // Автоматика и телемеханика. 2007. № 7. С. 103–115.
6. Фуртат И.Б. Робастная синхронизация динамической сети с компенсацией возмущений // Автоматика и телемеханика. 2011. № 12. С.104–114.
7. Фуртат И.Б. Консенсусное управление линейной динамической сетью по выходу с компенсацией возмущений // Мехатроника, автоматизация, управление. 2011. № 4. С. 12–18.
8. Фуртат И.Б. Субоптимальное управление нелинейными мультиагентыми системами // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 1 (83). С. 19–23.
9. Фуртат И.Б. Робастное управление определенным классом неминимально-фазовых динамических сетей // Известия РАН. Теория и системы управления. 2014. № 1. С. 35–48.

10. Цыкунов А.М. Адаптивное и робастное управление динамическими объектами по выходу. М.: ФИЗМАТЛИТ, 2009. 268 с.
11. Паршева Е.А. Адаптивное децентрализованное управление многосвязными объектами со скалярными входом и выходом с неминимальной реализацией эталонной модели // Автоматика и телемеханика. 2005. № 8. С. 118–127.
12. Mirkin B., Gutman P.-O. Lyapunov-based adaptive output-feedback control of MIMO nonlinear plants with unknown, time-varying state delays // Proc. 9th IFAC Workshop on Time Delay Systems. Prague, Czech Republic, 2010. Part. 1. P. 33–38.
13. Mirkin B., Gutman P.-O. Adaptive output-feedback tracking: the case of MIMO plants with unknown, time-varying state delay // Systems and Control Letters. 2009. V. 58. N 1. P. 62–68.
14. Cavallo A., Natale C. A robust output feedback control law for MIMO plants // Proc. 15th IFAC World Congress. Barcelona, Spain, 2002. V. 15, part. 1. P. 335–345.
15. Ge S.S., Li Z. Robust adaptive control for a class of MIMO nonlinear systems by state and output feedback // IEEE Transactions on Automatic Control. 2014. V. 59. N 6. P. 1624–1629.
16. Qi R., Tao G., Jiang B. Adaptive control of MIMO time-varying systems with indicator function based parametrization // Automatica. 2014. V. 50. N 5. P. 1369–1380.
17. Chen M., Ge S.S., Ren B. Adaptive tracking control of uncertain MIMO nonlinear systems with input constraints // Automatica. 2011. V. 47. N 3. P. 452–465.
18. Ambrose H., Qu Z. Model reference robust control for MIMO systems // International Journal of Control. 1997. V. 68. N 3. P. 599–623.
19. Фомин В.Н., Фрадков А.Л., Якубович В.А. Адаптивное управление динамическими объектами. М.: Наука, 1981. 448 с.
20. Narendra K.S., Annaswamy A. Stable Adaptive Systems. New Jersey: Prentice Hall, 2005. 512 p.
21. Бобцов А.А. Робастное управление по выходу линейной системой с неопределенными коэффициентами // Автоматика и телемеханика. 2002. № 11. С. 108–117.

Бобцов Алексей Алексеевич – доктор технических наук, профессор, декан факультета, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, bobtsov@mail.ru

Фаронов Максим Викторович – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, faronov_m@mail.ru

Фуртат Игорь Борисович – доктор технических наук, доцент, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; ведущий научный сотрудник, Институт проблем машиноведения РАН, Санкт-Петербург, 199178, Российская Федерация, cainenash@mail.ru

Пыркин Антон Александрович – кандидат технических наук, доцент, ведущий научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, a.pyrkin@gmail.com

Цзянь Ван – научный сотрудник, Институт автоматизации, Университет Ханчжоу Дяньцзы, Ханчжоу, 310018, Китай, wangjian@hdu.edu.cn

Aleksei A. Bobtsov – D.Sc., Professor, Dean, Department head, ITMO University, Saint Petersburg, 197101, Russian Federation, bobtsov@mail.ru

Maksim V. Faronov – postgraduate, ITMO University, Saint Petersburg, 197101,Russian Federation, faronov_m@mail.ru

Igor B. Furtat – D.Sc., Associate professor, Professor, ITMO University, Saint Petersburg, 197101, Russian Federation; leading scientific researcher, Institute of Problems of Mechanical Engineering Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, cainenash@mail.ru

Anton A. Pyrkin – PhD, Associate professor, ITMO University, Saint Petersburg, 197101, Russian Federation, a.pyrkin@gmail.com

Wang Jian – researcher, Hangzhou Dianzi University, Hangzhou, 310018, China, wangjian@hdu.edu.cn

Принято к печати 14.05.14
Accepted 14.05.14

4 НОВЫЕ МАТЕРИАЛЫ И НАНОТЕХНОЛОГИИ MATERIAL SCIENCE AND NANOTECHNOLOGIES

УДК 535.6

ЧАСТОТНЫЕ ХАРАКТЕРИСТИКИ СОВРЕМЕННЫХ СВЕТОДИОДНЫХ ЛЮМИНОФОРНЫХ МАТЕРИАЛОВ

М.С. Фудин[a], К.Д. Мынбаев[a,b], Х. Липсанен[a,c], К.Е. Айфантис[a,d], В.Е. Бугров[a], А.Е. Романов[a,b]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[b] ФТИ им. А.Ф. Иоффе, Санкт-Петербург, 194021, Российская Федерация, Karim.mynbaev@niuitmo.ru

[c] Университет Аалто, Аалто, 02150, Финляндия

[d] Университет Аризоны, Таксон, 85721, Аризона, США

Аннотация. Для оценки перспектив применения люминофорных светодиодов в системах беспроводной передачи данных в оптическом диапазоне рассмотрены частотные характеристики современных светодиодных люминофорных материалов. Проведены измерения зависимости интенсивности излучения одиночных светодиодов и светодиодных сборок с люминофорами на основе иттрий-алюминиевого и лютеций-алюминиевого граната (в том числе с добавлением нитридного люминофора), а также силикатных люминофоров, от частоты электрических импульсов, возбуждающих излучение светодиодов. Показано, что с точки зрения скорости передачи информации люминофоры на основе гранатов (в том числе с добавлением нитридных люминофоров) имеют бóльший потенциал, чем силикатные люминофоры. Материалы на основе гранатов могут быть использованы в оптических системах передачи данных с полосой пропускания (без дополнительной модуляции) до 3 МГц (в одночиповых светодиодах) и до 4,5 МГц (в 9-чиповых сборках). Результаты работы показывают, что значительная часть светодиодов, применяемых в системах общего освещения, уже сейчас может быть использована для передачи информации пользователям, например, в системах позиционирования в закрытых пространствах, для облегчения поиска нужных помещений и объектов и т.п.

Ключевые слова: белые светодиоды, передача информации, оптический диапазон, люминофоры, полоса пропускания.

Благодарности. Работа выполнена при государственной финансовой поддержке, выделяемой на реализацию Программы развития международных научных подразделений Университета ИТМО. Авторы благодарны Л.А. Никулиной за предоставленные образцы светодиодов.

FREQUENCY CHARACTERISTICS OF MODERN LED PHOSPHOR MATERIALS

M.S. Fudin[a], K.D. Mynbaev[a,b], H. Lipsanen[a,c], K.E. Aifantis[a,d], V.E. Bougrov[a], A.E. Romanov[a,b]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation

[b] Ioffe Institute, Saint Petersburg, 194021, Russian Federation, Karim.mynbaev@niuitmo.ru

[c] Aalto University, Aalto, 02150, Finland

[d] University of Arizona, Tucson, 85721, Arizona, USA

Abstract. Frequency characteristics of modern LED phosphor materials have been considered for the purpose of assessing the prospects of phosphor-based LEDs in wireless communication data systems which use optical wavelengths. The measurements have been carried out on the dependence of the emission intensity of single LEDs and LED chip-on-board modules with phosphors based on yttrium-aluminum and lutetium-aluminum garnets (with or without addition of nitride-based phosphors) as well as silicate-based phosphors, on the frequency of electric pulses exciting the emission. It was shown that from the point of view of data transmission rate, garnet-based phosphors (including systems with added nitride phosphors) are more promising than silicate–based ones. Garnet-based materials can be used in optical communication data systems with bandwidth (without extra modulation applied) up to 3 MHz with single–chip LEDs and up to 4.5 MHz with 9-chip LED chip-on-board modules. The results of the work indicate that a significant part of white LEDs used in general lighting systems can be even now used for data transfer, for example, in systems assisting positioning in closed spaces to facilitate people searching necessary rooms or objects.

Keywords: white LEDs, data transfer, optical wavelengths, phosphors, bandwidth.

Acknowledgements. This work was financially supported by the Russian Government via funds allotted for the implementation of the Program of International Scientific Departments development at ITMO University. The authors are thankful to L.A. Nikulina for supplying them with the samples of LEDs.

Введение

В последнее время активно развивается интерес к технологиям передачи данных при помощи видимого света (Visible Light Communication, VLC), использующим диапазон длин волны 375–780 нм. Основной задачей, решаемой в настоящее время разработчиками технологии VLC, является увеличение скорости передачи данных. Для этого используются различные аппаратные и программные средства [1, 2], рекорды скорости постоянно обновляются, и в идеальных условиях с использованием одиночного монохромного (длина волны 450 нм) светодиодного микрочипа на основе нитрида галлия (GaN) уже превышена скорость в 3 Гбит/с [3]. Однако одним из наиболее перспективных направлений технологий VLC является применение светодиодов одновременно и для передачи данных, и для освещения, для чего необ-

ходимо использовать светодиоды белого света. При использовании белого света максимальные скорости достигаются с помощью излучения, генерируемого так называемыми RGB-светодиодами, когда модуляция сигнала осуществляется раздельно по каждому из трех (R,G и B) каналов при средней скорости в 15 Мбит/с на канал. Это решение эффективно, но является весьма дорогостоящим. В то же время представляется, что многие задачи VLC можно решать с использованием уже существующих промышленных светодиодов, выпускаемых для систем общего освещения и обладающих низкой себестоимостью. Поскольку в системе «полупроводниковый чип–люминофор», используемой в большинстве современных белых светодиодов, наиболее инерционным является люминофор (так, стандартные синие чипы на основе GaN обеспечивают скорости передачи в 20 Мбит/с, в то время как добавление люминофора, как считается, снижает скорость на порядок, до 2 Мбит/с [4]), то для решения этих задач необходимо иметь представление о частотных характеристиках современных светодиодных люминофорных материалов.

Проблема заключается в том, что многие люминофорные материалы исходно создавались таким образом, чтобы обеспечить максимально длительное время высвечивания (с учетом эффекта насыщения люминофора), – в частности, чтобы скомпенсировать эффект мерцания из-за пульсаций тока в сетях питания светодиодов, – и, таким образом, по определению обладают низкой частотой переключения [5]. Время жизни люминесценции ионов церия в иттрий-алюминиевом гранате ($Y_3Al_5O_{12}:Ce^{3+}$, наиболее распространенный люминофор для белых светодиодов) составляет около 65 нс [6], а типичное время высвечивания у большинства светодиодных люминофоров составляет несколько миллисекунд. Существуют люминофоры, излучающие свет в течение минут и даже часов после выключения возбуждения [7]. Естественно, что время высвечивания зависит от физических процессов в люминофоре, – типа используемых излучательных переходов, эффективности передачи возбуждения от матрицы к люминесцентному центру, наличия ловушек и т.п. В белых светодиодах сейчас в основном используются внутри- и межоболочечные переходы в ионах редкоземельных металлов (Ce, Eu и др.), внедренных в матрицу из граната, но набирают популярность и люминофоры, использующие другие ионы (в частности, Mn, Cu, Co), а также другие типы матриц. Сведения о частотных характеристиках подобных материалов и о потенциальных возможностях их применения в системах VLC соответственно в литературе отсутствуют: разработчики этих систем в своих экспериментах используют монохромные светодиоды без люминофора, отфильтровывают длины волн, излучаемые люминофором, или ограничиваются общим понятием «белый светодиод» (см., например, [2, 8–14]). В настоящей работе мы сообщаем о результатах исследования частотных характеристик ряда современных люминофорных материалов, используемых при разработке и производстве светодиодов, и кратко обсуждаем перспективы использования этих материалов в системах VLC.

Экспериментальная часть

В работе исследовались системы, в которых люминофорные материалы были использованы в маломощных светодиодах SMD серии OLP (номинальный ток 20 мА) и мощных 9-чиповых светодиодных сборках X10 серии OCC (номинальный ток 1050 мА) производства компании «Оптоган»[1]. Источником электрических сигналов, возбуждавших излучение светодиодов, служил генератор Agilent 33522, выдававший импульсы прямоугольной формы. При исследовании мощных светодиодов в схему питания дополнительно включался линейный источник питания Matrix MPS-6003LK-2. Для регистрации сигнала использовался кремниевый фотодиод, подключенный к мультиметру MAST MY-65. Расстояние от светодиода до фотодиода при измерениях не превышало 10 мм. Дополнительная модуляция сигнала не проводилась.

В одночиповых маломощных светодиодах использовались следующие люминофоры: традиционный на основе иттрий-алюминиевого граната (ИАГ), легированного ионами церия $Y_3Al_5O_{12}:Ce^{3+}$; ИАГ в комбинации с $AlCaSrClN_3Si:Eu^{2+}$; люминофор на основе лютеций-алюминиевого граната (ЛАГ) $Lu_3Al_3O_{12}:Ce^{3+}$, ЛАГ в комбинации с $AlCaSrClN_3Si:Eu^{2+}$; ЛАГ в комбинации с $(Ba,Ca,Sr)SiO_4:Eu^{2+}$; силикатные люминофоры $CaSiO_4:Eu^{2+}$, $BaSiO_4:Eu^{2+}$, $SrSiO_4:Eu^{2+}$. Также были проведены измерения частотных характеристик светодиода с прозрачным силиконовым эластомером без частиц люминофора. В мощных светодиодных сборках X10 использовались люминофоры $Y_3Al_5O_{12}:Ce^{3+}$, $Lu_3Al_3O_{12}:Ce^{3+}$ и $Lu_3Al_3O_{12}:Ce^{3+}$ в комбинации с $AlCaSrClN_3Si:Eu^{2+}$.

Результаты и обсуждение

На рис. 1 представлена амплитудно-частотная характеристика (АЧХ) одночипового светодиода с силикатным люминофором $SrSiO_4:Eu^{2+}$. Как видно, характеристика имеет постоянный участок до частоты 2×10^5 Гц, после чего наблюдается довольно резкий спад. Спад до уровня в 3 дБ от исходного соответствовал частоте 3,2 МГц, что означало, что без дополнительной модуляции светодиод с данным люминофором позволял бы передавать данные со скоростью 3,2 Мбит/с (в идеальных условиях). Данное

[1] Наноэкопром. Светодиодные компоненты [Электронный ресурс]. Режим доступа: http://nanoecoprom.com/catalog/open/63 свободный. Яз. рус. (дата обращения 06.10.2014).

значение оказалось максимальным для исследованных люминофорных материалов, использованных в одночиповых светодиодах.

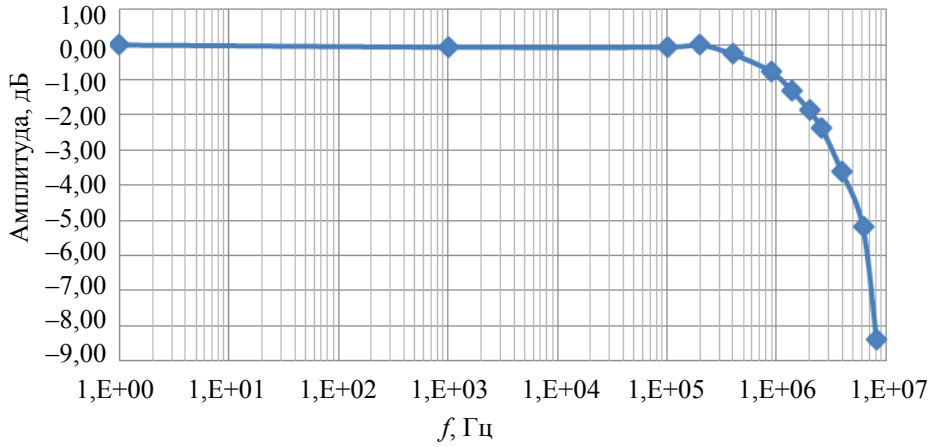


Рис. 1. АЧХ одночипового светодиода серии OLP с силикатным люминофором $SrSiO_4:Eu^{2+}$

На рис. 2 представлена АЧХ мощной светодиодной 9-чиповой сборки X10 с люминофором $Lu_3Al_3O_{12}:Ce^{3+}$. Здесь АЧХ демонстрирует постоянный спад, уровню спада в 3 дБ от исходного соответствовала частота 4,5 МГц. Это значение оказалось максимальным для исследованных люминофорных материалов, примененных в мощных светодиодных сборках.

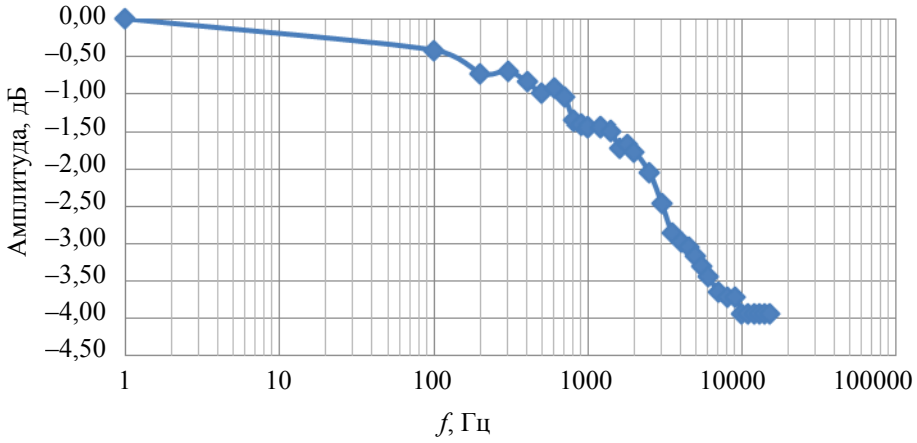


Рис. 2. АЧХ светодиодной сборки X10 с люминофором $Lu_3Al_3O_{12}:Ce^{3+}$

На рис. 3 представлены АЧХ светодиодов (одночипового и сборки X10) с люминофором $Lu_3Al_3O_{12}:Ce^{3+}$ в комбинации с нитридным люминофором $AlCaSrClN_3Si:Eu^{2+}$. Целью добавления нитридного люминофора является, как известно, получение более «теплого» белого света, востребованного в освещении жилых помещений, путем введения в «холодное» свечение ЛАГ красной (длина волны 580–650 нм) составляющей [15]. Для сравнения на рисунке также приведена АЧХ одночипового светодиода, в котором в силиконовый эластомер люминофор не добавлялся. В последнем случае и было получено максимальное значение ширины полосы пропускания. Оно составило 6,8 МГц, данная величина использовалась как референтная при интерпретации измерений, проведенных с использованием люминофоров. Как видно, при добавлении в эластомер люминофора частота среза АЧХ заметно уменьшается, причем для маломощного одночипового светодиода это уменьшение значительно сильнее, чем для светодиодной сборки. Ширины полос АЧХ всех исследованных светодиодов приведены в таблице.

Таким образом, для исследованных люминофорных материалов в одночиповых светодиодах ширина полосы пропускания варьировалась от 1,8 МГц до 3,2 МГц, причем максимальный результат был достигнут при использовании светодиода со следующими люминофорами: силикатный $SrSiO_4:Eu^{2+}$, ЛАГ $Lu_3Al_3O_{12}:Ce^{3+}$ и люминофор на ИАГ с примесью нитрида $Y_3Al_5O_{12}:Ce^{3+}+AlCaSrClN_3Si:Eu^{2+}$. Для светодиодных сборок X10 ширина полосы пропускания варьировалась от 3,5 МГц до 4,5 МГц. Наилучший результат был достигнут при использовании светодиодной сборки с люминофором на $Lu_3Al_3O_{12}:Ce^{3+}$, что соответствовало результатам, полученным при измерении одночиповых светодиодов.

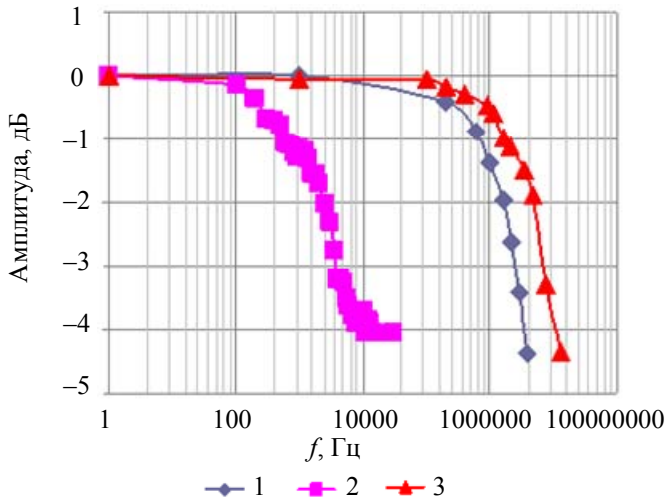


Рис. 3. АЧХ светодиодов (сборки X10 (1) и одночипового OLP (2)) с люминофором $Lu_3Al_3O_{12}:Ce^{3+}$ в комбинации с $AlCaSrClN_3Si:Eu^{2+}$ и АЧХ OLP светодиода с силиконовым эластомером без частиц люминофора (3)

Люминофор	Ширина полосы пропускания, МГц	
	Светодиод OLP	Сборка X10
Без люминофора	6,8	–
$Y_3Al_5O_{12}:Ce^{3+}$	2,7	3,5
$Y_3Al_5O_{12}:Ce^{3+}$ с $AlCaSrClN_3Si:Eu^{2+}$	3,2	–
$Lu_3Al_3O_{12}:Ce^{3+}$	3,2	4,5
$Lu_3Al_3O_{12}:Ce^{3+}$ с $AlCaSrClN_3Si:Eu^{2+}$	2,2	3,7
$Lu_3Al_3O_{12}:Ce^{3+}$ с $(Ba,Ca,Sr)SiO_4:Eu^{2+}$	3,0	–
$BaSiO_4:Eu^{2+}$	2,5	–
$SrSiO_4:Eu^{2+}$	3,2	–
$CaSiO_4:Eu^{2+}$	1,8	–

Таблица. Ширина полосы АЧХ исследованных люминофоров

Анализируя полученные данные, отметим, что в целом для люминофоров на основе гранатов, легированных церием, частотные характеристики оказались лучше, чем для силикатных люминофоров, легированных европием. Высокое быстродействие люминофорных смесей на основе легированных церием ЛАГ хорошо известно [16]. Добавление силикатных люминофоров с европием к ЛАГ предсказуемо уменьшило полосу пропускания, то же самое касается и добавления к ЛАГ нитридного люминофора. Силикатные люминофоры, легированные европием, известны относительно длительным временем высвечивания, это позволяет использовать их в приложениях, требующих долговременной фосфоресценции люминофорного материала [17]. При этом различие в частотных характеристиках люминофоров различного химического состава, но использующих одни и те же ионы-активаторы, объясняется наличием разных типов ловушек, способных захватывать носители и «задерживать» процесс передачи возбуждения люминесцентным центрам [17]. Результат, согласно которому добавление нитридного люминофора к ИАГ несколько улучшило частотные характеристики получившейся люминофорной смеси, требует дополнительной проверки.

Частотные характеристики мощных светодиодных сборок оказались лучше характеристик маломощных светодиодов, использующих аналогичные люминофорные смеси. Этот результат также ожидаем, поскольку известно, что в экспериментах, подобных проведенным в нашей работе, чем выше интенсивность света, попадающего на фотоприемник, тем больше скорость передачи данных [18]. Для монохромных безлюминофорных светодиодов тенденция является обратной – уменьшение размеров (и соответственно мощности) светодиодных чипов позволяет снизить паразитную емкость и импеданс приборной структуры и повысить частоту модуляции излучения [3].

Полученные в работе результаты впервые позволяют разработать рекомендации по возможному использованию современных светодиодных люминофорных материалов в системах передачи данных,

использующих оптический диапазон. Как отмечалось выше, до сегодняшнего дня существовало обобщенное мнение о том, что люминофорные белые светодиоды имеют АЧХ, ограниченную полосой 2 МГц, что делало перспективы их применения в системах VLC весьма туманными [3, 12, 19, 20]. Проведенное авторами исследование показало, что цифра в 2 МГц справедлива лишь в отношении классического ИАГ, легированного церием, и для некоторых силикатных люминофоров. При использовании ЛАГ полоса пропускания может быть увеличена до 3 МГц, а применение в системах освещения вместо маломощных одночиповых светодиодов мультичиповых сборок позволяет увеличить ширину полосы АЧХ до 4,5 МГц, что вполне достаточно для многих современных приложений – в системах позиционирования в закрытых помещениях, на транспорте и т.д. Разумеется, при использовании светодиодов общего освещения для передачи информации необходимо проведение исследований влияния частоты и способа модуляции на качество белого света, однако подобные эксперименты нужно проводить уже при использовании дополнительных схем модуляции сигнала, применяемых для расширения полосы АЧХ и коррекции ошибок в системах VLC.

Заключение

Рассмотрены частотные характеристики современных светодиодных люминофорных материалов, применяющихся при разработке как одиночных белых светодиодов, так и светодиодных сборок. Показано, что с точки зрения скорости передачи информации люминофоры на основе гранатов (в том числе с добавлением нитридных люминофоров) более перспективны, чем силикатные люминофоры. Материалы на основе гранатов могут быть использованы в оптических системах передачи данных с полосой пропускания (без дополнительной модуляции) до 3 МГц в одночиповых светодиодах и до 4,5 МГц в 9-чиповых сборках, что достаточно для многих бытовых приложений.

Литература

1. Elgala H., Mesleh R., Haas H. Indoor broadcasting via white LEDs and OFDM // IEEE Transactions on Consumer Electronics. 2009. V. 55. N 3. P. 1127–1134.
2. Zhang H., Yuan Y., Xu W. PAPR reduction for DCO-OFDM visible light communications via semidefinite relaxation // IEEE Photonics Technology Letters. 2014. V. 26. N 17. P. 1718–1721.
3. Tsonev D., Hyunchae Chun, Rajbhandari S., McKendry J.J.D., Videv S., Gu E., Haji M., Watson S., Kelly A.E., Faulkner G., Dawson M.D., Haas H., O'Brien D. A 3-Gb/s single-LED OFDM-based wireless VLC link using a gallium nitride μLED // IEEE Photonics Technology Letters. 2014. V. 26. N 7. P. 637–640.
4. Grubor J., Randel S., Langer K.-D., Walewski J.W. Broadband information broadcasting using LED-based interior lighting // Journal of Lightwave Technology. 2008. V. 26. N 24. P. 3883–3892.
5. Smet P.F., Parmentier A.B., Poelman D. Selecting conversion phosphors for white light-emitting diodes // Journal of the Electrochemical Society. 2011. V. 158. N 6. P. R37–R54.
6. Bachmann V., Ronda C., Meijerink A. Temperature quenching of yellow Ce^{3+} luminescence in YAG:Ce // Chemistry of Materials. 2009. V. 21. N 10. P. 2077–2084.
7. van den Eeckhout K., Poelman D., Smet P.F. Persistent luminescence in non-Eu^{2+}-doped compounds: a review // Materials. 2013. V. 6. N 7. P. 2789–2818.
8. Jovicic A., Li J., Richardson T. Visible light communication: opportunities, challenges and the path to market // IEEE Communications Magazine. 2013. V. 51. N 12. P. 26–32.
9. Feng L.-F., Li Y., Li D., Wang C.-D., Zhang G.-Y., Yao D.-S., Liu W.-F., Xing P.-F. Frequency response of modulated electroluminescence of light-emitting diodes // Chinese Physics Letters. 2011. V. 28. N 10. Art. 107801.
10. McKendry J.J.D., Massoubre D., Zhang S., Rae B.R., Green R.P., Gu E., Henderson R.K., Kelly A.E., Dawson M.D. Visible-light communications using a CMOS-controlled micro-light-emitting-diode array // Journal of Lightwave Technology. 2012. V. 30. N 1. P. 61–67.
11. Wu Y., Yang A., Feng L., Zuo L., Sun Y.-N. Modulation based cells distribution for visible light communication // Optics Express. 2012. V. 20. N 22. P. 24196–24208.
12. Khalid A.M., Cossu G., Corsini R., Choudhury P., Ciaramella E. 1-Gb/s Transmission over a phosphorescent white LED by using rate-adaptive discrete multitone modulation // IEEE Photonics Journal. 2012. V. 4. N 5. P. 1465–1473.
13. Das P., Park Y., Kim K.-D. Performance of color-independent OFDM visible light communication based on color space // Optics Communications. 2014. V. 324. P. 264–268.
14. Sung J.-Y., Chow C.-W., Yeh C.-H. Is blue optical filter necessary in high speed phosphor-based white light LED visible light communications? // Optics Express. 2014. V. 22. N 17. P. 20646–220651.
15. Асеев В.А., Колобкова Е.В., Некрасова Я.А., Никоноров Н.В., Рохмин А.С. Люминесценция марганца во фторфосфатных стеклах // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 6 (82). С. 36–39.

16. Babin V., Bichevin V., Gorbenko V., Kink M., Makhov A., Maksimov Y., Nikl M., Stryganyuk G., Zazubovich S., Zorenko Y. Time-resolved spectroscopy of exciton-related states in single crystals and single crystalline films of $Lu_3Al_5O_{12}$ and $Lu_3Al_5O_{12}$:Ce // Physica Status Solidi (B) Basic Research. 2011. V. 248. N 6. P. 1505–1512.
17. Van den Eeckhout K., Smet P.F., Poelman D. Persistent luminescence in Eu^{2+}-doped compounds: a review // Materials. 2010. V. 3. N 4. P. 2536–2566.
18. Grobe L., Paraskevopoulos A., Hilt J., Schulz D., Lassak F., Hartlieb F., Kottke C., Jungnickel V., Langer K.-D. High-speed visible light communication systems // IEEE Communications Magazine. 2013. V. 51. N 12. P. 60–66.
19. Komine T., Nakagawa M. Fundamental analysis for visible-light communication system using LED lights // IEEE Transactions on Consumer Electronics. 2004. V. 50. N 1. P. 100–107.
20. Vucic J., Kottke C., Nerreter S., Langer K.-D., Walewski J.W. 513 Mbit/s visible light communications link based on DMT modulation of a white LED // Journal of Lightwave Technology. 2010. V. 28. N 24. P. 3512–3518.

Фудин Максим Сергеевич – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Wizmaks1991@mail.ru

Мынбаев Карим Джафарович – доктор физико-математических наук, профессор, Санкт-Петербург, 197101, Российская Федерация; заведующий лабораторией, ФТИ им. А.Ф. Иоффе, Санкт-Петербург, 194021, Российская Федерация, Karim.mynbaev@niuitmo.ru

Липсанен Харри – D.Sci., профессор, профессор-исследователь, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; профессор, Университет Аалто, Аалто, 02150, Финляндия, harri.lipsanen@aalto.fi

Айфантис Катерина – PhD, профессор-исследователь, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; профессор, Университет Аризоны, Таксон, 85721, Аризона, США, aifantis@email.arizona.edu

Бугров Владислав Евгеньевич – доктор физико-математических наук, заведующий кафедрой, Санкт-Петербург, 197101, Российская Федерация, Vladislav.bougrov@niuitmo.ru

Романов Алексей Евгеньевич – доктор физико-математических наук, профессор, заведующий кафедрой, Санкт-Петербург, 197101, Российская Федерация; главный научный сотрудник, ФТИ им. А.Ф. Иоффе, Санкт-Петербург, 194021, Российская Федерация, Alexey.romanov@niuitmo.ru

Maxim S. Fudin – student, ITMO University, Saint Petersburg, 197101, Russian Federation, Wizmaks1991@mail.ru

Karim D. Mynbaev – D.Sc., Professor, ITMO University, Saint Petersburg, 197101, Russian Federation; Head of laboratory, Ioffe Institute, Saint Petersburg, 194021, Russian Federation, Karim.mynbaev@niuitmo.ru

Harri Lipsanen – D.Sc., Professor, Professor-researcher, ITMO University, Saint Petersburg, 197101, Russia; Professor, Aalto University, Aalto, 02150, Finland, harri.lipsanen@aalto.fi

Katerina E. Aifantis – PhD, Professor-researcher, ITMO University, Saint Petersburg, 197101, Russian Federation; Professor, University of Arizona, Tucson, 85721, Arizona, USA, aifantis@email.arizona.edu

Vladislav E. Bougrov – D.Sc., Department head, ITMO University, Saint Petersburg, 197101, Russian Federation, Vladislav.bougrov@niuitmo.ru

Alexei E. Romanov – D. Sc., Professor, Department head, ITMO University, Saint Petersburg, 197101, Russian Federation; leading scientific researcher, Ioffe Institute, Saint Petersburg, 194021, Russian Federation, Alexey.romanov@niuitmo.ru

Принято к печати 01.09.14
Accepted 01.09.14

УДК 535.343, 539.213.27

СПЕКТРАЛЬНО-ЛЮМИНЕСЦЕНТНЫЕ ХАРАКТЕРИСТИКИ ФТОРОФОСФАТНЫХ СТЕКОЛ, АКТИВИРОВАННЫХ МАРГАНЦЕМ И КВАНТОВЫМИ ТОЧКАМИ СУЛЬФИДА КАДМИЯ

Ж.О. Липатова[a], В.А. Асеев[a], Е.В. Колобкова[a]

[a]Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, zluka_yo@mail.ru

Аннотация. Разработка и исследование люминофоров на основе квантовых точек является перспективной задачей фотоники. Введение квантовых точек во фторофосфатные стекла обеспечивает жесткую матрицу, высокий коэффициент поглощения, широкую полосу люминесценции и высокий квантовый выход. Поскольку ионы марганца обладают интенсивной полосой люминесценции в красной области спектра, то добавление их в стекла с квантовыми точками приводит к уширению спектра в длинноволновую область спектра. Такое излучение по своему спектральному составу ближе к естественному солнечному излучению и обеспечивает источникам излучения на его основе более высокий индекс цветопередачи. Целью работы является исследование спектрально-люминесцентных свойств фторофосфатных стекол, активированных марганцем и квантовыми точками CdS. Для этого были синтезированы фторофосфатные стекла состава $47NaPO_3$-$30H_3PO_4$-$10Ga_2O_3$-5ZnO-*x*MnS-7,5$NaALF_6$-4,2CdS, где *x*=3, 6, 8 мол.%. Квантовые точки были получены путем вторичной термообработки стекол при температуре 430 °С в течение 90 мин. Измерены спектры поглощения в видимой области спектра (300–600 нм). Показано смещение края фундаментального поглощения в видимую область спектра при термообработке. Данное изменение обусловлено ростом квантовых точек. Экспериментально установлено, что при возбуждении лазером на длине волны 410 нм максимумы интенсивности люминесценции сдвигаются в красную область спектра (с 620 нм до 660 нм). Максимальный сдвиг наблюдался у образца с концентрацией марганца 3 мол.%, минимальный сдвиг – у образца с концентрацией 8 мол.%. По результатам измерений кинетики затухания люминесценции марганца (620 нм) были получены значения времен жизни от 18 мс для образца с концентрацией MnS 3 мол.% до 15 мс – для MnS 8 мол.%. Уменьшение времени жизни с ростом концентрации связано с концентрационным тушением люминесценции марганца. Рост квантовых точек CdS при термообработке приводит к снижению значений времен жизни люминесценции до значений 9–3 мс (3 и 8 мол.% MnS соответственно). Результаты проведенных исследований показали, что данные фторофосфатные стекла, активированные марганцем и квантовыми точками CdS, являются перспективными для применения в качестве люминофоров для белых светодиодов.

Ключевые слова: квантовые точки, фторофосфатные стекла, CdS, люминофоры.

Благодарности. Работа выполнена при государственной финансовой поддержке Российского научного фонда (Соглашение № 14-23-00136).

SPECTRAL-LUMINESCENT CHARACTERISTICS OF FLUOROPHOSPHATE GLASSES ACTIVATED WITH MANGANESE AND CADMIUM SULPHIDE QUANTUM DOTS

Zh.O. Lipatova[a], V.A. Aseev[a], E.V. Kolobkova[a]

[a]ITMO University, Saint Petersburg, 197101, Russian Federation, zluka_yo@mail.ru

Abstract. Research and development of phosphors based on quantum dots (QD) is a perspective problem of photonics. The main advantages of fluorophosphate glass with quantum dots are: high absorption coefficient, solid matrix and a broad band luminescence with high quantum efficiency of QD. Manganese ions have an intense band luminescence in the red region of the spectrum. Thus, the addition of manganese ions in the glass with quantum dots leads to a broadening of the spectrum in the long wavelength region. Such emission is closer to natural sunlight and has a high color rendering index. The work objective is the study of the spectral and luminescent properties of fluorophosphate glasses doped with manganese and CdS quantum dots. Fluorophosphate glasses ($47NaPO_3$-$30H_3PO_4$-$10Ga_2O_3$-5ZnO-xMnS-7,5$NaALF_6$-4,2CdS, where x = 3, 6, 8 mol. %) were synthesized. The secondary heat treatment at the temperature of 430 ° C for 90 minutes has led to the growth of quantum dots in glass volume. Absorption spectra have been measured in the visible range (from 300 to 600 nm). Heat treatment has led to a shift of the fundamental absorption edge in the visible region of the spectrum. This change is due to the growth of quantum dots. Maximum intensity of luminescence is shifted to the red region of the spectrum from 620 nm to 660 nm under laser excitation at 410 nm. The maximum shift was observed in the glass with a concentration of 3 mol. % of manganese, the minimum one - in the glass with a concentration of 8 mol. %. Values of manganese ions lifetime from18 ms for a sample with a concentration of MnS 3 mol. % to15 ms for MnS 8 mol % were obtained. The decrease in the lifetime with concentration increasing of manganese ions is due to the concentration quenching of the luminescence. The growth of CdS quantum dots in the heat treatment leads to a decrease of the lifetimes to the values below 9-3 ms (3 and 8 - mol. % MnS, respectively). Obtained findings prove that fluorophosphate glasses doped with manganese and CdS quantum dots are perspective materials for phosphors in white LEDs.

Keywords: quantum dots, fluorophosphate glasses, CdS, phosphors.

Acknowledgements. This work was financially supported by the Russian Scientific Foundation (Agreement №14-23-00136).

Введение

В настоящее время перспективной областью в фотонике является разработка и исследование люминофоров. Одним из перспективных материалов являются полупроводниковые наночастицы, называемые квантовыми точками (КТ), размеры которых составляют от 1 до 20 нм. Благодаря эффекту размерного квантования можно целенаправленно управлять оптическими и электрическими параметрами КТ, что помогает создать на их основе структурированные наноматериалы с требуемыми оптическими свойствами [1, 2]. КТ

на основе полупроводников представляют практический интерес в качестве люминесцирующих материалов. КТ обладают широкой полосой возбуждения и излучают различные цвета в зависимости от их размера и природы полупроводника. Таким образом, можно получить КТ с любой длиной волны люминесценции – от ультрафиолетового до ближнего инфракрасного диапазона. Сульфид кадмия (CdS) является широкозонным полупроводником и люминесцирует в видимой области спектра [3].

Одним из методов получения КТ является метод коллоидного синтеза. Он подразумевает синтез КТ из жидкой фазы. Недостатком способа является низкий квантовый выход флуоресценции за счет дефектности поверхности нанокристаллов, что приводит к появлению энергетических уровней, лежащих внутри запрещенной зоны [4–6]. Также возникают трудности при введении КТ в твердые тела.

Чтобы избежать этих недостатков, предлагается получение КТ в стекле. Это более простой и экономичный способ. Синтез КТ происходит в результате высокотемпературной термообработки, обеспечивая небольшой разброс КТ по размерам. Изменение температуры и времени термообработки позволяет управлять размерами КТ и за счет этого сдвигать полосы люминесценции [7–9]. Люминофоры на базе КТ обладают более высокой эффективностью излучения.

В качестве матрицы в основном выбираются силикатные и фторофосфатные стекла. Последние более перспективны, так как в них можно ввести CdS в большей концентрации, обеспечивая тем самым более эффективные оптические свойства. Такие стекла имеют высокий коэффициент поглощения, бóльшую интенсивность люминесценции, больший квантовый выход и больший сдвиг Стокса, чем у силикатных стекол, так как температурно-временные режимы формирования квантовых точек у фторофосфатных стекол более мягкие [10–12].

Ионы марганца обладают широкой полосой люминесценции в красной области спектра. Добавление ионов марганца в стекло с КТ дает возможность сместить интегральные спектры люминесценции в длинноволновую область. Такое излучение по своему спектральному составу ближе к естественному солнечному излучению. При применении данных стекол в качестве люминофоров это даст возможность получить более низкие цветовые температуры и более высокий индекс цветопередачи [13, 14].

Целью работы является исследование спектрально-люминесцентных свойств фторофосфатных стекол, активированных марганцем и КТ CdS.

Методическая часть

Были исследованы фторофосфатные стекла следующего состава: $47NaPO_3$-$30H_3PO_4$-$10Ga_2O_3$-5ZnO-xMnS-7,5$NaALF_6$-4,2CdS, где x=3, 6, 8 мол.%. Для синтеза стекол применялись марки «ХЧ» и «ОСЧ». Синтез проводился в течение 40 мин в закрытых стеклоуглеродных тиглях в атмосфере аргона при температуре T=950–1000 °С. Навеска составляла 50 г. Был произведен отжиг при температуре несколько ниже температуры стеклования для снятия остаточных напряжений. После отжига толщины образцов не превышали 2 мм.

КТ синтезировались путем вторичной термообработки при температуре 430 °С. Термообработка производилась поэтапно через каждые 15 мин. Суммарное время термообработки составило 90 мин.

Были измерены спектры поглощения на спектрофотометре Varian Cary 500 в видимой области спектра 300–600 нм (шаг 0,1 нм, время интеграции 0,5 с). Спектры люминесценции были измерены с помощью лазера Solar Laser Systems LQ 529 B (длина волны возбуждения 410 нм), регистрировались с помощью монохроматора (длина волны 620 нм), фотоэлектронного умножителя и цифрового синхронного усилителя. Измерения проводились в видимой области спектра 400–800 нм.

Кинетика затухания люминесценции измерялась с помощью лазера Solar Laser Systems LQ 529 B на длине волны возбуждения 410 нм, монохроматора (длина волны 620 нм), фотоэлектронного умножителя и подключенного на выходе осциллографа. Все измерения проводились при комнатной температуре.

Эксперимент

Рассмотрим спектры поглощения, представленные на рис. 1. В спектре исходных стекол (рис. 1, а) наблюдается полоса поглощения с максимумом на 410 нм, связанная с переходом иона Mn^{2+} $^6A_1(^6S)\rightarrow{}^4T_2(^4G)$ [15]. Край фундаментального поглощения у исходных стекол находится в ультрафиолетовой области спектра (310 нм). Видно (рис. 1, а, вставка), что при росте концентрации марганца пропорционально увеличивается коэффициент поглощения. Термообработка приводит к сдвигу края фундаментального поглощения в видимую область спектра (440 нм). Увеличение коэффициента поглощения свидетельствует о росте размера КТ CdS. Максимальная величина сдвига наблюдается для образца с концентрацией сульфида марганца (MnS) 3 мол.% и составляет 140 нм, минимальная – для образца с MnS 8 мол.% и составляет 115 нм (рис. 1, б).

Влияние термообработки на спектры поглощения представлено на рис. 2. На данных спектрах наблюдаются пики, обусловленные увеличением размера КТ. Так как для исходных стекол коэффициент поглощения при максимальной концентрации MnS 8 мол.% равен 0,3 $см^{-1}$, а для квантовых точек он превышает величину 50 $см^{-1}$, то полосы поглощения марганца перекрываются, и их не видно. Также сечение

поглощения КТ CdS больше сечения марганца, поэтому происходит сдвиг положения экситонного максимума в длинноволновую область. Максимумы поглощения сдвигаются с 409 нм до 422 нм. После термообработки в течение 75 мин край спектров поглощения перестает сдвигаться.

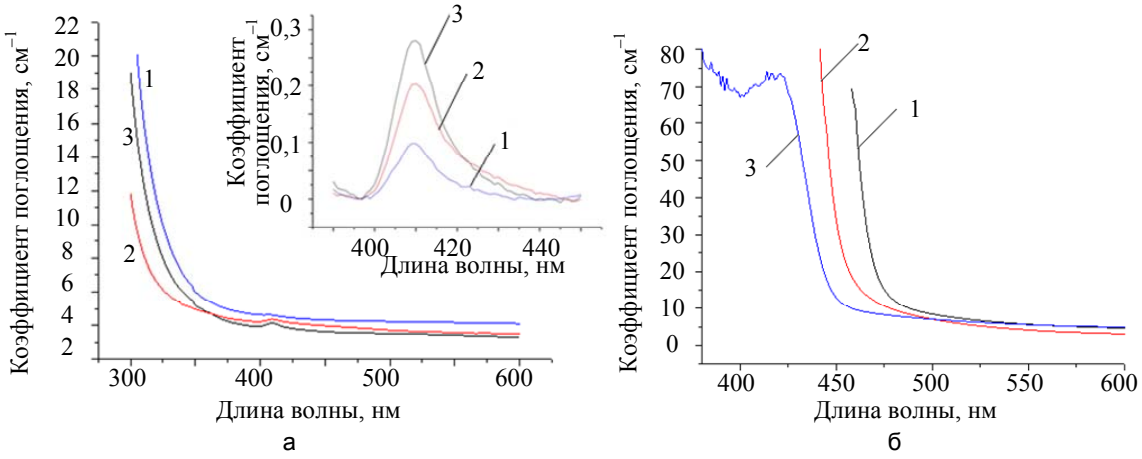


Рис. 1. Спектры поглощения: исходных образцов (вставка – полосы поглощения MnS для перехода $^6A_1(^6S) \rightarrow {}^4T_2(^4G)$) (а); после термообработки в течение 90 мин при различном содержании MnS (б): 1 – 3 мол.%; 2 – 6 мол.%; 3 – 8 мол.%

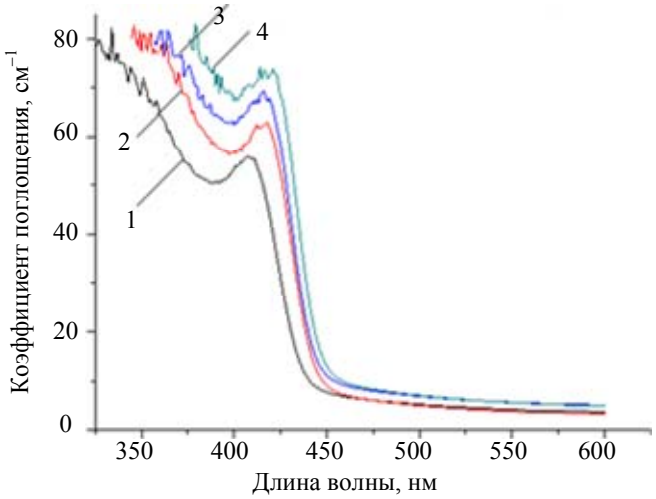


Рис. 2. Спектры поглощения для образца MnS 8 мол.% при различных термообработках: 1 – в течение 30 мин; 2 – в течение 45 мин; 3 – в течение 60 мин; 4 – в течение 75 мин

Рассмотрим спектры люминесценции для образцов с различной концентрацией марганца. В исходных образцах КТ отсутствуют, поэтому наблюдается только люминесценция марганца (рис. 3, а). Максимум интенсивности приходится на область 620 нм. После термообработки в течение 90 мин наблюдается суммарная люминесценция КТ CdS и марганца (рис. 3, б). В итоге видно уширение полосы на 50 нм. Максимальный сдвиг полосы наблюдается для образца MnS (3 мол.%), минимальный сдвиг – для образца MnS (8 мол.%), что характерно и для спектров поглощения. Это обусловлено увеличением размера КТ CdS.

На рис. 4 представлены спектры люминесценции для образца с MnS 3 мол.% при термообработках длительностью 30, 60 и 75 мин. В данном случае наблюдается люминесценция КТ CdS. Видно, что после термообработки наблюдается смещение полос люминесценции в красную область спектра, а также уширение ее полосы. При термообработке 30 мин максимальная интенсивность наблюдается на 600 нм, при 60 мин – на 610 нм, при 75 мин – на 660 нм. После 75 минутной термообработки появляется небольшой пик в области длин волн 500–570 нм, обусловленный, скорее всего, неравномерным распределением КТ по размерам.

Рассмотрим кинетические свойства. На рис. 5, а, представлено влияние термообработки на время затухания люминесценции при различных концентрациях марганца. В исходных образцах наблюдается уменьшение времени затухания люминесценции при увеличении концентрации марганца, которое находится в пределах 18–15 мс. После термообработки в течение 45 мин наблюдаются одновременные процессы затухания как КТ CdS, так и марганца. За счет этого происходит уменьшение времени жизни – с 14 до 8 мс. В результате термообработки в течение 90 мин время жизни резко уменьшается, поскольку происходит рост диаметра КТ CdS и тушение марганца соответственно. Измеренная величина составила от 9 до 3 мс.

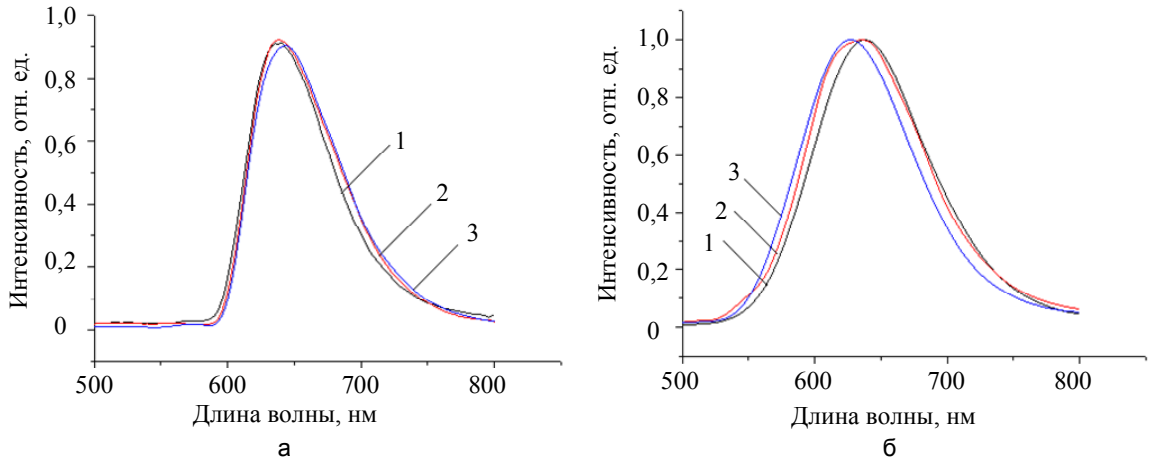


Рис. 3. Спектры люминесценции исходных образцов (а) и образцов после термообработки в течение 90 мин (б) при различных концентрациях MnS: 1 – 3 мол.%, 2 – 6 мол.%, 2 – 8 мол.%

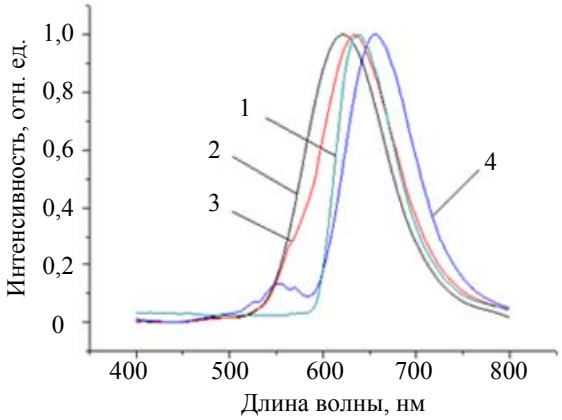


Рис. 4. Влияние термообработки на спектры люминесценции для образца MnS 3 мол.%: 1 – исходный образец; 2 – термообработка в течение 30 мин; 3 – термообработка в течение 60 мин; 4 – термообработка в течение 75 мин

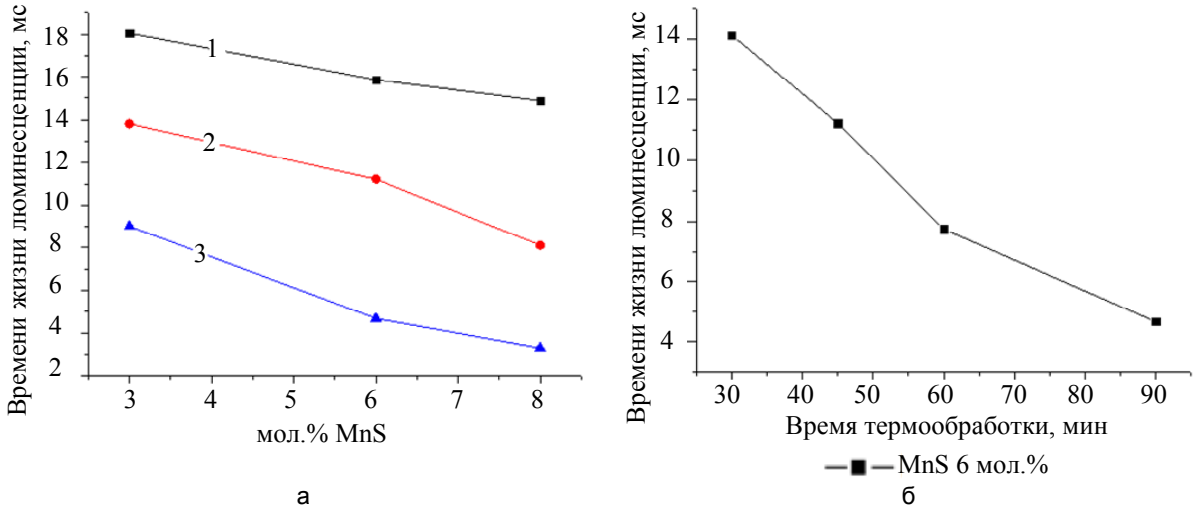


Рис. 5. Влияние термообработки на кинетику затухания люминесценции: 1 – исходные образцы; 2 – термообработка 45 мин; 3 – термообработка 90 мин (а). Изменение времени жизни люминесценции для образца с концентрацией MnS 6 мол.% в зависимости от времени термообработки (б)

На рис. 5, б, представлена зависимость времени жизни люминесценции для образца MnS 6 мол.% в результате термообработок длительностью 30, 45, 60 и 90 мин. При увеличении времени термообработки наблюдается уменьшение времени жизни люминесценции, связанное с увеличением диаметра и концентрации КТ CdS. Эта величина изменяется от 14 до 4 мс.

Заключение

Исследованы люминофоры, активированные в различных концентрациях MnS и КТ CdS. Получены спектрально-люминесцентные характеристики этих материалов. Выявлено, что влияние термообработки приводит к увеличению размеров и числа КТ, вследствие чего наблюдается сдвиг краев спектров поглощения в длинноволновую область спектра.

Рост КТ CdS и влияние марганца увеличивают ширину полосы люминесценции на 50 нм, а также ее интенсивность. Наблюдается сдвиг максимумов интенсивности люминесценции в сравнении с исходным образцом. Для спектров поглощения и люминесценции наблюдается следующая закономерность: чем меньше концентрация марганца, тем больше смещение в красную область. Кинетика затухания люминесценции стекол с КТ уменьшается в сравнении с исходными образцами на 11 мс.

Литература

1. Ушакова Е.В. Особенности эволюции фотовозбуждений в квантовых точках халькогенидов кадмия и свинца: автореф. дис. … канд. физ.-мат. наук. СПб.: НИУ ИТМО, 2012. 24 с.
2. Васильев Р.Б., Дирин Д.Н. Квантовые точки: синтез, свойства, применение. М.: ФНМ, 2007. 34 с.
3. Пак В.Н., Левкин А.Н. Оптические свойства наночастиц сульфидов цинка и кадмия в силикагеле // Известия РГПУ им. А.И. Герцена. 2008. № 64. С. 74–85.
4. Новичков Р.М., Вакштейн М.С., Нодова Е.Л., Маняшин А.О., Тараскина И.И. Способ синтеза полупроводниковых квантовых точек. Патент РФ № 2381304. Бюл. 2010. № 4.
5. Принц А.В., Принц В.Я. Способ изготовления квантовых структур: квантовых точек, проволок, элементов квантовых приборов. Патент РФ № 2278815. Бюл. 2006. № 18.
6. Витухновский А.Г., Ващенко А.А., Лебедев В.С., Васильев Р.Б. Квантово-точечный светоизлучающий органический диод. Патент РФ № 2506667. Бюл. 2014. № 4.
7. Агафонова Д.С., Колобкова Е.В., Сидоров А.И. Люминесцентные волокна с квантовыми точками CdS(Se) для волоконно-оптического датчика искры // Письма в журнал технической физики. 2012. Т. 38. № 22. С. 65–70.
8. Колобкова Е.В., Никоноров Н.В., Асеев В.А. Влияние серебра на рост квантовых точек во фторофосфатных стеклах // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 5 (81). С. 1–5.
9. Рempель С.В., Углинских М.Ю., Ворох А.С. Технология стекла, содержащего наночастицы сульфида кадмия // Исследовано в России. 2010. № 79. С. 930–933.
10. Borelly N.F., Smith D.W. Quantum confinement of PbS microcrystals in glass // Journal of Non-Crystalline Solids. 1994. V. 180. N 1. P. 25–31.
11. Олейников В.А. Квантовые точки в биологии и медицине // Природа. 2010. № 3. С. 22–28.
12. Vorobjev I.A., Rafalovskaya-Orlovskaya E.P., Gladkih A.A., Potashnikova D.M., Barteneva N.S. Applications of fluorescent semiconductor nanocrystals in microscopy and cytometry // Tsitologiya. 2011. V. 53. N 5. P. 392–403.
13. Асеев В.А., Колобкова Е.В., Некрасова Я.А., Никоноров Н.В., Рохмин А.С. Люминесценция марганца во фторфосфатных стеклах // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 6 (82). С. 36–39.
14. Reisfeld R., Kisilev A., Jorgensen C.K. Luminescence of manganese (II) in 24 phosphate glasses // Chemical Physics Letters. 1984. V. 111. N 1–2. P. 19–24.
15. Шамшурин А.В., Маскалюк Л.Г., Репин А.В. Люминофоры на основе твердых растворов фосфатов цинка и магния, активированные ионами марганца // Труды Одесского политехнического института. 1999. № 3. С. 230–232.

Липатова Жанна Олеговна – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, zluka_yo@mail.ru

Асеев Владимир Анатольевич – кандидат физико-математических наук, доцент кафедры, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Aseev@oi.ifmo.ru

Колобкова Елена Вячеславовна – доктор химических наук, профессор кафедры, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Kolobok106@rambler.ru

Zhanna O. Lipatova – student, ITMO University, Saint Petersburg, 197101, Russian Federation, zluka_yo@mail.ru

Vladimir A. Aseev – PhD, Associate professor, ITMO University, Saint Petersburg, 197101, Russian Federation, Aseev@oi.ifmo.ru

Elena V. Kolobkova – D.Sc., Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, Kolobok106@rambler.ru

Принято к печати 25.09.14
Accepted 25.09.14

5 КОМПЬЮТЕРНЫЕ СИСТЕМЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
COMPUTER SCIENCE

УДК 004.932

МЕТОД ПОВЫШЕНИЯ РЕЗКОСТИ ЦИФРОВЫХ ИЗОБРАЖЕНИЙ

В.В. Беззубик[a], Н.Р. Белашенков[a], Г.В. Вдовин[a,b,c], Н.С. Кармановский[a], О.А. Соловьев[c]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, bezzubik@mail.ru
[b] Дельфтский Технический Университет, Дельфт, 2600 АА, Нидерланды
[c] Flexible Optical BV, Рясвяк, 2288 GG, Нидерланды

Аннотация. Предложен и апробирован метод улучшения резкости цифровых изображений, основанный на выполнении многомасштабного анализа изображения, вычислении значений дифференциальных откликов его яркости по различным пространственным масштабам и последующем синтезе восстанавливающей функции, с помощью которой повышение резкости изображения производится путем простого поэлементного вычитания значений этой функции из массива значений яркости искаженного изображения. Особенностью метода является использование принципа транспозиции элементов восстанавливающей функции, ее нормировка и учет знака градиента дифференциального отклика яркости изображения в областях вблизи границ объектов. Алгоритм, реализующий предложенный метод, допускает применение целочисленной арифметики, что значительно сокращает время вычислений. В работе показано, что для изображений с небольшой величиной размытия границ, соответствующих неустранимым аберрациям изображающих систем, при синтезе восстанавливающей функции резкости достаточно ограничить рассмотрение двумя первыми масштабами. Предлагаемый метод не требует априорной информации о характере и величине ядра размытия, что соответствует представлениям о «слепой» деконволюции изображений, но практическая реализация данного метода существенно проще и не требует значительных вычислительных ресурсов. Наиболее перспективной областью применения метода являются цифровые системы машинного зрения и интеллектуальные системы наблюдения, предназначенные для работы в составе информационных комплексов, непосредственно связанных с распознаванием образов и выработкой решений на их основе в режиме реального времени.

Ключевые слова: цифровое изображение, контраст, резкость, ядро размытия, деконволюция.
Благодарности. Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации.

MULTISCALE DIFFERENTIAL METHOD FOR DIGITAL IMAGE SHARPENING

V.V. Bezzubik[a], N.R. Belashenkov[a], G.V. Vdovin[a,b,c], N.S. Karmanovsky[a], O.A. Soloviev[c]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, bezzubik@mail.ru
[b] Delft Technical University, Delft, 2600 AA, The Netherlands
[c] Flexible Optical BV, Rijaswijk, 2288 GG, The Netherlands

Abstract. We have proposed and tested a novel method for digital image sharpening. The method is based on multi-scale image analysis, calculation of differential responses of image brightness in different spatial scales, and the subsequent calculation of a restoration function, which sharpens the image by simple subtraction of its brightness values from those of the original image. The method features spatial transposition of the restoration function elements, its normalization, and taking into account the sign of the brightness differential response gradient close to the object edges. The calculation algorithm for the proposed method makes use of integer arithmetic that significantly reduces the computation time. The paper shows that for the images containing small amount of the blur due to the residual aberrations of an imaging system, only the first two scales are needed for the calculation of the restoration function. Similar to the blind deconvolution, the method requires no *a priori* information about the nature and magnitude of the blur kernel, but it is computationally inexpensive and is much easier in practical implementation. The most promising applications of the method are machine vision and surveillance systems based on real-time intelligent pattern recognition and decision making.

Keywords: digital image, image contrast, sharpness, blur kernel, deconvolution.
Acknowledgements. The research has been carried out under financial support of the Ministry of Education & Science of the Russian Federation.

Введение

Одной из ключевых характеристик, определяющих качество цифровых изображений, является резкость, которую принято связывать с величиной градиента яркости в областях вблизи границ объектов на изображении. При этом чем меньше расстояние, на котором происходит изменение яркости и чем больше величина самого изменения, тем выше резкость. Качество изображения непосредственно связано с его информационной наполненностью – способностью отчетливо отображать мелкие детали. Если рассматривать факторы, влияющие на резкость цифрового изображения, то к ним следует отнести в первую очередь аберрации изображающей оптической системы, главной из которых является дефокусировка, а также пространственный шум, определяемый свойствами матрицы фотоприемного сенсора [1]. Пренебрегая влиянием шумов на качество цифрового изображения, математическое выражение, описывающее преоб-

разование неискаженного изображения $f(x,y)$ из-за дефокусировки или аберраций, вносимых оптической системой или трактом передачи изображения, можно записать в виде

$$g(x,y) = f(x,y) \otimes h(x,y),$$

где x, y – поперечные координаты в плоскости изображения, $h(x,y)$ – искажающее ядро, а символ $\otimes$ обозначает операцию свертки.

Если считать, что границы объектов на неискаженном изображении $f(x,y)$ представляют собой области, в пределах которых функция распределения яркости изменяется скачком, то выполнение операции свертки этой функции с искажающим ядром приводит к «размытию» границ объектов и появлению так называемых «транзитных зон», в пределах которых яркость изменяется плавно. При этом контраст мелких деталей уменьшается, а некоторые из них могут вообще исчезнуть. Зрительное восприятие «размытых» изображений в целом затрудняется, а их информационная наполненность существенно снижается.

В литературе существует два основных подхода к решению задачи повышения резкости цифровых изображений, подвергшихся воздействию ядра размытия $h(x,y)$. Первый заключается в модификации профиля распределения яркости пикселей в пределах транзитной зоны при сохранении ее ширины, что приводит к повышению доли высокочастотных компонент в спектре сигнала и кажущемуся повышению информационной наполненности изображения. Применение данного подхода улучшает зрительное восприятие подвергшегося обработке изображения, однако степень улучшения оставляет желать большего, так как создается лишь иллюзия повышения резкости, главным образом, за счет появления инвертированных областей яркости на противоположных сторонах границ перепадов яркости на изображениях. Среди этих методов наиболее распространенным является метод нерезкого маскирования (unsharp masking) [2–8], состоящий в том, что исходное изображение подвергают дополнительному «размытию» путем усреднения значений яркости элементов изображения (пикселей) в пределах окрестностей определенного размера. Далее производят операцию вычитания полученного изображения из исходного, а результат складывают с исходным изображением. Этот простой способ является наиболее распространенным в технике обработки цифровых изображений, однако ряд его существенных недостатков, таких как отсутствие требований к параметрам преобразования, приводит к нежелательным результатам в виде модификации локальных контрастов и появлению артефактов.

Другой подход – это восстановление «размытых» изображений, состоящее в поиске функции профиля распределения яркости, наиболее близкой к профилю яркости неискаженного изображения. Реализация данного метода позволяет добиться уменьшения ширины транзитной зоны. Математически это означает выполнение операции обратной свертки (деконволюции), для осуществления которой необходимо знание ядра размытия $h(x,y)$. В некоторых случаях априорная информация о характеристиках этого ядра (даже приблизительная) существенно упрощает процедуру восстановления изображения благодаря хорошо разработанному математическому аппарату, позволяющему выполнять процедуру деконволюции достаточно эффективно, хотя и ценой трудоемких вычислений [9–11]. В случаях, когда априорная информация об ядре размытия $h(x,y)$ полностью отсутствует, задача качественного восстановления изображения («слепая» деконволюция) становится намного сложнее, так как в этом случае результат сильно зависит от начальных условий длительного итерационного процесса предсказания–коррекции при поиске неизвестного ядра. Также стремительно возрастает общая трудоемкость вычислений [12–15]. При этом нет никаких гарантий того, что итерационный процесс восстановления изображения сойдется к истинному значению распределения яркости неискаженного изображения.

Таким образом, задача поиска новых способов эффективного восстановления качества искаженных изображений на основе простых алгоритмов продолжает оставаться весьма актуальной.

В настоящей работе предложен метод повышения резкости цифровых изображений, который сочетает в себе простоту и высокую производительность с высокой степенью приближения к «идеальным» характеристикам изображений, сохранением величин локальных и глобальных контрастов, а также отсутствием искусственно создаваемых артефактов. Данный метод позволяет найти приближенное значение $\tilde{f}(x,y)$ невозмущенной функции $f(x,y)$ путем простого вычитания из известного сигнала $g(x,y)$ восстанавливающей функции $\varphi(x,y)$:

$$\tilde{f}(x,y) = g(x,y) - \varphi(x,y).$$

Ниже будет описан процесс построения такой восстанавливающей функции $\varphi(x,y)$.

Многомасштабный анализ цифрового изображения

Рассмотрим цифровое полутоновое изображение шириной M (координата x, индекс i) и высотой N (координата y, индекс j) пикселей со значениями яркости i, j-го пикселя $f_{i,j}$, подвергшееся воздействию неиз-

вестного ядра размытия $h_{i,j}$. Исходное искаженное («размытое») изображение $g_{i,j}$ может быть представлено в виде результата применения операции дискретной свертки к невозмущенному изображению $f_{i,j}$:

$$g_{i,j} = \sum_p \sum_q f_{i+p,j+q} \cdot h_{p,q} \ .$$

Расположим начало координат изображения в левом верхнем углу так, что индекс *i* по координате *x* возрастает слева направо от 1 до *M*, а индекс *j* по координате *y* возрастает сверху вниз от 1 до *N*.

Введем отдельно для координат *x* и *y* наборы квадратных цифровых фильтров с нечетным числом элементов $S \times S$ различного размера ($S \geq 3$), элементы которых имеют следующий вид:

$$Kx_{p,q}^{(S)} = \begin{cases} -1, & -S_c \leq p < 0, \quad -S_c \leq q \leq S_c \\ 0, & p = 0, \quad -S_c \leq q \leq S_c \\ +1, & 0 < p \leq S_c, -S_c \leq q \leq S_c \end{cases}, \tag{1}$$

$$Ky_{p,q}^{(S)} = \begin{cases} -1, & -S_c \leq p \leq S_c, \ -S_c \leq q < 0 \\ 0, & -S_c \leq p \leq S_c, \quad q = 0 \\ +1, & -S_c \leq p \leq S_c, \quad 0 < q \leq S_c \end{cases}, \tag{2}$$

где $S_c = (S-1)/2$, p и q – индексы по координатам *x* и *y* соответственно, диапазон изменения которых составляет от $-S_c$ до S_c. В дальнейшем по аналогии с вейвлет-анализом будем называть величину *S* масштабом и обозначать верхним индексом, заключенным в скобки. Ввиду того, что записи (1) и (2) для цифровых фильтров не совсем обычны, для наглядности приведем их вид при *S* = 3 и *S* = 5:

$$Kx^{(3)} = \begin{vmatrix} -1 & 0 & +1 \\ -1 & 0 & +1 \\ -1 & 0 & +1 \end{vmatrix}, \ Ky^{(3)} = \begin{vmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ +1 & +1 & +1 \end{vmatrix},$$

$$Kx^{(5)} = \begin{vmatrix} -1 & -1 & 0 & +1 & +1 \\ -1 & -1 & 0 & +1 & +1 \\ -1 & -1 & 0 & +1 & +1 \\ -1 & -1 & 0 & +1 & +1 \\ -1 & -1 & 0 & +1 & +1 \end{vmatrix}, \ Ky^{(5)} = \begin{vmatrix} -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 & +1 \end{vmatrix}.$$

Следует отметить, что для $S = 3$ фильтры $Kx^{(3)}$ и $Ky^{(3)}$ представляют собой известные операторы Prewitt, часто используемые при детекции границ объектов в цифровой обработке изображений. Очевидно, что фильтры (1) и (2) удовлетворяют условию нулевого среднего:

$$\sum_{p=-S_c}^{S_c} \sum_{q=-S_c}^{S_c} Kx_{p,q}^{(S)} = 0 \ , \ \sum_{p=-S_c}^{S_c} \sum_{q=-S_c}^{S_c} Ky_{p,q}^{(S)} = 0 \ . \tag{3}$$

Путем последовательного вычисления дискретной свертки исходного изображения $g_{i,j}$ с набором фильтров (1) и (2) различного масштаба *S*, получим величины $Rx_{i,j}^{(S)}$ и $Ry_{i,j}^{(S)}$,

$$Rx_{i,j}^{(S)} = A^{-1}(S) \sum_{p=-S_c}^{p=S_c} \sum_{q=-S_c}^{q=S_c} g_{i+p,j+q} \cdot Kx_{p,q}^{(S)} \ , \tag{4}$$

$$Ry_{i,j}^{(S)} = A^{-1}(S) \sum_{p=-S_c}^{p=S_c} \sum_{q=-S_c}^{q=S_c} g_{i+p,j+q} \cdot Ky_{p,q}^{(S)} \ , \tag{5}$$

которые в дальнейшем будем называть нормированными дифференциальными откликами яркости цифрового изображения в точке с координатами (i, j) по соответствующей координате и масштабу *S*. Необходимо отметить, что в результате выполнения операции дискретной свертки из-за граничных эффектов размеры массивов дифференциальных откликов уменьшаются и равны $(M - S - 1) \times (N - S - 1)$ для соответствующего масштаба *S*. В выражениях (4) и (5) нормирующий множитель $A(S)$ равен $A(S) = S(S^2 - 1)/4$. Величина $A(S)$ выбрана таким образом, чтобы для любого масштаба *S* было обеспечено равенство интегральных величин модулей дифференциальных откликов по всему изображению:

$$\sum_{i,j} \left| Rx_{i,j}^{(S)} \right| = C_x = \text{const} \ \forall S \, , \quad \sum_{i,j} \left| Ry_{i,j}^{(S)} \right| = C_y = \text{const} \ \forall S \ .$$

Величины C_x и C_y имеют смысл суммы абсолютных значений перепадов яркости на всех границах объектов на изображении.

В дальнейшем при анализе нам потребуются также массивы производных по соответствующим координатам и масштабам от нормированных дифференциальных откликов, которые мы определим следующим образом:

$$Px_{i,j}^{(S)} = \frac{1}{2}\left(Rx_{i+1,j}^{(S)} - Rx_{i-1,j}^{(S)} \right),\ Py_{i,j}^{(S)} = \frac{1}{2}\left(Ry_{1,j+1}^{(S)} - Ry_{i,j-1}^{(S)} \right). \quad (6)$$

Синтез восстанавливающей функции резкости

Восстанавливающая функция $\varphi_{i,j}$ в каждой точке i, j изображения рассчитывается как сумма вкладов в нее $\varphi_{i,j}^{(S)}$ по соответствующим масштабам S:

$$\varphi_{i,j} = \varphi_{i,j}^{(3)} + \varphi_{i,j}^{(5)} + \varphi_{i,j}^{(7)} + \ldots, \quad (7)$$

а качество восстановления $f_{i,j}$ в общем случае определяется числом учитываемых при расчете масштабов. Перед вычислением вклада $\varphi_{i,j}^{(S)}$ в восстанавливающую функцию от каждого масштаба S создадим массивы величин $\overline{R}x_{i,j}^{(S)}$ и $\overline{R}y_{i,j}^{(S)}$ путем транспозиции элементов массивов $Rx_{i,j}^{(S)}$ и $Ry_{i,j}^{(S)}$ по следующим правилам:

$$\overline{R}\,x_{i,j}^{(S)} = Rx_{i-k^{(S)},j}^{(S)}, \text{ где } k^{(S)} = S_c \cdot sign(Rx_{i,j}^{(S)}) \cdot sign(Px_{i,j}^{(S)}), \quad (8)$$

$$\overline{R}\,y_{i,j}^{(S)} = Ry_{i,j-l^{(S)}}^{(S)}, \text{ где } l^{(S)} = S_c \cdot sign(Ry_{i,j}^{(S)}) \cdot sign(Py_{i,j}^{(S)}), \quad (9)$$

после чего вклады $\varphi_{i,j}^{(S)}$ по каждому из масштабов вычислим следующим образом:

$$\varphi_{i,j}^{(S)} = \begin{cases} \left|\overline{R}\,x_{i,j}^{(S)}\right| \cdot sign(Px_{i,j}^{(S)}), \ для\ \left|\overline{R}\,x_{i,j}^{(S)}\right| \geq \left|\overline{R}\,y_{i,j}^{(S)}\right| \\ \left|\overline{R}\,y_{i,j}^{(S)}\right| \cdot sign(Py_{i,j}^{(S)}), \ для\ \left|\overline{R}\,x_{i,j}^{(S)}\right| < \left|\overline{R}\,y_{i,j}^{(S)}\right| \end{cases}, \quad (10)$$

а восстановленное изображение $\tilde{f}_{i,j}$ найдем путем простого поэлементного вычитания $\varphi_{i,j}$ из исходного искаженного изображения $g_{i,j}$:

$$\tilde{f}_{i,j} = g_{i,j} - \varphi_{i,j}. \quad (11)$$

Рассмотрим последовательность действий при вычислении восстанавливающей функции $\varphi_{i,j}$ на примере двух модельных изображений. Первое из них представляет собой границу перепада яркости от 100 до 200 единиц градации серого по координате x. Второе изображение представляет собой результат преобразования первого изображения «размывающим» гауссовым фильтром с радиусом r, равным 1 пикселю. Профили обоих изображений показаны на рис. 1, а, пунктирной и сплошной линиями соответственно. После вычисления по формулам (4) и (5) величин нормированных дифференциальных откликов $f_{i,j}$ и $g_{i,j}$ по первым четырем масштабам S построим спектрограммы (зависимости дифференциального отклика от пространственной координаты и масштаба), аналогичные тем, которые приняты в вейвлет-анализе. На рис. 1, б, пунктирная линия отображает дифференциальный отклик изображения $f_{i,j}$, сплошная – изображения $g_{i,j}$, а точечная линия представляет собой производную дифференциального отклика $g_{i,j}$, рассчитанную по формуле (6). Из спектрограмм видно, что дифференциальные отклики на участках изображений с постоянным значением яркости равны нулю на всех масштабах. Это является следствием условия (3). На рис. 1, б, серым цветом закрашены области ненулевых значений нормированных дифференциальных откликов яркости цифрового изображения $g_{i,j}$ для тех значений пространственной координаты x, для которых значения дифференциальных откликов яркости $f_{i,j}$ равны 0. Иными словами, для синтеза восстанавливающей функции резкости интерес представляют значения нормированных дифференциальных откликов яркости размытого цифрового изображения $g_{i,j}$, расположенных в тех областях, в которые ядро размытия «рассеяло» часть яркости «резкого» изображения. Чем больше размер ядра размытия, тем протяженнее будут эти области и тем значительнее будут вклады в восстанавливающую функцию от более высоких масштабов. По форме и размерам спектрограммы в периферийных областях можно судить о характеристиках ядра размытия. Задача восстановления резкости, таким образом, заключается в том, чтобы вернуть рассеянную ядром яркость из периферийной области «размытого» изображения на свои места в область границы с резким перепадом яркости. Отметим, что в нашем примере величина C_x равна величине перепада яркости на границе, т.е. 100 единицам, и одинакова как для резкой, так и для размытой границы на любом масштабе S. Именно это обстоятельство и позволит нам в дальнейшем произвести реконструкцию искаженного изображения с помощью простого вычитания из него восстанавливающей функции по формуле (11).

Отметим, что производная от дифференциальных откликов по масштабам от размытого изображения (точечная кривая на рис. 1, б) используется при построении вкладов от масштабов в восстанавливающую функцию по выражениям (8)–(10). На рис. 1, в, проиллюстрирован процесс транспозиции, который определяется выражениями (8) и (9), а также на этом же рисунке сплошной линией изображены вклады $\varphi_{i,j}^{(S)}$ для ряда масштабов в восстанавливающую функцию $\varphi_{i,j}$ (выражение (7)). Видно, что с ростом масштаба S величина вклада $\varphi_{i,j}^{(S)}$ в восстанавливающую функцию $\varphi_{i,j}$ уменьшается. На рис. 1, г, показаны результаты вычитания значений восстанавливающей функции $\varphi_{i,j}$ из размытого изображения $g_{i,j}$ при разных значениях количества слагаемых N_S, принятых к учету в выражении (7), или, другими словами, количества масштабов, использованных при синтезе восстанавливающей функции $\varphi_{i,j}$. На рис. 1, г, кривая 1 соответствует $N_S = 1$, кривая 2 – $N_S = 2$ и т.д. Видно, что применение предлагаемого метода не приводит к появлению артефактов изменения локального контраста изображения в виде «гало», что является существенным преимуществом при обработке цифровых изображений в системах машинного зрения.

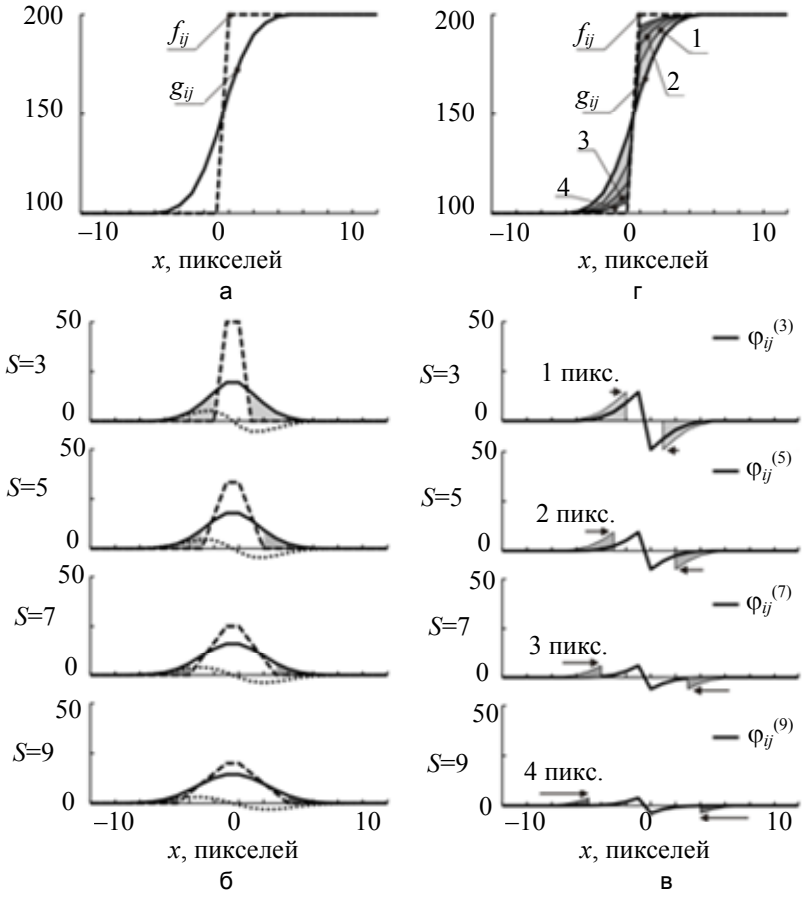


Рис. 1. Последовательность восстановления резкости изображения: многомасштабный анализ – этапы (а) и (б); транспозиция и синтез восстанавливающей функции – этап (в); восстановление резкости границ объектов – этап (г)

Восстановление резкости тестовых и реальных изображений

Предложенный метод был апробирован на тестовых и реальных изображениях. Тестовые изображения были построены с помощью генератора графических образов, а реальный снимок получен с камеры мобильного телефона среднего качества.

В качестве тестового мы использовали изображение круга диаметром 25 пикселей яркостью 200 единиц градации серого на фоне 100 единиц градаций серого.

На рис. 2, а, представлены увеличенное изображение тестового объекта и соответствующее сечение профиля интенсивности по координате x, а на рис. 2, б, – изображение того же тестового объекта, подвергнутого воздействию ядра размытия в виде гауссова фильтра с радиусом r = 1 пиксель и сечение его профиля яркости по координате x.

На рис. 3 представлены результаты восстановления размытого изображения (рис. 2, б) методом unsharp masking с радиусом ядра дополнительного размытия R_{UM} = 1 пиксель (рис. 3, а) и предложенным методом. Столбец (рис. 3, б) соответствует случаю, когда в сумму (7) входит только один член (N_S = 1). Результаты, приведенные в столбцах (рис. 3, в, г), получены при учете двух (N_S = 2) и трех (N_S = 3) членов суммы (7) соответственно.

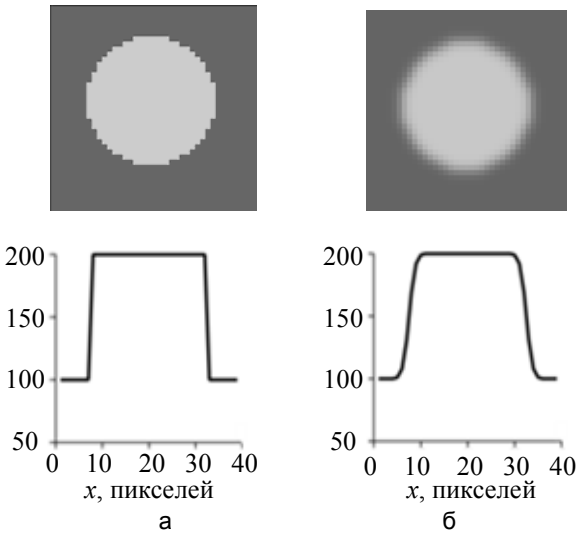


Рис. 2. Тестовые изображения и сечения профиля яркости по координате *x*: идеальное изображение (а); изображение, размытое гауссовым ядром с радиусом *r*, равным 1 пикселю (б)

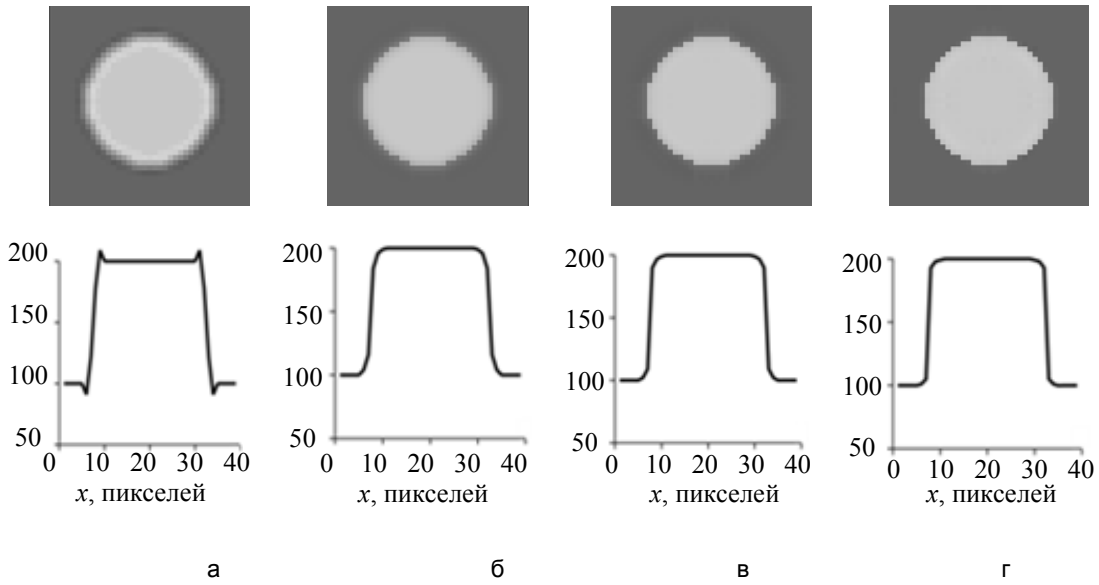


Рис. 3. Результаты восстановления размытого тестового изображения (рис. 2, б) методом unsharp masking с радиусом ядра дополнительного размытия R_{UM}, равным 1 пикселю (а), и предложенным методом при N_S = 1 (б), N_S = 2 (в) и N_S = 3 (г)

Для оценки эффективности восстановления резкости размытого изображения предложенным методом и его сравнения с методом unsharp masking воспользуемся количественной мерой [16]

$$\sigma^2 = \frac{\sum_M \sum_N (\tilde{f}_{i,j} - f_{i,j})^2}{M\,N},$$

которая представляет собой квадрат среднеквадратического отклонения значений яркости пикселей восстановленного изображения от значений яркости пикселей исходно резкого изображения. Результаты расчета σ^2 для случаев восстановления резкости тестовых изображений, размытых гауссовыми ядрами различного радиуса, в нормированном виде представлены на рис. 4. На рис. 4, а, по оси абсцисс отложе-

но количество N_S членов суммы (7), учитываемых при синтезе восстанавливающей функции $\varphi_{i,j}$, причем $N_S = 0$ соответствует случаю $\tilde{f}_{i,j} \equiv g_{i,j}$. Из рисунка видно, что с увеличением числа учитываемых масштабов N_S мера σ^2 быстро достигает минимального значения ($N_S = 3$ для размытого изображения с $r = 1{,}0$ – кривая 1, $N_S = 4$ для размытых изображений с $r = 1{,}5$ и $r = 2{,}0$ – кривые 2 и 3), а затем наблюдается ухудшение качества восстановления резкости. Следует обратить внимание на тот факт, что для $r = 1{,}0$ значения σ^2 при $N_S = 2$ и $N_S = 3$ отличаются незначительно, а вычислительные затраты метода с ростом масштаба растут квадратично. Учет двух членов в сумме (7) в этом случае будет предпочтительным. Для изображений, размытых гауссовыми ядрами с $r > 1{,}0$, синтез восстанавливающей функции требует учета дополнительных слагаемых.

Для сравнения предложенного метода восстановления резкости цифровых изображений с широко используемым на практике методом unsharp masking было произведено восстановление тестовых размытых изображений с $r = 1{,}0$; $r = 1{,}5$ и $r = 2{,}0$ с различными значениями радиуса ядра дополнительного размытия R_{UM}. Результаты расчета σ^2 для этих случаев представлены на рис. 4, б. Как видно из рисунка, количественная мера восстановления также имеет минимумы, но абсолютные значения этих минимумов значительно больше тех, которые достигнуты в предложенном методе. Столь существенное различие в значениях $\sigma^2_{\min}$ рассмотренных методов имеет достаточно простое объяснение – unsharp masking не изменяет ширины транзитной зоны в области границ объектов, а лишь повышает локальные контрасты за счет изменения значений яркости в области границ объектов. Этот метод ориентирован на физиологический отклик в зрительной системе и поэтому нашел достаточно широкое применение при улучшении качества восприятия изображения человеком. Однако в системах машинного зрения появление искусственных артефактов может приводить к дисфункциям и нарушениям работы алгоритмов.

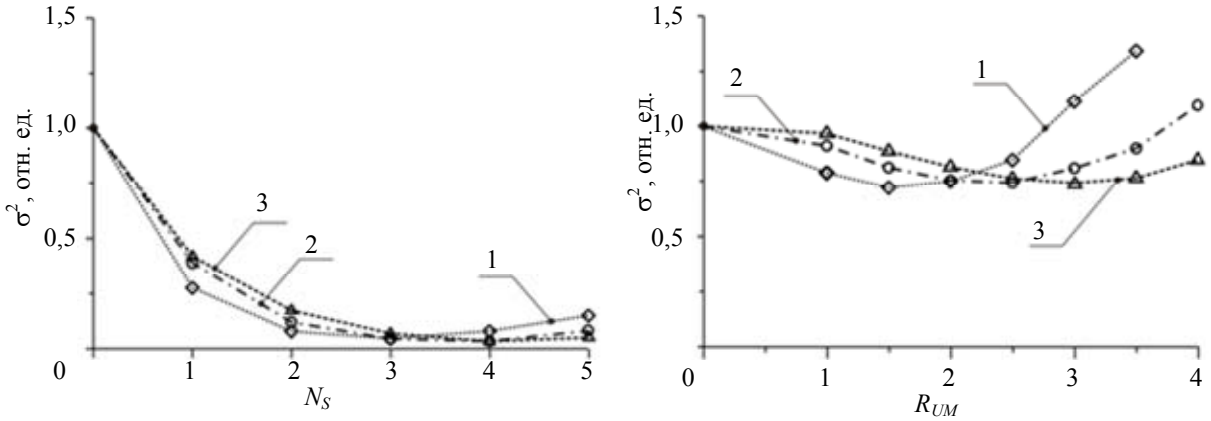


Рис. 4. Зависимости эффективности восстановления резкости изображений, размытых гауссовыми ядрами с $r = 1{,}0$; $r = 1{,}5$ и $r = 2{,}0$ (кривые 1, 2 и 3): предложенным методом от числа N_S (а); методом unsharp masking от радиуса R_{UM} (б)

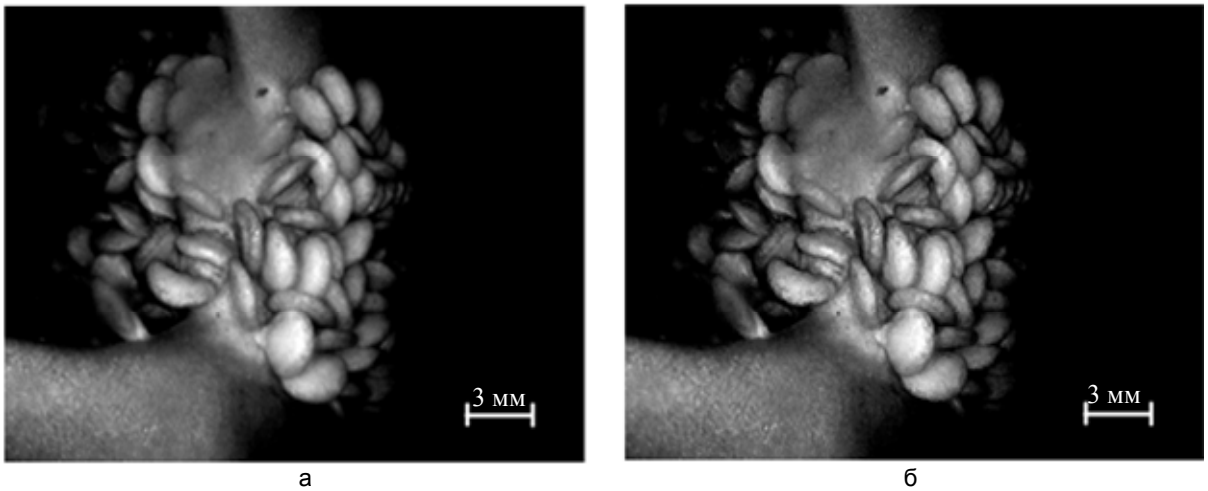


Рис. 5. Изображение внутренней полости перца, полученное с помощью цифрового видеоэндоскопа (а), и результат его восстановления предложенным методом при $N_S = 2$ (б)

На рис. 5 приведен результат восстановления резкости изображения внутренней полости перца (Capsicum annuum), полученного с помощью цифрового видеоэндоскопа. Ввиду отсутствия эталонного изображения количественную оценку степени улучшения в данном случае привести невозможно. Тем не менее, следует отметить, что на восстановленном изображении артефакты не наблюдаются, а величина контраста, рассчитанного по классическому определению как отношение среднеквадратического отклонения значений яркости пикселей к среднему значению градации серого [17] по всему кадру, сохраняется неизменной. В приведенном примере размер изображения составлял 720×564 пикселей, а время вычислений на процессоре с тактовой частотой 3 ГГц составило 0,11 мс.

Заключение

В работе предложен эффективный метод улучшения резкости цифровых изображений, не приводящий к появлению артефактов. Метод основан на вычислении двумерных массивов данных, отвечающих дифференциальным откликам пространственного распределения яркости цифрового изображения в виде результатов свертки изображения с дифференцирующими фильтрами одного или нескольких линейно изменяющихся размеров. Сокращение ширины зоны «размытия» границ объектов на изображении достигается за счет пространственной транспозиции этих данных и их линейной комбинации с исходным изображением. Показана роль нормировки значений дифференциальных откликов, рассчитанных по различным пространственным масштабам. Применение метода не требует предварительного знания ядра размытия искаженного изображения и определения каких-либо параметров алгоритма. Возможности метода продемонстрированы на примерах тестовых изображений в сравнении с широко используемым на практике методом повышения резкости путем нерезкого маскирования исходного изображения – unsharp masking.

Литература

1. Беззубик В.В., Белашенков Н.Р., Никифоров В.О. Метод количественной оценки контраста цифрового изображения // Научно-технический вестник СПбГУ ИТМО. 2010. № 6 (70). С. 86–88.
2. Polesel A., Ramponi G., Mathews V.J. Image enhancement via adaptive unsharp masking // IEEE Transactions on Image Processing. 2000. V. 9. N 3. P. 505–510.
3. Cao G., Zhao Y., Ni R., Kot A.C. Unsharp masking sharpening detection via overshoot artifacts analysis // IEEE Signal Processing Letters. 2011. V. 18. N 10. P. 603–606.
4. Kim S.H., Allebach J.P. Optimal unsharp mask for image sharpening and noise removal // Journal of Electronic Imaging. 2005. V. 14. N 2. Art. 023005. P. 1–13.
5. Kotera H., Wang H. Multiscale image sharpening adaptive to edge profile // Journal of Electronic Imaging. 2005. V. 14. N 1. Art. 013002. P. 1–17.
6. Kwok N.M., Shi H.Y., Fang G., Ha Q.P. Intensity-based gain adaptive unsharp masking for image contrast enhancement // Proc. 5th Int. Congress on Image and Signal Processing (CISP 2012). Chongqing, China, 2012. P. 529–533.
7. Hong H., Li L., Park I.K., Zhang T. Universal deblurring method for real images using transition region // Optical Engineering. 2012. V. 51. N 4. Art. 047006.
8. Loza A., Bull D.R., Hill P.R., Achim A.M. Automatic contrast enhancement of low-light images based on local statistics of wavelet coefficients // Digital Signal Processing. 2013. V. 23. N 6. P. 1856–1866.
9. Morigi S., Reichel L., Sgallari F., Shyshkov A. Cascadic multiresolution methods for image deblurring // SIAM Journal on Imaging Sciences. 2008. V. 1. N 1. P. 51–74.
10. Levin A., Weiss Y., Durand F., Freeman W.T. Understanding and evaluating blind deconvolution algorithms // Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Miami, US, 2009. P. 1964–1971.
11. Barone F., Rossi C. Deconvolution with partially known kernel of nonnegative signals // Machine Vision and Applications. 1990. V. 3. N 2. P. 107–115.
12. Markham J., Conchello J.-A. Parametric blind deconvolution: a robust method for the simultaneous estimation of image and blur // Journal of Optical Society of America A: Optics and Image Science, and Vision. 1999. V. 16. N 10. P. 2377–2391.
13. Kundur D., Hatzinakos D. Blind image deconvolution // IEEE Signal Processing Magazine. 1996. V. 13. N 3. P. 43–64.
14. Mo X., Jiao J., Shen C. PSF-constraints based iterative blind deconvolution method for image deblurring // Lecture Notes in Computer Science. 2010. V. 5916 LNCS. P. 141–151.
15. Laligant O., Truchetet F., Dupasquier A. Edge enhancement by local deconvolution // Pattern Recognition. 2005. V. 38. N 5. P. 661–672.
16. The Oxford Dictionary of Statistical Terms. 6th ed. Ed. Y. Dodge. Oxford: Oxford University Press, 2003. 498 p.
17. Вудс Р., Гонсалес Р. Цифровая обработка изображений. М.: Техносфера, 2005. 1072 с.

Беззубик Виталий Вениаминович – ведущий инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, bezzubik@mail.ru

Белашенков Николай Романович – кандидат физико-математических наук, доцент, начальник ДНИР, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, belashenkov@gmail.com

Вдовин Глеб Валерьевич – PhD, профессор, Дельфтский Технический Университет, Дельфт, 2600 AA, Нидерланды; директор, Flexible Optical BV, Рясвяк, 2288 GG, Нидерланды, gleb@okotech.com

Кармановский Николай Сергеевич – кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, karmanov50@mail.ru

Соловьев Олег Александрович – PhD, ведущий научный сотрудник, Flexible Optical BV, Рясвяк, 2288 GG, Нидерланды, oleg@okotech.com

Vitaly V. Bezzubik – leading engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, bezzubik@mail.ru

Nickolai R. Belashenkov – PhD, Associate professor, Head of Research and Development Department, ITMO University, Saint Petersburg, 197101, Russian Federation, belashenkov@gmail.com

Gleb V. Vdovin – PhD, Professor, Delft Technical University, Delft, 2600 AA, The Netherlands; Director, Flexible Optical BV, Rijaswijk, 2288 GG, The Netherlands, gleb@okotech.com

Nikolai S. Karmanovsky – PhD, Associate professor, Associate professor, ITMO University, Saint Petersburg, 197101, Russian Federation,karmanov50@mail.ru

Oleg A. Soloviev – PhD, leading scientific researcher, Flexible Optical BV, Rijaswijk, 2288 GG, The Netherlands, oleg@okotech.com

Принято к печати 09.10.14
Accepted 09.10.14

УДК 519.872

ДИСЦИПЛИНЫ ОБСЛУЖИВАНИЯ НА ОСНОВЕ МАТРИЦЫ ПРИОРИТЕТОВ

Т.И. Алиев[a], Э. Махаревс[b]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, aliev@cs.ifmo.ru
[b] Балтийская международная академия, Рига, LV-1019, Латвия

Аннотация. Рассматриваются дисциплины обслуживания заявок общего вида в системах массового обслуживания с неоднородной нагрузкой. Для математического описания таких дисциплин предлагается использовать матрицу приоритетов, отображающую вид приоритета (относительный, абсолютный или его отсутствие) между двумя любыми классами заявок. Такой способ описания, обладая наглядностью и простотой задания приоритетов, позволяет получить математические зависимости характеристик функционирования системы от параметров. Сформулированы требования к формированию матрицы приоритетов, введено понятие канонической матрицы приоритетов. Показано, что не всякая матрица, построенная в соответствии с этими требованиями, является корректной. Понятие некорректности матрицы приоритетов проиллюстрировано на примере; показано, что такие матрицы не обеспечивают однозначности и определенности при разработке алгоритма, реализующего соответствующие им дисциплины обслуживания. Для канонических матриц приоритетов сформулированы правила построения корректных матриц. В качестве одной из основных характеристик рассматривается время пребывания в системе заявок разных классов, которое складывается из времени ожидания начала обслуживания и времени нахождения заявки на обработке. Для этих характеристик с использованием метода введения дополнительного события получены преобразования Лапласа, на основе которых выведены математические зависимости для расчета двух первых начальных моментов соответствующих характеристик обслуживания заявок.

Ключевые слова: система обслуживания, дисциплина обслуживания, смешанные приоритеты, матрица приоритетов, канонические матрицы приоритетов, корректные и некорректные матрицы приоритетов.

QUEUEING DISCIPLINES BASED ON PRIORITY MATRIX

T.I. Aliev[a], E. Maharevs[b]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, aliev@cs.ifmo.ru
[b] Baltic International Academy, Riga, LV-1019, Latvia

Abstract. The paper deals with queueing disciplines for demands of general type in queueing systems with multivendor load. A priority matrix is proposed to be used for the purpose of mathematical description of such disciplines, which represents the priority type (preemptive priority, not preemptive priority or no priority) between any two demands classes. Having an intuitive and simple way of priority assignment, such description gives mathematical dependencies of system operation characteristics on its parameters. Requirements for priority matrix construction are formulated and the notion of canonical priority matrix is given. It is shown that not every matrix, constructed in accordance with such requirements, is correct. The notion of incorrect priority matrix is illustrated by an example, and it is shown that such matrixes do not ensure any unambiguousness and determinacy in design of algorithm, which realizes corresponding queueing discipline. Rules governing construction of correct matrixes are given for canonical priority matrixes. Residence time for demands of different classes in system, which is the sum of waiting time and service time, is considered as one of the most important characteristics. By introducing extra event method Laplace transforms for these characteristics are obtained, and mathematical dependencies are derived on their basis for calculation of two first moments for corresponding characteristics of demands queueing.

Keywords: queueing system, queueing discipline, mixed priorities, priority matrix, canonical priority matrix, correct and incorrect priority matrix.

Введение

Системы массового обслуживания (СМО) с неоднородным потоком заявок широко применяются в качестве моделей систем различного назначения, включая вычислительные системы и компьютерные сети [1–5]. Качество функционирования таких систем определяется значениями характеристик обслуживания заявок, таких как время пребывания и ожидания заявок, число заявок в системе. Для обеспечения требуемого качества функционирования используются различные стратегии управления поступающими в систему потоками заявок, задаваемые в виде дисциплин обслуживания (ДО). Важное место среди них занимают приоритетные дисциплины, в частности, с относительными (ОП) или с абсолютными (АП) приоритетами [6–8].

Дисциплины обслуживания заявок с одним классом приоритетов (ОП или АП) не всегда позволяют достичь требуемого качества функционирования системы. Исходя из этого, в реальных системах часто используются ДО общего вида, в которых один и тот же класс заявок может иметь АП по отношению к одной группе классов заявок, ОП – к другой группе и не иметь приоритета к остальным классам заявок (обслуживание без приоритетов в порядке поступления, БП). Такие ДО относятся к дисциплинам со смешанными приоритетами (ДО СП) [9–12], которые при наличии ограничений на время пребывания заявок в системе позволяют наилучшим образом обеспечить выполнение этих ограничений.

При построении математических зависимостей характеристик обслуживания заявок от параметров системы при использовании ДО СП возникает задача математического описания таких дисциплин. Ранее авторами для описания ДО СП было предложено использовать матрицу приоритетов (МП), отражающую вид приоритета между двумя любыми классами заявок [10]. В отличие от других подходов к описанию ДО с несколькими уровнями приоритетов [11, 12], рассматриваемое матричное представление позволяет

назначать приоритеты классам заявок в произвольном порядке, что расширяет множество возможных ДО и охватывает большое количество различных ДО даже при небольшом числе классов заявок. Отметим, что математические зависимости для расчета характеристик функционирования СМО в [10] получены только для средних значений, в частности, для среднего времени ожидания заявок разных классов.

Ниже сформулированы требования к построению МП и представлены аналитические зависимости характеристик обслуживания заявок разных классов от параметров системы на уровне преобразований Лапласа для непрерывных величин и производящих функций для дискретных величин.

Матрица приоритетов

МП представляет собой квадратную матрицу $\mathbf{Q} = [q_{ij}\ (i, j = 1, \ldots, H)]$, размерность которой определяется числом классов заявок H, поступающих в систему. Элемент q_{ij} матрицы задает приоритет заявок класса i (i-заявок) по отношению к заявкам класса j (j-заявкам) и может принимать следующие значения: 0 – нет приоритета, 1 – приоритет относительный и 2 – приоритет абсолютный.

С помощью МП можно описать большое множество ДО, в том числе с одним классом приоритетов. Так, например, в случае четырех классов заявок $(H = 4)$ матрицы, соответствующие традиционным ДО ОП, ДО АП и произвольной ДО СП, будут иметь следующий вид:

$$\mathbf{Q}^{\text{ОП}} = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 1 & 1 \\ 2 & 0 & 0 & 1 & 1 \\ 3 & 0 & 0 & 0 & 1 \\ 4 & 0 & 0 & 0 & 0 \end{array};\quad \mathbf{Q}^{\text{АП}} = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 2 & 2 & 2 \\ 2 & 0 & 0 & 2 & 2 \\ 3 & 0 & 0 & 0 & 2 \\ 4 & 0 & 0 & 0 & 0 \end{array};\quad \mathbf{Q}^{\text{СП}} = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 2 & 2 & 0 & 1 \\ 4 & 0 & 1 & 0 & 0 \end{array}.$$

В отличие от традиционных ДО с одним классом приоритетов (ОП или АП), в которых обычно предполагается, что приоритеты назначены по правилу «у класса с меньшим номером – более высокий приоритет» [10], с помощью МП приоритеты могут быть назначены произвольным образом, как это показано выше для МП $\mathbf{Q}^{\text{СП}}$, где наивысший приоритет имеют заявки класса 3, а самый низкий – заявки класса 2.

Элементы МП должны удовлетворять следующим требованиям:

1. $q_{ii} = 0\ \ (i = 1, \ldots, H)$;
2. если $q_{ij} = 1$ или 2, то $q_{ji} = 0\ \ (i, j = 1, \ldots, H)$.

Матрица приоритетов называется канонической, если $q_{ij} = 0$ для всех $i \geq j\ \ (i, j = 1, \ldots, H)$. Канонические МП описывают дисциплины, в которых заявки классов с меньшим номером имеют приоритет не ниже, чем заявки классов с бóльшим номером.

Число вариантов заполнения канонической МП $\xi = 3^{H(H-1)/2}$, где H – число классов заявок, определяющее размерность матрицы. Однако из этого числа должны быть исключены так называемые некорректные матрицы приоритетов.

Корректность МП предполагает однозначность и определенность алгоритма, реализующего соответствующую МП. Не всякая матрица приоритетов, удовлетворяющая перечисленным выше требованиям, является корректной. МП является некорректной, если при ее реализации может возникнуть неоднозначная ситуация, причем любое принятое решение будет противоречить заданной МП.

Проиллюстрируем понятие некорректности на примере МП для трех классов заявок $(H = 3)$ вида

$$\mathbf{Q} = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0 & 1 & 1 \\ 2 & 0 & 0 & 2 \\ 3 & 0 & 0 & 0 \end{array}.$$

Рассмотрим следующую ситуацию. Пусть в момент поступления в систему заявки класса 2 в приборе обслуживается заявка класса 3, а в очереди находится одна или несколько заявок класса 1, которые не могут прервать обслуживание заявки класса 3, так как в соответствии с заданной МП они имеют только ОП по отношению к заявкам класса 3 ($q_{13} = 1$). При этом возникает следующая неопределенность. С одной стороны, поступившая заявка класса 2 должна прервать обслуживание заявки класса 3, поскольку по отношению к заявкам класса 3 имеет АП $(q_{23} = 2)$. С другой стороны, поступившая заявка класса 2 не может начать обслуживаться раньше заявок класса 1, находящихся в очереди и имеющих ОП по отношению к заявкам класса 2 ($q_{12} = 1$). Эта неопределенность приводит к неоднозначности при построении алгоритма, реализующего данную ДО. Любое решение из двух возможных (прервать и не прервать об-

служивание заявки класса 3) приведет к дисциплине, не соответствующей заданной МП. Такие МП в дальнейшем исключим из рассмотрения, относя их к некорректным. При этом резко уменьшается число возможных вариантов заполнения МП (табл. 1).

Размерность МП	2×2	3×3	4×4	5×5	6×6
Число вариантов ξ заполнения канонических МП	3	27	729	59049	>14 млн
Число корректных канонических МП	3	13	75	541	4683
Оценка числа корректных канонических МП $\tilde{\xi}_H$	3	13	75	537	4644

Таблица 1. Число матриц приоритетов

Отметим, что с ростом числа классов заявок число различных ДО растет экспоненциально.

Для приближенной оценки числа корректных канонических МП при $H > 6$ экспериментальным путем с использованием метода математической индукции получено следующее выражение: $\tilde{\xi}_H \approx 1,5 \times 1,44^{H-2} H!$ $(H \geq 2)$. Это выражение, как видно из табл. 1, дает нижнюю оценку числа корректных канонических МП. Заметим, что $\tilde{\xi}_H$ определяет только число канонических МП. Если допустить произвольные варианты заполнения МП, то число допустимых вариантов заполнения МП возрастает примерно еще в $H!$ раз.

Поскольку не всякий вариант заполнения МП является корректным, необходимо сформулировать правила, позволяющие формировать только корректные МП.

1. правило строки. После ненулевого элемента в строке не должен быть ноль, т.е. если $q_{ij} = 1$ или 2, то $q_{ik} \neq 0$ для всех $k > j$ $(i, j, k = 1, \ldots, H)$;
2. правило столбца. Элементы в пределах одного столбца должны образовывать невозрастающую последовательность: $q_{i+1\,j} \leq q_{ij}$ $(i = 1, \ldots, H-1;\ \ j = 1, \ldots, H)$;
3. правило БП-группы. Классы заявок, образующие БП-группу, должны иметь одинаковые приоритеты по отношению к остальным классам заявок, т.е. если $q_{ii+1} = 0,$ то $q_{i+1\,j} = q_{ij}$ для всех $j = 1, \ldots, H$.

Если не выполняется хотя бы одно из перечисленных правил, то матрица приоритетов будет некорректной. В случае неканонической МП для ее проверки на соответствие перечисленным правилам необходимо преобразовать исходную матрицу в канонический вид путем перестановки строк и столбцов в соответствии с уровнем приоритетности, который подсчитывается как сумма всех элементов исходной матрицы в пределах каждой строки. Чем больше полученное значение, тем выше уровень приоритетности у соответствующего класса заявок.

Характеристики обслуживания заявок

Рассмотрим характеристики одноканальной СМО, в которую поступают H классов заявок, образующих простейшие потоки с интенсивностями $\lambda_1, \ldots, \lambda_H$. Длительность τ_{b_k} обслуживания заявок класса k распределена по произвольному закону с плотностью распределения вероятностей $b_k(\tau)$. Выбор заявок из очереди на обслуживание осуществляется в соответствии с ДО СП, заданной с помощью МП.

В качестве основной характеристики, описывающей эффективность функционирования системы, будем рассматривать время пребывания в системе τ_{u_k} заявок класса $k = 1, \ldots, H$, которое складывается из времени ожидания начала обслуживания τ_{x_k} и времени нахождения заявки на обработке τ_{v_k}, включающего в себя время ожидания заявки в прерванном состоянии:

$$\tau_{u_k} = \tau_{x_k} + \tau_{v_k}. \qquad (1)$$

Ниже при записи выражений для определения характеристик обслуживания заявок используются следующие обозначения.

1. $F_k(\tau) = \Pr(\tau_{f_k} < \tau)$ – функция распределения непрерывной случайной величины $\tau_{f_k} > 0$, причем f_k – один из следующего набора символов: b_k, x_k, z_k, v_k, w_k, u_k.
2. $f_k(\tau) = F_k'(\tau)$ – плотность распределения случайной величины τ_{f_k}.
3. $f_k^{(n)} = \int_0^\infty \tau^n f_k(\tau)\, d\tau$ – начальный момент порядка n $(n = 1, 2, \ldots)$ случайной величины τ_{f_k}, причем для простоты при записи первого начального момента (математического ожидания) верхний индекс будем опускать: $f_k = f_k^{(1)}$.

4. $F_k^*(s) = \int_0^\infty e^{-s\tau} f_k(\tau)\, d\tau \quad (s > 0)$ – преобразование Лапласа плотности $f_k(\tau)$.

5. $M_k^*(z) = \sum_{m=0}^{\infty} z^m P_k(m)$ – производящая функция дискретной случайной величины – числа заявок класса *k*, находящихся в системе, где $P_k(m)$ – вероятность того, что в системе находится *m* заявок класса *k*.

6. $R = \sum_{i=1}^{H} \rho_i$ – суммарная загрузка системы, где $\rho_i = \lambda_i b_i$ – загрузка, создаваемая *i*-заявками, причем предполагается, что $R < 1$, т.е. система работает без перегрузок.

7. $r_g(i,k)$ – коэффициенты, принимающие значения 0 или 1 в зависимости от значений элементов q_{ik} и q_{ki} матрицы приоритетов и позволяющие выделить классы заявок *i* и *k*, имеющие между собой один и тот же вид приоритета (ОП, АП, БП или любое их сочетание):

– $r_0(i,k) = 1$, если между заявками классов *i* и *k* нет приоритетов;

– $r_1(i,k) = 1$, если *i*-заявки имеют ОП по отношению к *k*-заявкам;

– $r_2(i,k) = 1$, если *i*-заявки имеют АП по отношению к *k*-заявкам.

Формулы для расчета коэффициентов $r_g(i,k)$ и их значения приведены в табл. 2.

Коэффициенты $r_0(i,k)$, $r_1(i,k)$ и $r_2(i,k)$ являются основными. На их основе формируются дополнительные коэффициенты $r_3(i,k)$, $r_4(i,k)$, $r_5(i,k)$ и $r_6(i,k)$, представляющие собой комбинацию основных коэффициентов. Отметим, что на значения коэффициентов оказывает влияние местоположение номеров классов в круглых скобках, и в общем случае $r_g(i,k) \neq r_g(k,i)$.

Коэффициенты $r_g(i,k)$	Значения элементов МП					
	q_{ik}	0	0	0	1	2
	q_{ki}	0	1	2	0	0
$r_0(i,k) = 0,5(1-q_{ik}-q_{ki})(2-q_{ik}-q_{ki})$		1	0	0	0	0
$r_1(i,k) = q_{ik}(2-q_{ik})$		0	0	0	1	0
$r_2(i,k) = 0,5\,q_{ik}(q_{ik}-1)$		0	0	0	0	1
$r_3(i,k) = r_0(i,k) + r_1(i,k)$		1	0	0	1	0
$r_4(i,k) = r_1(i,k) + r_2(i,k)$		0	0	0	1	1
$r_5(i,k) = r_0(i,k) + r_4(i,k)$		1	0	0	1	1
$r_6(i,k) = r_5(i,k) + r_1(k,i)$		1	1	0	1	1

Таблица 2. Значения коэффициентов $r_g(i,k)$

8. $\Lambda_k^{(g)} = \sum_{i=1}^{H} r_g(i,k)\lambda_i$ – частичные суммарные интенсивности потоков заявок.

9. $R_k^{(g)} = \sum_{i=1}^{H} r_g(i,k)\rho_i$ – частичные суммарные загрузки.

Величины $\Lambda_k^{(g)}$ и $R_k^{(g)}$ имеют также простое физическое толкование:

– $\Lambda_k^{(0)}$ и $R_k^{(0)}$ представляют собой суммарную интенсивность потоков и суммарную загрузку, создаваемые заявками всех классов, имеющих такой же приоритет, что и *k*-заявки, т.е. заявками тех классов, которые вместе с заявками класса *k* образуют БП-группу;

– $\Lambda_k^{(1)}$, $R_k^{(1)}$ и $\Lambda_k^{(2)}$, $R_k^{(2)}$ – суммарные интенсивности и загрузки, создаваемые заявками всех классов, которые обладают по отношению к *k*-заявкам более высоким ОП и АП соответственно, и т.д.

Для одноканальной системы с ДО СП, в которую поступают простейшие потоки заявок разных классов, аналитические зависимости для расчета характеристик функционирования системы получены с использованием метода введения дополнительного события [13] на уровне преобразований Лапласа и производящих функций.

Из (1) следует, что плотность распределения времени пребывания $u_k(\tau)$ заявок класса k представляет собой свертку плотностей $x_k(\tau)$ и $v_k(\tau)$. Тогда преобразование Лапласа плотности распределения $u_k(\tau)$: $U_k^*(s) = X_k^*(s)V_k^*(s)$.

Преобразование Лапласа плотности распределения $x_k(\tau)$ времени ожидания начала обслуживания имеет вид

$$X_k^*(s) = \frac{R_k^{(6)}\sigma_k + \sum_{i=1}^{H} r_1(k,i)\lambda_i [1 - B_k^*(\sigma_k)]}{s - \Lambda_k^{(0)} + \sum_{i=1}^{H} r_0(i,k)\lambda_i B_i^*(\sigma_k)]}, \tag{2}$$

где $\sigma_k = s + \Lambda_k^{(4)} - \Lambda_k^{(4)} D_k^*(s)$, а $D_k^*(s)$ определяется из уравнения:

$$\Lambda_k^{(4)} D_k^*(s) = \sum_{i=1}^{H} r_4(i,k)\lambda_i B_i^*(s + \Lambda_k^{(4)} - \Lambda_k^{(4)} D_k^*(s)). \tag{3}$$

Преобразование Лапласа плотности распределения $v_k(\tau)$ времени нахождения k-заявки на обработке

$$V_k^*(s) = B_k^*(s + \Lambda_k^{(2)} - \Lambda_k^{(2)} C_k^*(s)), \tag{4}$$

где $C_k^*(s)$ определяется из уравнения

$$\Lambda_k^{(2)} C_k^*(s) = \sum_{i=1}^{H} r_2(i,k)\lambda_i B_i^*(s + \Lambda_k^{(2)} - \Lambda_k^{(2)} C_k^*(s)). \tag{5}$$

В выражениях (2)–(5) используются следующие обозначения:

– $C_k^*(s)$ – преобразование Лапласа плотности распределения $c_k(\tau)$ периода занятости τ_{c_k} прибора заявками с более высоким АП, чем рассматриваемая k-заявка:

$$C_k^*(s) = \int_0^{\infty} e^{-s\tau} c_k(\tau)\, d\tau; \tag{6}$$

– $D_k^*(s)$ – преобразование Лапласа плотности распределения $d_k(\tau)$ периода занятости τ_{d_k} прибора заявками с более высоким ОП и АП, чем рассматриваемая k-заявка:

$$D_k^*(s) = \int_0^{\infty} e^{-s\tau} d_k(\tau)\, d\tau. \tag{7}$$

Выражения (2)–(7) позволяют определить все основные характеристики обслуживания заявок в системе, в частности, их начальные моменты.

Из (1) следует, что математическое ожидание u_k и второй начальный момент $u_k^{(2)}$ времени пребывания в системе k-заявок ($k = \overline{1,H}$) равны соответственно

$$u_k = x_k + v_k; \qquad u_k^{(2)} = x_k^{(2)} + 2x_k v_k + v_k^{(2)}.$$

Значения $x_k, x_k^{(2)}$ и $v_k, v_k^{(2)}$ определяются путем дифференцирования соответствующих выражений (2) и (4) в точке $s = 0$:

$$\left.\begin{aligned}
x_k &= \frac{\sum_{i=1}^{H} r_6(i,k)\lambda_i b_i^{(2)}}{2(1-R_k^{(4)})(1-R_k^{(5)})}; \qquad v_k = \frac{b_k}{1-R_k^{(2)}};\\
x_k^{(2)} &= \frac{\sum_{i=1}^{H} r_6(i,k)\lambda_i b_i^{(3)}}{3(1-R_k^{(4)})^2(1-R_k^{(5)})} + \frac{\sum_{i=1}^{H} r_5(i,k)\lambda_i b_i^{(2)} \sum_{i=1}^{H} r_6(i,k)\lambda_i b_i^{(2)}}{2(1-R_k^{(4)})^2(1-R_k^{(5)})^2} +\\
&+ \frac{\sum_{i=1}^{H} r_4(i,k)\lambda_i b_i^{(2)} \sum_{i=1}^{H} r_6(i,k)\lambda_i b_i^{(2)}}{2(1-R_k^{(4)})^3(1-R_k^{(5)})};\\
v_k^{(2)} &= \frac{b_k^{(2)}}{(1-R_k^{(2)})^2} + \frac{\sum_{i=1}^{H} r_2(i,k)\lambda_i b_i^{(2)}}{(1-R_k^{(2)})^3},
\end{aligned}\right\} \tag{8}$$

где $b_i^{(n)}$ – n-й начальный момент длительности обслуживания i-заявок ($i = \overline{1,H}$; $n = 1, 2, \ldots$).

На основе выражений (8) могут быть вычислены дисперсии и коэффициенты вариации соответствующих временных характеристик. При необходимости по двум моментам можно выполнить аппроксимацию вероятностных распределений [14].

Выражения (1)–(8) позволяют определить время ожидания в прерванном состоянии и полное время ожидания заявок каждого класса. В частности, их средние значения соответственно равны $(k=\overline{1,H})$

$$\left.\begin{aligned} z_k &= v_k - b_k = \frac{R_k^{(2)} b_k}{1-R_k^{(2)}}; \\ w_k &= x_k + z_k = \frac{\sum_{i=1}^{H} r_6(i,k)\lambda_i b_i^{(2)}}{2(1-R_k^{(4)})(1-R_k^{(5)})} + \frac{R_k^{(2)} b_k}{1-R_k^{(2)}}. \end{aligned}\right\} \qquad (9)$$

Характеристики обслуживания заявок ДО с одним классом приоритетов (ОП и АП) и ДО БП, как частные случаи ДО СП, могут быть получены на основе выражений (2)–(9).

При решении задач синтеза реальных систем, связанных, например, с оценкой емкости накопителя (буферной памяти), требуется определять число заявок, находящихся в системе.

Производящая функция числа заявок класса $k=\overline{1,H}$, находящихся в системе, связана с преобразованием Лапласа времени пребывания следующей зависимостью [15]:

$$M_k^*(z) = U_k^*(\lambda_k - \lambda_k z) \qquad (k=\overline{1,H}). \qquad (10)$$

Дифференцируя соответствующее число раз выражение (10) по z в точке $z = 1$, получим зависимости, связывающие начальные моменты распределений числа заявок и времени их пребывания в системе, в частности, для двух первых моментов:

$$m_k = \lambda_k u_k; \qquad m_k^{(2)} = \lambda_k^2 u_k^{(2)} + m_k \qquad (k=\overline{1,H}). \qquad (11)$$

Первое выражение в (11) представляет собой формулу Литтла [15], связывающую среднее число заявок в системе со средним временем их пребывания.

Заключение

Рассмотренные дисциплины обслуживания, построенные на основе матрицы приоритетов, обобщают традиционные дисциплины с одним классом приоритетов и расширяют множество возможных дисциплин обслуживания. Наибольший эффект от применения этих дисциплин состоит в возможности наилучшим образом обеспечить требуемое качество функционирования системы при наличии ограничений на времена пребывания в системе заявок разных классов. При этом задача сводится к определению значений элементов матрицы приоритетов, при которых выполняются заданные ограничения.

Литература

1. Олифер В.Г., Олифер Н.А. Компьютерные сети. Принципы, технологии, протоколы: Учебник для вузов. 3-е изд. СПб.: Питер, 2006. 944 с.
2. Aliev T.I., Nikulsky I.Y., Pyattaev V.O. Modeling of packet switching network with relative prioritization for different traffic types // Proc. 10th International Conference on Advanced Communication Technology (ICACT-2008). Phoenix Park, South Korea, 2008. Art. 4494220. P. 2174–2176.
3. Bogatyrev V.A. An interval signal method of dynamic interrupt handling with load balancing // Automatic Control and Computer Sciences. 2000. V. 34. N 6. P. 51–57.
4. Bogatyrev V.A. On interconnection control in redundancy of local network buses with limited availability // Engineering Simulation. 1999. V. 16. N 4. P. 463–469.
5. Вишневский В.М., Семенова О.В. Системы поллинга: теория и применение в широкополосных беспроводных сетях. М.: Техносфера, 2007. 312 с.
6. Муравьева-Витковская Л.А. Обеспечение качества обслуживания в мультисервисных компьютерных сетях за счет приоритетного управления // Изв. вузов. Приборостроение. 2012. Т. 55. № 10. С. 64–68.
7. Алиев Т.И. Проектирование систем с приоритетами // Изв. вузов. Приборостроение. 2014. Т. 57. № 4. С. 30–35.
8. Алиев Т.И., Муравьева Л.А. Система с динамически изменяющимися смешанными приоритетами и ненадежным прибором // Автоматика и телемеханика. 1988. Т. 49. № 7. С. 99–106.
9. Alfa A.S. Matrix-geometric solution of discrete time MAP/PH/1 priority queue // Naval Research Logistics. 1998. V. 45. N 1. P. 23–50.
10. Основы теории вычислительных систем / Под ред. С.А. Майорова. М.: Высшая школа, 1978. 408 с.
11. Гольдштейн Б.С. Об оптимальном приоритетном обслуживании в программном обеспечении ЭАТС. В кн. Системы управления сетями. М.: Наука, 1980. С. 73–78.
12. Zhao J.-A., Li B., Cao X.-R., Ahmad I. A matrix-analytic solution for the DBMAP/PH/1 priority queue // Queueing Systems. 2006. V. 53. N 3. P. 127–145.

13. Климов Г.П. Стохастические системы обслуживания. М.: Наука, 1966. 242 с.
14. Алиев Т.И. Аппроксимация вероятностных распределений в моделях массового обслуживания // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 2 (84). С. 88–93.
15. Клейнрок Л. Вычислительные системы с очередями: Пер. с англ. М.: Мир, 1979. 600 с.

Алиев Тауфик Измайлович – доктор технических наук, профессор, заведующий кафедрой, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, aliev@cs.ifmo.ru

Махаревс Эдуардс – доктор технических наук, хабилитированный доктор инженерных наук Латвийской Республики, профессор, профессор, Балтийская международная академия, Рига, LV-1019, Латвия, eduard@rostourism.lv

Taufik I. Aliev – D.Sc., Professor, Department head, ITMO University, Saint Petersburg, 197101, Russian Federation, aliev@cs.ifmo.ru

Eduards Maharevs – D.Sc., Professor, Baltic International Academy, Riga, LV-1019, Latvia, eduard@rostourism.lv

Принято к печати 11.09.14
Accepted 11.09.14

УДК 004.4'242

ПОСТРОЕНИЕ АВТОМАТНЫХ ПРОГРАММ ПО СПЕЦИФИКАЦИИ С ПОМОЩЬЮ МУРАВЬИНОГО АЛГОРИТМА НА ОСНОВЕ ГРАФА МУТАЦИЙ

Д.С. Чивилихин[a], В.И. Ульянцев[a], В.В. Вяткин[a,b,c], А.А. Шалыто[a]

[a]Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, chivdan@rain.ifmo.ru
[b] Университет Аалто, Хельсинки, FI-00076, Финляндия
[c] Технологический университет Лулео, Лулео, SE-971 87, Швеция

Аннотация. Традиционный в области разработки программ процесс тестирования не может гарантировать их корректности, поэтому при повышенных требованиях к надежности программ прибегают к верификации. Верификация позволяет проверять некоторые свойства программы во всех возможных ее состояниях, однако сам процесс верификации сложен. В методе проверки моделей (model checking) строится модель программы (обычно вручную), а требования к ней записываются на языке темпоральной логики. Выполнение или невыполнение этих требований к модели может быть проверено автоматически. Основной проблемой такого подхода является разрыв между программой и ее моделью. Парадигма автоматного программирования позволяет устранить указанный разрыв. В автоматном программировании логика работы программ описывается управляющими конечными автоматами, модели которых могут быть построены автоматически. В работе рассматривается применение муравьиного алгоритма на основе графа мутаций для решения задач построения автоматных программ по их спецификации, заданной сценариями работы и темпоральными свойствами. Апробация предложенного подхода проведена на примере задачи построения автомата управления дверьми лифта, а также на случайных данных. Полученные результаты показывают, что муравьиный алгоритм в два–три раза эффективнее ранее применявшегося генетического. Предложенный подход может быть рекомендован для автоматизированного построения управляющих программ для ответственных систем.

Ключевые слова: конечный автомат, муравьиный алгоритм, верификация.

Благодарности. Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01), при поддержке РФФИ в рамках научного проекта № 14-07-31337 мол_а.

AUTOMATA PROGRAMS CONSTRUCTION FROM SPECIFICATION WITH AN ANT COLONY OPTIMIZATION ALGORITHM BASED ON MUTATION GRAPH

D.S. Chivilikhin[a], V.I. Ulyantsev[a], V.V. Vyatkin[a,b,c], A.A. Shalyto[a]

[a]ITMO University, Saint Petersburg, 197101, Russian Federation, chivdan@rain.ifmo.ru
[a]Aalto University, Helsinki, FI-00076, Finland
[c] Luleå University of Technology, Luleå, SE-971 87, Sweden, valeriy.vyatkin@aalto.fi

Abstract. The procedure of testing traditionally used in software engineering cannot guarantee program correctness; therefore verification is used at the excess requirements to programs reliability. Verification makes it possible to check certain properties of programs in all possible computational states; however, this process is very complex. In the model checking method a model of the program is built (often, manually) and requirements in terms of temporal logic are formulated. Such temporal properties of the model can be checked automatically. The main issue in this framework is the gap between the program and its model. Automata-based programming paradigm gives the possibility to overcome this limitation. In this paradigm, program logic is represented using finite-state machines. The advantage of finite-state machines is that their models can be constructed automatically. The paper deals with the application of mutation-based ant colony optimization algorithm to the problem of finite-state machine construction from their specification, defined by test scenarios and temporal properties. The presented approach has been tested on the elevator doors control problem as well as on randomly generated data. Obtained results show the ant colony algorithm is two-three times faster than the previously used genetic algorithm. The proposed approach can be recommended for inferring control programs for critical systems.

Keywords: finite-state machine, ant colony algorithm, verification.

Acknowledgements. The work is partially financially supported by the Government of the Russian Federation (grant 074-U01), and also partially supported by RFBR (scientific project № 14-07-31337 мол_а).

Введение

Проектирование программного обеспечения является сложной задачей, особенно когда область применения требует высокого уровня надежности программ. В таких областях, как космическая и военная промышленность, авиация, энергетика и медицина, цена ошибок очень высока, так как их возникновение может привести к большим потерям ресурсов или нанести вред здоровью человека. Программы в таких случаях не могут подвергаться одной лишь процедуре тестирования, поскольку она может только указать на наличие ошибок, но не может гарантировать их отсутствия.

Метод проверки моделей (model checking [1]) применяется для проверки некоторых свойств программы во всех возможных ее состояниях. При традиционном подходе сначала строится модель рассматриваемой программной системы, а требования к модели записываются на языке темпоральной логики. Модель чаще всего строится вручную. Темпоральные свойства проверяются для модели, а не для исходной программы, что в общем случае указывает на разрыв между программой и ее моделью.

Парадигма автоматного программирования [2], в которой логика работы программ представляется в виде одного или нескольких управляющих конечных автоматов, позволяет преодолеть это ограничение. Программы, разработанные с помощью этой парадигмы, могут быть автоматически преобразованы в используемые в методе проверки моделей структуры [3, 4]. Таким образом, нет никакого разрыва между автоматными программами и их моделями.

Разрабатываемые в настоящей работе методы можно применять для построения систем управления, основанных на автоматах. Например, в стандарте проектирования распределенных систем управления IEC 61499 [1] конечные автоматы используются как ключевые управляющие элементы функциональных блоков. Методы, рассматриваемые в данной работе, могут быть применены для автоматизации различных этапов построения распределенных систем управления [5].

Основным достоинством автоматного программирования является возможность создавать программы автоматически по спецификации, которая может содержать, например, сценарии работы и темпоральные свойства. Для построения программ, в частности, применяются алгоритмы поисковой оптимизации, выполняющие направленный перебор решений-кандидатов. Так, в [3, 4] для этого применялись эволюционные алгоритмы. К сожалению, эти алгоритмы не обладают достаточной эффективностью, хотя и позволяют находить решение во много раз быстрее, чем прямым перебором. В [6] был предложен муравьиный алгоритм построения управляющих конечных автоматов, который для известной модельной задачи об «умном» муравье оказался существенно эффективнее генетического алгоритма.

Целью данной работы является экспериментальная проверка эффективности указанного муравьиного алгоритма на задаче построения автоматов по спецификации, состоящей из сценариев работы и темпоральных формул [4]. Эксперименты показали, что муравьиный алгоритм обладает более высокой производительностью по сравнению с ранее применявшимися генетическими алгоритмами. Настоящая работа основана на [7]. Был применен более современный метод настройки параметров алгоритмов, проведены более обширные экспериментальные исследования, в том числе на случайных данных.

Построение автоматных программ по спецификации

В работе управляющим конечным автоматом будем называть семерку $\langle X, E, Y, Z, y_0, \phi, \delta \rangle$, где X – множество булевых входных переменных; E – множество входных событий; Y – множество состояний; $y_0 \in Y$ – начальное состояние; Z – множество выходных воздействий; $\phi : Y \times E \times 2^X \to Y$ – функция переходов, а $\delta : Y \times E \times 2^X \to Z^*$ – функция выходов.

Проще говоря, управляющий конечный автомат – это такой детерминированный автомат, каждый переход которого помечен событием, булевой формулой от входных переменных и последовательностью выходных воздействий. Пример графа переходов управляющего автомата с тремя состояниями приведен на рис. 1. Начальное состояние выделено жирной рамкой. На каждом переходе перед косой чертой записано входное событие (*A* и *H*), в квадратных скобках записана булева формула от входных переменных x_1 и x_2, а после косой черты записана последовательность выходных воздействий.

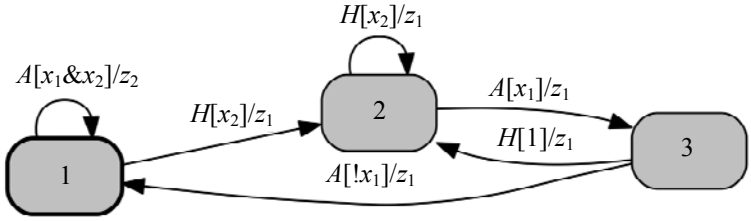


Рис. 1. Пример управляющего конечного автомата. Здесь $X = \{x_1, x_2\}$, $E = \{A, H\}$, $Y = \{1,2,3\}$, $y_0 = 1$, $Z = \{z_1, z_2\}$

В работе рассматриваются спецификации, содержащие сценарии работы и темпоральные формулы. Сценарием работы называется последовательность троек $\langle e, \varphi, O \rangle$, называемых элементами сценария, где $e \in E$ – событие, $\varphi \in 2^X$ – булева формула от входных переменных и $O \in Z^*$ – последовательность выходных воздействий (возможно, пустая). Говорят, что автомат удовлетворяет элементу сценария $\langle e, \varphi, O \rangle$ в состоянии y, если в этом состоянии автомат содержит переход, помеченный событием e, последовательностью выходных воздействий O и охранным условием, равным φ как булева формула. Автомат удовлетворяет сценарию работы, если он удовлетворяет всем последовательно обработанным элементам сценария в соответствующих состояниях.

[1] Function blocks: International Standard IEC 61499 Part 1: Architecture. International Electrotechnical Commission, 2011-12.

Для представления темпоральных свойств автомата в работе используется логика линейного времени (Linear Temporal Logic, LTL). Язык LTL состоит из пропозициональных переменных **Prop**, логических операторов **and**, **or**, **not** и набора темпоральных операторов, таких как, например, G (глобально в будущем), X (в следующий момент времени) и F (когда-либо в будущем). Для проверки того, удовлетворяет ли автомат LTL-формуле, применяется верификатор автоматных программ, разработанный в [3, 8]. Указанный верификатор принимает на вход автомат и LTL-формулу и либо выдает на выход контрпример (путь в модели Крипке [9], нарушающий формулу), либо сообщение о том, что автомат удовлетворяет формуле.

Задачу построения управляющего автомата по спецификации можно сформулировать так: требуется построить автомат с не более чем N_{states} состояниями, удовлетворяющий набору сценариев работы и множеству LTL-формул. В [10] показано, что уже задача построения автомата только по сценариям является, как минимум, NP-трудной. Этим обусловливается необходимость применения метаэвристических алгоритмов [11] при решении поставленной задачи.

Рассматриваемый в настоящей работе муравьиный алгоритм [6], как и генетический, представляет собой алгоритм направленного перебора. Степень соответствия каждого решения-кандидата поставленной задаче в таких алгоритмах оценивается с помощью так называемой функции приспособленности (ФП), которую необходимо максимизировать.

Функция приспособленности

Для оценки соответствия управляющих автоматов заданной спецификации используется ФП, предложенная в [8]. Первый компонент ФП f_{tests} оценивает, насколько хорошо автомат соответствует заданному набору сценариев работы, второй компонент f_{LTL} оценивает соответствие автомата темпоральным формулам. Выражение для ФП учитывает также число переходов автомата и имеет вид

$$f = f_{tests} + f_{LTL} + \frac{M - n_{transitions}}{100 \cdot M},$$

где $n_{transitions}$ – число всех переходов автомата, а M – число, гарантированно превосходящее $n_{transitions}$. Величины f_{tests} и f_{LTL} принимают значения от нуля до единицы, а последний член в указанной формуле для f не превышает 0,01. Таким образом, автоматы со значением ФП, большим двух, удовлетворяют всем сценариям работы и темпоральным формулам.

Муравьиный алгоритм построения автоматов

Основой метода построения конечных автоматов из [6] является представление пространства поиска (множества всех автоматов с заданными параметрами) в виде ориентированного графа G. Каждая вершина этого графа, который будем называть графом мутаций, ассоциирована с конечным автоматом, а ребра соответствуют мутациям автоматов.

Под мутацией конечного автомата понимается небольшое изменение его структуры – функций переходов φ и выходов δ. В данной работе используются следующие операторы мутации конечных автоматов.

1. **Изменение состояния, в которое ведет переход**. Для случайно выбранного перехода в автомате состояние y, в которое он ведет, заменяется другим состоянием, выбранным случайным образом из множества $Y \setminus \{y\}$.
2. **Добавление или удаление переходов**. Наличие некоторых переходов в состоянии автомата может сделать его противоречащим LTL-формуле. В связи с этим необходимо иметь возможность периодически удалять и соответственно добавлять переходы. Оператор мутации, предложенный в [4], сканирует состояния автомата и с определенной вероятностью изменяет набор переходов в выбранном состоянии. Случайным образом решается, добавить или удалить переход. Переход добавляется лишь в том случае, когда в этом состоянии нет перехода, помеченного какой-либо встречающейся в сценариях комбинацией входного события и булевой формулы. Тогда в состояние добавляется новый переход, который ведет в случайно выбранное состояние. В случае удаления перехода из текущего состояния удаляется случайно выбранный переход.

Пример графа мутаций приведен на рис. 2. Вершины соответствуют автоматам, а ребра – мутациям автоматов. Запись на ребре (2, H [x_2]) → 1 означает, что мутация изменила состояние, в которое ведет переход из состояния 2 по событию H и формуле x_2, на состояние 1. На каждом ребре (u, v) графа, как и в классических муравьиных алгоритмах [12], задаются две величины: так называемое значение эвристической информации η_{uv} и значение феромона $\tau_{\min}$. Эвристическая информация вычисляется по формуле

$$\eta_{uv} = \max\left(\eta_{\min}, f(v) - f(u)\right),$$

где $\eta_{\min}$ – небольшая положительная константа, например, 10^{-3}. В начале работы алгоритма всем ребрам графа сопоставляется начальное значение феромона $\tau_{\min} = 10^{-3}$.

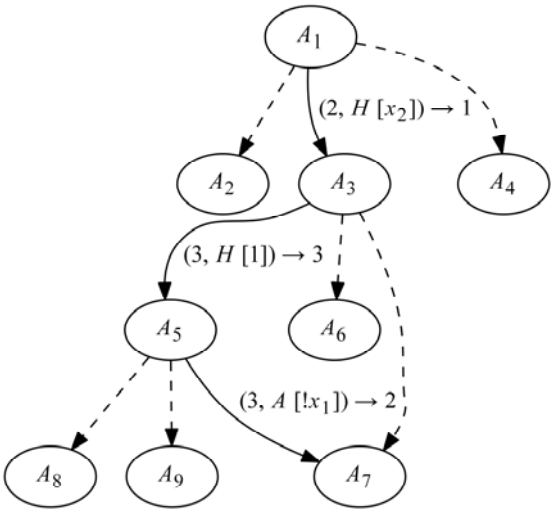


Рис. 2. Пример графа мутаций. Вершины соответствуют автоматам, а ребра – мутациям автоматов. Запись на ребре (2, H [x_2]) → 1 означает, что мутация изменила состояние, в которое ведет переход автомата из состояния 2 по событию H и формуле x_2, на состояние 1

В процессе работы алгоритма значения феромона изменяются муравьями, а значения эвристической информации остаются неизменными. На каждой итерации алгоритма выполняется две операции: построение решений муравьями и обновление значений феромона. Также происходит проверка заданных условий останова.

Построение решений муравьями

Процедуру построения решений муравьями можно разделить на два этапа. На первом этапе выбираются вершины графа, из которых муравьи начнут построение путей. В качестве единственной стартовой вершины для всех муравьев выбирается вершина, ассоциированная с лучшим найденным решением.

На втором этапе совершается одна итерация муравьиной колонии, в ходе которой каждый муравей перемещается по графу, начиная с соответствующей стартовой вершины. Пусть муравей находится в вершине u, ассоциированной с автоматом A. Если у вершины u существуют инцидентные ей ребра, то муравей выполняет одно из двух действий – построение новых решений либо вероятностный выбор. Если у вершины u нет инцидентных ей ребер, то муравей всегда выполняет построение новых решений.

1. **Построение новых решений**. С вероятностью p_{new} муравей пытается создать новые ребра и вершины графа G путем выполнения фиксированного числа N_{mut} мутаций автомата A. После выполнения муравьем всех мутаций он выбирает лучшую из построенных вершин и переходит в нее. Процесс обработки одной мутации автомата A таков:
   - выполнить мутацию автомата A, получить автомат A_{mut};
   - найти в графе G вершину t, ассоциированную с автоматом A_{mut}. Если такой вершины не существует, то создать ее и добавить в граф;
   - добавить в граф ребро (u, t).

   Отметим, что переход осуществляется даже в том случае, когда значение ФП у лучшего из построенных с помощью мутаций автоматов меньше значения ФП автомата A. Это свойство алгоритма позволяет ему выходить из локальных оптимумов.
2. **Вероятностный выбор**. С вероятностью $1 - p_{new}$ муравей выбирает следующую вершину из множества N_u ребер, инцидентных вершине u. Вершина v выбирается с вероятностью, определяемой классической в муравьиных алгоритмах формулой [12]:

$$p_{uv} = \frac{\tau_{uv}^{\alpha} \cdot \eta_{uv}^{\beta}}{\sum\limits_{w \in N_u} \tau_{uw}^{\alpha} \cdot \eta_{uw}^{\beta}},$$

   где $\alpha, \beta \in [0, 1]$ – параметры, определяющие значимость значений феромона и эвристической информации при выборе пути.

Каждый муравей в колонии делает по одному шагу до тех пор, пока все муравьи не остановятся. Каждый муравей может выполнить максимум n_{stag} шагов, на каждом из которых происходит построение новых решений либо вероятностный выбор без увеличения своего значения ФП. После этого муравей будет остановлен. Аналогично, колония муравьев может выполнить максимум N_{stag} итераций без увеличения максимального значения ФП. Затем алгоритм перезапускается.

Обновление значений феромона

Значение феромона, которое муравей откладывает на ребрах своего пути, равно максимальному значению ФП всех автоматов, посещенных этим муравьем. Для каждого ребра (u, v) графа G хранится число τ_{uv}^{best} – наибольшее из значений феромона, когда-либо отложенных муравьями на этом ребре. Последовательно рассматриваются все пути муравьев на текущей итерации алгоритма. Для каждого пути выделяется отрезок от начала пути до вершины, содержащей автомат с наибольшим на пути значением ФП $f_{\max}$. Для всех ребер из этого отрезка обновляются значения τ_{uv}^{best}: $\tau_{uv}^{best} \leftarrow \max\left(f_{\max}, \tau_{uv}^{best}\right)$. Затем значения феромона на всех ребрах графа G обновляются по формуле

$$\tau_{uv} \leftarrow \max\left(\tau_{\min}, (1-\rho)\tau_{uv} + \tau_{uv}^{best}\right),$$

где $\rho \in [0, 1]$ – скорость испарения феромона; $\tau_{\min} = 10^{-3}$ – минимальное разрешенное значение феромона. Введение нижней границы $\tau_{\min}$ исключает чрезмерное испарение феромона с ребер графа.

Настройка значений параметров алгоритмов

Для обеспечения адекватного сравнения эффективности муравьиного и генетического алгоритмов перед запуском экспериментов была проведена автоматическая настройка значений параметров алгоритмов с помощью программного средства *irace* [13]. На вход средству подается информация об интервалах возможных значений параметров настраиваемого алгоритма и множество экземпляров задачи, в данном случае – множество наборов сценариев и темпоральных формул. Это множество экземпляров задачи является обучающим для процесса настройки значений параметров. Средство *irace* реализует итеративную процедуру, в которой эффективность того или иного набора параметров настраиваемого алгоритма оценивается путем запуска этого алгоритма на экземплярах задачи из обучающего набора. По ходу работы *irace* с помощью статистического теста определяются неэффективные наборы параметров, которые затем отбрасываются.

Обучающее множество экземпляров задачи состояло из 200 наборов сценариев и темпоральных формул. При создании каждого из экземпляров задачи сначала генерировался случайный конечный автомат, число состояний N_{states} которого выбиралось случайно из отрезка [5, 10]. Остальные параметры автоматов имели следующие значения: $|E| = 2, |X| = 1, |Z| = 2$. По каждому автомату был сгенерирован набор из $5 \cdot |Y|$ сценариев общей длиной $100 \cdot |Y|$, а также две LTL-формулы.

На настройку параметров каждого алгоритма было отведено 12 ч на компьютере с процессором AMD Phenom(tm) II x4 955 3.2 GHz. В результате были получены значения параметров муравьиного и генетического алгоритмов, указанные в табл. 1, которые далее использовались во всех экспериментах.

Генетический алгоритм	
Параметр	**Значение**
Размер популяции	201
Доля элитных особей	0,25
Вероятность мутации	0,06
Число итераций до малой мутации поколения	483
Число итераций до большой мутации поколения	100

Муравьиный алгоритм	
Параметр	**Значение**
Число муравьев N_{ants}	4
Число мутаций N_{mut}	44
Максимальное число итераций колонии без увеличения значения ФП N_{stag}	28
Максимальное число шагов муравья без увеличения значения ФП n_{stag}	45
Скорость испарения феромона ρ	0,52

Таблица 1. Значения параметров муравьиного и генетического алгоритмов, полученные с помощью программного средства *irace*

Экспериментальное исследование предлагаемого подхода

Построение автомата управления дверьми лифта. В первой части экспериментального исследования рассматривалась задача управления конкретной системой (дверьми лифта) [8]. Система характеризуется пятью входными событиями: e_{11} (нажата кнопка «Открыть двери»), e_{12} (нажата кнопка «Закрыть двери»), e_2 (двери успешно открыты или закрыты), e_3 (препятствие помешало дверям закрыться), e_4 (двери заклинило). Существует три возможных выходных воздействия: z_1 (начать открывать двери), z_2 (начать закрывать двери), z_3 (послать аварийный сигнал). Спецификация, являющаяся входными данными для генетического алгоритма [3, 4] и применяемого в работе муравьиного алгоритма, состоит из

z_2 (начать закрывать двери), z_3 (послать аварийный сигнал). Спецификация, являющаяся входными данными для генетического алгоритма [3, 4] и применяемого в работе муравьиного алгоритма, состоит из 9 сценариев работы и 11 LTL-формул. Поиск осуществлялся среди автоматов с не более чем $N_{states} = 6$ состояниями. Автомат, решающий задачу, изображен на рис. 3, а.

В каждом эксперименте было проведено по 1000 запусков каждого алгоритма, каждый запуск продолжался до достижения лучшим решением значения ФП, равного 2,0075, что соответствует автомату, полностью удовлетворяющему заданной спецификации.

Для сравнения указанных алгоритмов измерялось медианное время работы и число вычислений ФП. Медианное время работы генетического алгоритма составило 7,3 с, а муравьиного – 3,5 с. Медианное число вычислений ФП генетического алгоритма равно 41902, а муравьиного – 10119. На рис. 3, б, приведена ящичная диаграмма, изображающая распределения запусков муравьиного и генетического алгоритмов по времени работы. Нижний и верхний края «ящика» изображают первый и третий квартили распределения, черта внутри ящика соответствует медиане, а горизонтальные линии снизу и сверху от ящика обозначают края статистически значимой выборки. Точки сверху от этих линий соответствуют выбросам.

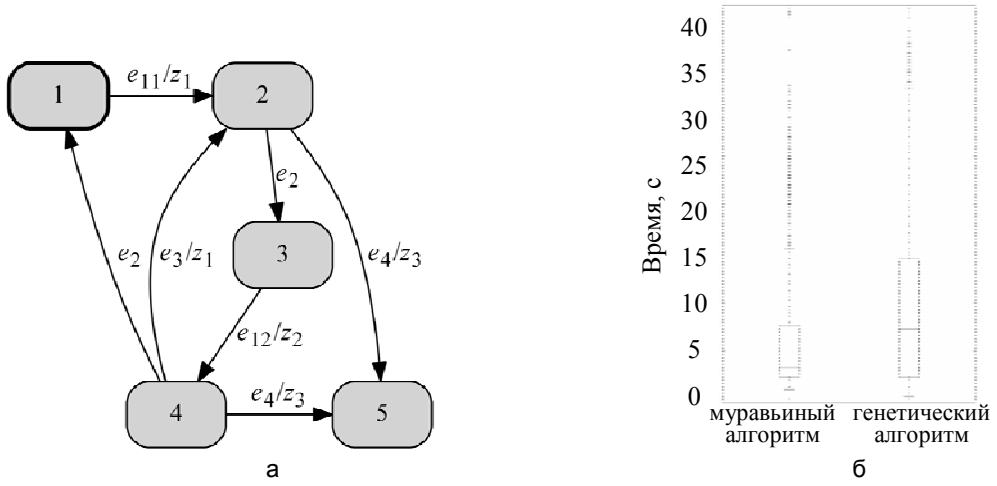


Рис. 3. Автомат управления дверьми лифта (а) и ящичная диаграмма распределения запусков муравьиного и генетического алгоритмов по времени работы (б)

Полученные в эксперименте результаты позволяют утверждать, что муравьиный алгоритм решает рассматриваемую задачу в два–три раза быстрее генетического алгоритма. Более того, как видно из рис. 3, б, муравьиный алгоритм, по сравнению с генетическим алгоритмом, обеспечивает меньший разброс времени работы. Статистическая значимость полученных результатов была подтверждена с помощью теста Уилкоксона [14]. Значение p-value, равное $3{,}93 \cdot 10^{-13}$, указывает на то, что вероятность совпадения среднего времени работы алгоритмов крайне мала.

Построение автоматов по случайным данным. Во второй части экспериментального исследования рассматривались случайно сгенерированные автоматы, содержащие от 4 до 10 состояний.

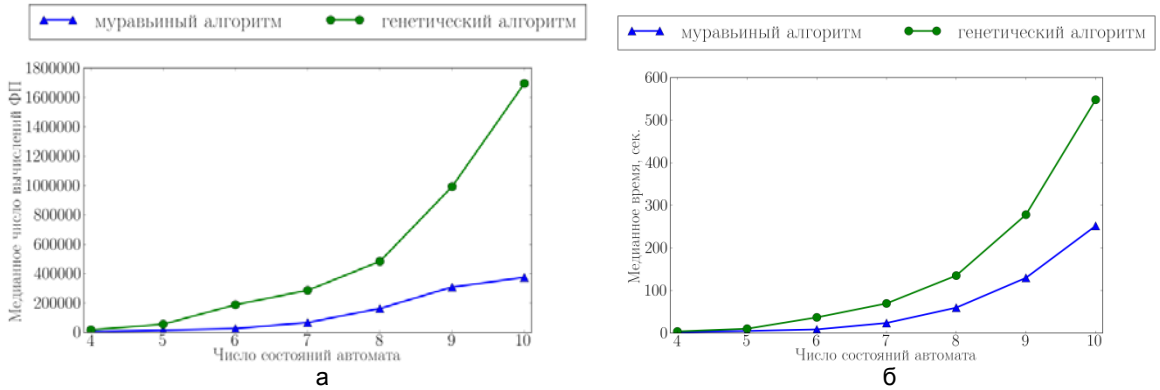


Рис. 4. Графики зависимости медианных значений числа вычислений ФП (а) и времени работы (б) от числа состояний автомата для генетического и муравьиного алгоритмов

Для каждого значения числа состояний N_{states} было сгенерировано по 100 автоматов со следующими параметрами: два входных события, два выходных воздействия, длина последовательности выходных воздействий от нуля до двух, булевы формулы зависят от единственной входной переменной. По каждому

из построенных автоматов был построен набор из $5 \cdot N_{states}$ сценариев работы общей длиной $100 \cdot N_{states}$ и две LTL-формулы. Формулы генерировались таким образом, чтобы они выполнялись для автомата, но не выполнялись для большого числа других автоматов.

N_{states}	p-value
4	$1 \cdot 10^{-7}$
5	$4 \cdot 10^{-6}$
6	$1 \cdot 10^{-10}$
7	$2 \cdot 10^{-6}$
8	$2 \cdot 10^{-4}$
9	$2 \cdot 10^{-6}$
10	$5 \cdot 10^{-4}$

Таблица 2. Результаты статистического теста Уилкоксона для экспериментов со случайно сгенерированными автоматами

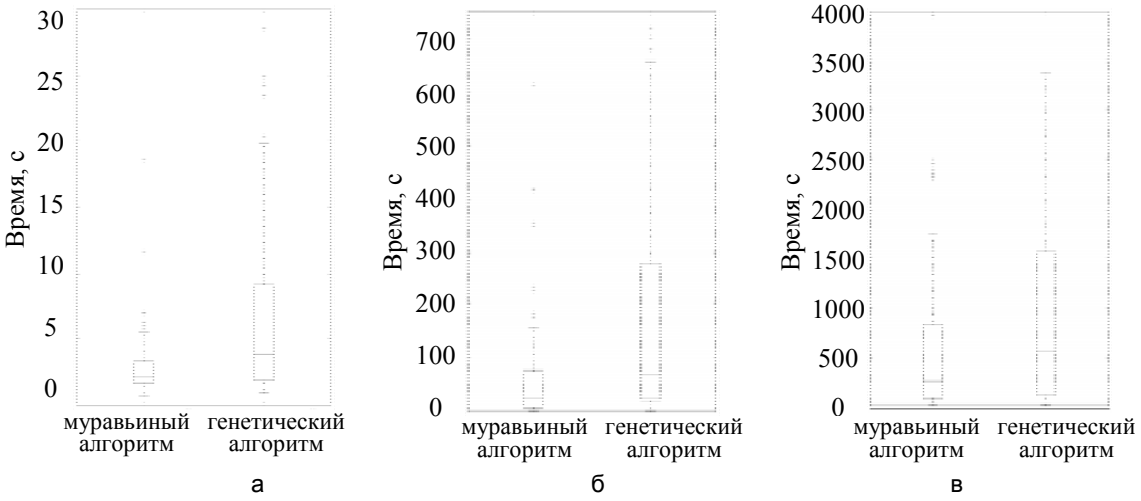


Рис. 5. Ящичные диаграммы распределения запусков муравьиного и генетического алгоритмов по времени работы для экспериментов со случайными автоматами с N_{states}: 4 (а); 7 (б); 10 (в)

Каждый эксперимент продолжался до нахождения алгоритмом решения, удовлетворяющего всем сценариям и темпоральным формулам. Замерялось число совершенных вычислений ФП, а также время работы алгоритмов. Графики медианных значений этих величин в зависимости от числа состояний автомата приведены на рис. 4, ящичные диаграммы распределений запусков алгоритмов по времени работы изображены на рис. 5. Результаты проведения статистического теста приведены в табл. 2. Полученные экспериментальные результаты позволяют сделать вывод о том, что на случайных данных муравьиный алгоритм существенно эффективнее генетического.

Заключение

В работе рассмотрено применение метода построения управляющих конечных автоматов на основе муравьиного алгоритма и графа мутаций к задаче построения автоматов по спецификации, включающей сценарии работы и темпоральные формулы. Проведенные экспериментальные исследования позволяют сделать вывод о том, что муравьиный алгоритм в два–три раза превосходит по производительности ранее применявшийся метод, основанный на генетическом алгоритме.

Литература

1. Clarke E.M., Grumberg O., Peled D.A. Model Checking. MIT press, 1999. 330 p.
2. Поликарпова Н.И., Шалыто А.А. Автоматное программирование. СПб: Питер, 2009. 176 с.
3. Егоров К.В., Царев Ф.Н., Шалыто А.А. Применение генетического программирования для построения автоматов управления системами со сложным поведением на основе обучающих примеров и спецификации // Научно-технический вестник СПбГУ ИТМО. 2010. № 5 (69). С. 81–86.
4. Tsarev F., Egorov K. Finite state machine induction using genetic algorithm based on testing and model checking // Proc. 13th Annual Genetic and Evolutionary Computation Conference, GECCO'11. Dublin, Ireland, 2011. P. 759–762.

5. Yang C.-H., Vyatkin V., Pang C. Model-driven development of control software for distributed automation: a survey and an approach // IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2014. V. 44. N 3. P. 292–305.
6. Chivilikhin D., Ulyantsev V. MuACOsm – a new mutation-based ant colony optimization algorithm for learning finite-state machines // Proc. 15th Genetic and Evolutionary Computation Conference, GECCO 2013. Amsterdam, Netherlands, 2013. P. 511–518.
7. Чивилихин Д.С., Ульянцев В.И., Шалыто А.А. Муравьиный алгоритм для построения автоматных программ по спецификации // Труды XII Всероссийского совещания по проблемам управления ВСПУ-2014. Москва, 2014. С. 4531–4542.
8. Егоров К.В. Генерация управляющих конечных автоматов на основе генетического программирования и верификации: дис. … канд. техн. наук. СПб.: НИУ ИТМО, 2013. 150 с.
9. Вельдер С.Э., Лукин М.А., Шалыто А.А., Яминов Б.Р. Верификация автоматных программ. СПб: Наука, 2011. 244 с.
10. Ульянцев В.И. Построение управляющих конечных автоматов по сценариям работы на основе решения задачи удовлетворения ограничений: магистерская диссертация [Электронный ресурс]. Режим доступа: http://is.ifmo.ru/diploma-theses/2013/master/ulyantsev/ulyantsev.pdf, свободный. Яз. рус. (дата обращения 30.09.2014).
11. Скобцов Ю.А., Федоров Е.Е. Метаэвристики. Донецк: НОУЛИДЖ, 2013. 426 с.
12. Dorigo M., Stützle T. Ant Colony Optimization. MIT Press, 2004. 319 p.
13. López-Ibáñez M., Dubois-Lacoste J., Stützle T., Birattari M. The irace package: iterated race for automatic algorithm configuration. Technical Report TR/IRIDIA/2011-004. IRIDIA, Belgium, 2011. 20 p.
14. Wilcoxon F. Probability tables for individual comparisons by ranking methods // Biometrics. 1947. V. 3. P. 119–122.

Чивилихин Даниил Сергеевич – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, chivdan@rain.ifmo.ru

Ульянцев Владимир Игоревич – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ulyantsev@rain.ifmo.ru

Вяткин Валерий Владимирович – доктор технических наук, профессор, Университет Аалто, Хельсинки, FI-00076, Финляндия; заведующий кафедрой, Технологический университет Лулео, Лулео, SE-971 87, Швеция, valeriy.vyatkin@aalto.fi

Шалыто Анатолий Абрамович – доктор технических наук, профессор, профессор, главный научный сотрудник, заведующий кафедрой, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, shalyto@mail.ifmo.ru

Daniil S. Chivilikhin – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, chivdan@rain.ifmo.ru

Vladimir I. Ulyantsev – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, ulyantsev@rain.ifmo.ru

Valeriy V. Vyatkin – D.Sc., Professor, Aalto University, Helsinki, FI-00076, Finland; Chaired Professor of Dependable Communications and Computations, Luleå University of Technology, Luleå, SE-971 87, Sweden, valeriy.vyatkin@aalto.fi

Anatoly A. Shalyto – D.Sc., Professor, Department head, ITMO University, Saint Petersburg, 197101, Russian Federation, shalyto@mail.ifmo.ru

Принято к печати 22.08.14
Accepted 22.08.14

УДК 004.056

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ ОРТОГОНАЛЬНЫХ ПРЕОБРАЗОВАНИЙ В ЧАСТОТНЫХ АЛГОРИТМАХ МАРКИРОВАНИЯ ЦИФРОВЫХ ИЗОБРАЖЕНИЙ

В.А. Батура[a], А.Ю. Тропченко[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, batu-vladimir@yandex.ru

Аннотация. Рассматривается эффективность использования ортогональных преобразований в частотных алгоритмах цифрового маркирования неподвижных изображений. Выбраны дискретное преобразование Адамара, дискретное косинусное преобразование и дискретное преобразование Хаара. Их эффективность определяется незаметностью встроенного с их помощью в изображение цифрового водяного знака, его стойкостью к наиболее распространенным операциям обработки изображения: JPEG-сжатию, зашумлению, изменению яркости и размера, эквализации гистограммы изображения. При использовании указанных ортогональных преобразований алгоритм цифрового маркирования его параметры встраивания остаются неизменными. Незаметность встраивания определяется величиной пикового отношения сигнала к шуму, стойкость водяного знака – коэффициентом корреляции Пирсона. Встраивание считается незаметным, если величина пикового отношения сигнала к шуму не ниже 43 дБ. Встроенный водяной знак считается устойчивым к определенной атаке, если коэффициент корреляции Пирсона не ниже 0,5. Для вычислительного эксперимента выбран алгоритм Elham, основанный на энтропии изображения. Вычислительный эксперимент проводится по следующему алгоритму: встраивание цифрового водяного знака при помощи алгоритма Elham в низкочастотную область изображения (контейнера), применение вредоносного воздействия на защищаемом изображении (стеганоконтейнере), извлечение цифрового водяного знака. Данные действия сопровождаются оценкой качества стеганоконтейнера и водяного знака, на основе которых и определяется эффективность ортогонального преобразования. В результате вычислительного эксперимента было установлено, что выбор указанных ортогональных преобразований при одинаковом алгоритме и параметрах встраивания не влияет на степень незаметности водяного знака. Основываясь на показателях корреляции, была установлена эффективность дискретного преобразования Адамара и дискретного косинусного преобразования по отношению к выбранным для эксперимента атакам. При этом использование дискретного преобразования Адамара повышает устойчивость встроенного водяного знака к изменению яркости и эквализации гистограммы стеганоконтейнера. Использование преобразования Хаара показало наименьшую эффективность. Полученные результаты будут полезны при разработке частотного алгоритма встраивания цифрового водяного знака в изображение.

Ключевые слова: преобразование Адамара, сжатие JPEG, стеганография, цифровое маркирование, цифровой водяной знак.

COMPARATIVE ANALYSIS OF APPLICATION EFFICIENCY OF ORTHOGONAL TRANSFORMATIONS IN FREQUENCY ALGORITHMS FOR DIGITAL IMAGE WATERMARKING

V.A. Batura[a], A.Yu. Tropchenko[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, batu-vladimir@yandex.ru

Abstract. The efficiency of orthogonal transformations application in the frequency algorithms of the digital watermarking of still images is examined. Discrete Hadamard transform, discrete cosine transform and discrete Haar transform are selected. Their effectiveness is determined by the invisibility of embedded in digital image watermark and its resistance to the most common image processing operations: JPEG-compression, noising, changing of the brightness and image size, histogram equalization. The algorithm for digital watermarking and its embedding parameters remain unchanged at these orthogonal transformations. Imperceptibility of embedding is defined by the peak signal to noise ratio, watermark stability– by Pearson's correlation coefficient. Embedding is considered to be invisible, if the value of the peak signal to noise ratio is not less than 43 dB. Embedded watermark is considered to be resistant to a specific attack, if the Pearson's correlation coefficient is not less than 0.5. Elham algorithm based on the image entropy is chosen for computing experiment. Computing experiment is carried out according to the following algorithm: embedding of a digital watermark in low-frequency area of the image (container) by Elham algorithm, exposure to a harmful influence on the protected image (cover image), extraction of a digital watermark. These actions are followed by quality assessment of cover image and watermark on the basis of which efficiency of orthogonal transformation is defined. As a result of computing experiment it was determined that the choice of the specified orthogonal transformations at identical algorithm and parameters of embedding doesn't influence the degree of imperceptibility for a watermark. Efficiency of discrete Hadamard transform and discrete cosine transformation in relation to the attacks chosen for experiment was established based on the correlation indicators. Application of discrete Hadamard transform increases stability of embedded watermark to the brightness changing and histogram equalization of the cover image. Haar transform application showed the lowest efficiency. These results will be useful in creation of frequency algorithm for embedding a digital watermark into an image.

Keywords: Hadamard transformation, JPEG compression, steganography, digital watermarking, digital watermark.

Введение

В последние годы для защиты авторских прав на мультимедийную продукцию активно применяются средства цифровой стеганографии, одним из которых является цифровое маркирование, при котором в объект защиты (контейнер) внедряется невидимый человеческому глазу цифровой водяной знак (ЦВЗ), представляющий собой двоичный код [1]. В настоящей работе для удобства восприятия ЦВЗ

представлен в виде бинарного изображения – логотипа. В качестве контейнера используем неподвижное (статическое) цифровое изображение.

Существует большое разнообразие методов цифрового маркирования, подразделяющихся на пространственные и частотные [1]. Пространственные методы основаны на изменении параметров пикселей изображения, как например, в работе [2]. Однако по сравнению с частотными пространственные методы не способны извлечь встроенный водяной знак из изображения после разных вредоносных воздействий, сильно искажающих изображение, например, зашумления, изменения размера, фильтрация. Данного недостатка лишены частотные методы, при которых ЦВЗ внедряется в частотные коэффициенты контейнера, полученные путем его разложения определенным ортогональным преобразованием. Однако из-за использования ортогональных преобразований частотные методы характеризуются высокой вычислительной сложностью [1]. Соответственно при разработке частотного метода цифрового маркирования неподвижных изображений возникает проблема выбора преобразования, обеспечивающего высокую скрытность внедрения ЦВЗ и его стойкость к вредоносным воздействиям.

На сегодняшний день существует относительно мало исследований, посвященных сравнению эффективности ортогональных преобразований разной вычислительной сложности в алгоритмах цифрового маркирования неподвижных изображений. Эффективность подобных алгоритмов определяется их способностью извлекать идентичный внедренному ЦВЗ или максимально похожий на него знак после вредоносных воздействий различного типа на защищаемое изображение (стеганоконтейнер). Как правило, сравнению подлежит использование дискретного косинусного преобразования (ДКП) и дискретного вейвлет-преобразования (ДВП). В подобных работах мало внимания уделяется эффективности алгоритмов, основанных на использовании ортогональных преобразований с малой вычислительной сложностью, что привело к необходимости проведения данного исследования. Целью работы является оценка эффективности использования ортогональных преобразований с малой вычислительной сложностью по сравнению с ДКП и ДВП в алгоритмах цифрового маркирования неподвижных изображений.

Выбор ортогональных преобразований

В качестве ортогональных преобразований в алгоритмах цифрового маркирования наиболее часто применяют дискретное косинусное преобразование [3, 4] и дискретное вейвлет-преобразование [5, 6] (в частности, дискретное преобразование Хаара (ДПХ) [7]), что связано с их использованием в форматах JPEG и JPEG 2000 соответственно.

Однако в ряде работ [8–16] для цифрового маркирования неподвижных изображений используется дискретное преобразование Адамара (ДПА). В некоторых исследованиях [9, 11, 12, 15, 16] для уменьшения вычислительной сложности используется двухмерное преобразование Адамара, которое реализуется строчно-столбцовым способом. При данном подходе к строкам матрицы исходных данных, а затем и к столбцам полученной матрицы применяются одномерные преобразования:

$$F_N = \frac{1}{N} A_N \left[X_N A_N \right] ,$$

где X_N – исходное изображение; F_N – преобразованное в набор коэффициентов изображение; N – размер изображения; A_N – матрица Адамара (ядро преобразования Адамара).

Матрица Адамара – квадратная двухуровневая матрица Hn порядка n, кратного 4, состоящая из элементов из множества $\{1, -1\}$, столбцы которой ортогональны [17].

В большинстве стеганографических исследований, как правило, применяется матрица порядка $N = 2^n$, где n – целое число. Подобные матрицы формируются на основе кронекеровского умножения матриц по формуле

$$A_{2N} = A_N \otimes A_2 = \begin{bmatrix} A_N & A_N \\ A_N & -A_N \end{bmatrix} , \qquad (1)$$

где $A_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ является матрицей Адамара наименьшего порядка.

В цифровом маркировании неподвижных изображений наиболее часто применяется матрица Адамара порядка 8 (рис. 1), что связано с использующимися при JPEG-компрессии блоками размером 8×8.

Одно из главных свойство матрицы Адамара – симметричность:

$A_2 = A_N^T$.

Число изменений знака строк матрицы аналогично частотной концепции преобразования Фурье [9]. Чем больше в строке матрицы Адамара перемен знака, тем более высокочастотными будут являться коэффициенты преобразования. Хотя стоит отметить, что в случае преобразования Адамара полученные коэффициенты считаются частотными условно. Коэффициент, находящийся в верхнем левом углу матрицы, содержит наибольшую энергию изображения (DC-коэффициент), остальные – AC-коэффициенты.

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}$$

Рис. 1. Матрица Адамара 8×8

По сравнению с ДКП и ДВП преимуществами ДПА являются [10] более низкая вычислительная сложность, простота аппаратной реализации, а также более низкий уровень искажения контейнера при условии учета особенностей зрительной системы человека. В настоящей работе произведена сравнительная оценка эффективности применения ДПА, ДКП и ДПХ в системах цифрового маркирования неподвижных цифровых изображений. Выбор ДКП и ДПХ для сравнения с ДПА обусловлен огромной распространенностью формата JPEG и большой перспективой формата JPEG 2000, показывающего лучшие результаты, особенно на больших коэффициентах сжатия [18].

Методика оценки эффективности

Оценка эффективности основана на выявлении отличий в устойчивости внедренного ЦВЗ к атакам на защищенное изображение (стеганоконтейнер) при использовании в выбранном алгоритме цифрового маркирования различных ортогональных преобразований. При этом параметры алгоритма, например, сила встраивания, тип ЦВЗ и контейнера, частотная область встраивания, должны оставаться неизменными. Кроме устойчивости, оценке подлежит уровень искажений, вносимых ЦВЗ в стеганоконтейнер.

Тип контейнера и ЦВЗ выбираем исходя из того, для каких изображений предназначен алгоритм цифрового маркирования. В качестве атак на стеганоконтейнер используем наиболее часто встречающиеся при обработке изображения операции:

- JPEG-сжатие с различным коэффициентом качества;
- зашумление (гауссов и спекл-шум с различным уровнем дисперсии, шум «соль и перец» с разной плотностью);
- среднечастотная фильтрация;
- изменение яркости и размера;
- эквализация гистограммы, приводящая к изменению контраста изображения.

Среди перечисленных атак JPEG-сжатие – самое распространенное вредоносное воздействие. В связи с этим ЦВЗ встраиваем в низкочастотную область контейнера, обеспечивающую наибольшую устойчивость к сжатию с потерями.

Существует достаточно большое количество метрик оценки качества изображения, большинство составляют разностные и корреляционные [19]. Пиковое отношение сигнала к шуму (PSNR) – одна из наиболее распространенных разностных характеристик определения качества изображения. PSNR определяется по формуле

$$\mathrm{PSNR} = 10\log_{10}\frac{255^2 \cdot M \cdot N}{\sum\limits_{x,y}^{M,N}(f(x,y)-\hat{f}(x,y))^2}, \qquad (2)$$

где $f(x,y)$ – контейнер; $\hat{f}(x,y)$ – стеганоконтейнер; x, y – координаты пикселей; M и N – высота и ширина изображения. PSNR измеряется в децибелах (дБ). Считается, что искажения незаметны для системы человеческого зрения в случае, если PSNR > 43 дБ.

В качестве меры качества извлеченного ЦВЗ используем коэффициент корреляции Пирсона – одну из наиболее часто применяемых корреляционных метрик в исследованиях стойкости внедренного ЦВЗ. Данный показатель измеряется по формуле

$$k = \frac{\sum\limits_c\sum\limits_r (A(c,r)-A_m)\cdot(B(c,r)-B_m)}{\sqrt{\sum\limits_c\sum\limits_r\left((A(c,r)-A_m)^2\right)\cdot\left(\sum\limits_c\sum\limits_r (B(c,r)-B_m)^2\right)}}, \qquad (3)$$

где c, r – координаты пикселя изображения; $A(c,r)$ – исходный ЦВЗ; $B(c,r)$ – извлеченный ЦВЗ; A_m, B_m – среднее арифметическое пикселей исходного и извлеченного ЦВЗ.

Считаем, что выбранное ортогональное преобразование эффективно к определенной атаке, если коэффициент корреляции Пирсона больше 0,5.

Вычислительный эксперимент

За основу тестирования взят однокоэффициентный алгоритм цифрового маркирования Elham [11], встраивающий черно-белый монохромный. Данный алгоритм использует показатель энтропии для сокрытия ЦВЗ в наиболее значимых областях изображения и является комбинацией двух ортогональных преобразований – ДПА и ДКП. ДПА – основное преобразование, предназначенное для перевода изображения-контейнера в частотную область после его предварительной декомпозиции на блоки размером 8×8. Причем ДПА подвергаются только блоки, у которых средняя энтропия пикселей превышает выбранный порог. ДКП является вспомогательным преобразованием, которое необходимо для уменьшения объема встраиваемого ЦВЗ путем сохранения лишь 15 низкочастотных коэффициентов, расположенных в порядке «зигзаг»-сканирования, начиная от DC-компонента. ЦВЗ также предварительно подвергается декомпозиции на блоки размером 8×8. Сохранившиеся 15 коэффициентов каждого блока ЦВЗ выстраивают в вектора с их последующим объединением в единый вектор. Каждый коэффициент вектора встраивается в блок частотных коэффициентов контейнера по формуле

$$c_i' = c_i + \alpha \cdot w_i, \quad (4)$$

где c_i – коэффициент контейнера, подлежащий изменению; c_i' – измененный коэффициент контейнера; w_i – встраиваемый элемент водяного знака; α – коэффициент усиления.

Данный метод относится к стегосистеме закрытого типа, так как для извлечения ЦВЗ требуется оригинальное изображение. Процесс извлечения происходит в порядке, обратном встраиванию, с той лишь разницей, что ДПА подвергается не только стеганоконтейнер, но и изображение-оригинал для определения местоположения модифицированных блоков. Извлечение осуществляется в соответствии с выражением

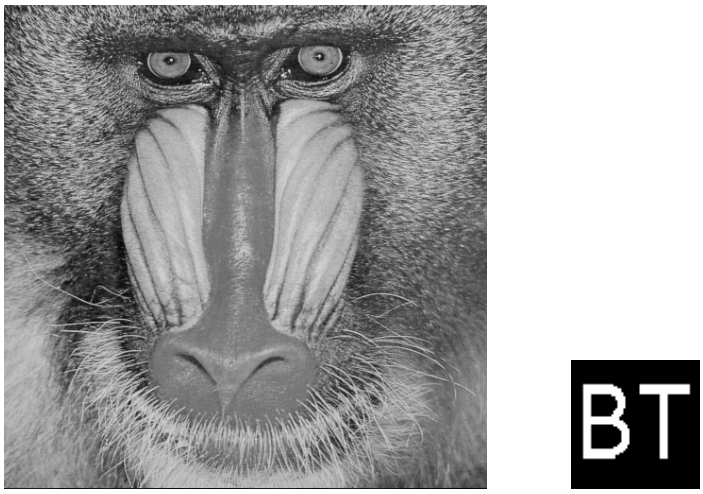
$$w_i = (c_i' - c_i)/\alpha. \quad (5)$$

В качестве программного средства реализации алгоритма выбрана система MATLAB, которая на данный момент является стандартом де-факто в области инженерных расчетов и включает в себя большой набор средств для реализации цифровой обработки сигналов и изображения.

Основные параметры вычислительной машины, выбранной для эксперимента:

- процессор: Intel Core i7 2630QM @ 2,00GHz;
- оперативная память: 4,00ГБ 1-канальная DDR3 @ 665 МГц.

В качестве контейнера выбрано полутоновое 8-битное изображение размером 512×512 пикселей (рис. 2, а), в качестве ЦВЗ – черно-белое монохромное изображение размером 64×64 пикселей (рис. 2, б). Пороговое значение энтропии равно 4,5, коэффициент усиления равен 35. Выбор данных значений обусловлен их оптимальностью для алгоритма Elham.



а б

Рис. 2. Контейнер (8-битное изображение размером 512×512 пикселей) (а); ЦВЗ (черно-белое монохромное изображение размером 64×64 пикселей) (б)

Результаты эксперимента

Для каждого ортогонального преобразования показатель PSNR равен 43,2104 дБ, что означает незаметность внедренного ЦВЗ. Показатель PSNR одинаков при использовании ДПА, ДКП и ДПХ первого и второго уровней, следовательно, выбор ортогонального преобразования не влияет на незаметность внедренного ЦВЗ при использовании одного и того же алгоритма внедрения.

На рис. 3 изображены графики зависимости качества ЦВЗ от различных показателей фактора качества JPEG (безразмерный коэффициент) при использовании ДПА, ДКП, ДПХ. Наилучшие показатели извлечения ЦВЗ при сжатии JPEG были достигнуты при использовании ДКП, что объясняется применением данного ортогонального преобразования в соответствующем алгоритме сжатия. При использовании ДПА были достигнуты близкие по значению показатели качества. Наихудший результат наблюдается при использовании ДПХ первого и второго уровней.

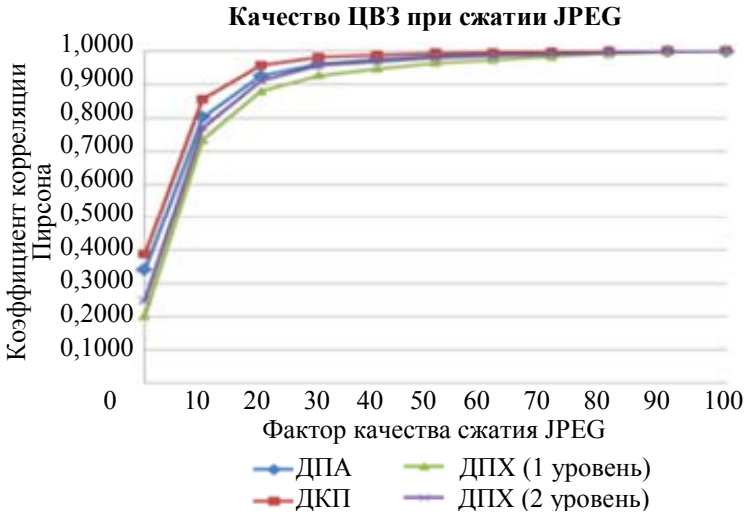


Рис. 3. Корректность извлеченного ЦВЗ в зависимости от показателя качества JPEG

Результаты сравнительного анализа эффективности ортогональных преобразований против других атак, выбранных в методике, приведены в таблице.

Тип атаки	ДПА	ДКП	ДПХ	ДПХ (двухуровневое)
Гауссов шум (0,001)	0,9562	0,9567	0,9535	0,9584
Гауссов шум (0,01)	0,7141	0,7166	0,6837	0,7136
Гауссов шум (0,1)	0,3258	0,2970	0,2873	0,3195
Шум «Соль и перец» (0,001)	0,9886	0,9857	0,9863	0,9882
Шум «Соль и перец» (0,01)	0,8823	0,8804	0,8738	0,8833
Шум «Соль и перец» (0,1)	0,4882	0,4692	0,4905	0,4668
Спекл-шум (0,001)	0,9877	0,9894	0,9885	0,9886
Спекл-шум (0,01)	0,9001	0,9083	0,9016	0,9026
Спекл-шум (0,1)	0,5476	0,5885	0,5207	0,5304
Изменение яркости (0,2)	0,9140	0,9018	0,5522	0,3356
Изменение яркости (0,3)	0,8510	0,8324	0,4227	0,2483
Изменение размера (256×256)	0,9214	0,9254	0,7205	0,9048
Изменение размера (128×128)	0,5375	0,5470	0,1639	0,5026
Среднечастотная фильтрация (3×3)	0,7440	0,7591	0,5077	0,6837
Эквализация гистограммы	0,5976	0,5741	0,4953	0,3120

Таблица. Показатели эффективности использования ортогональных преобразований

В таблице представлены показатели корреляции между встроенным и извлеченным ЦВЗ. Для гауссова и спекл-шумов с различным уровнем дисперсии, а также шума «соль и перец» с разной плотностью наблюдаем приблизительно одинаковую устойчивость при использовании ДПА, ДКП и преобразования Хаара первого и второго уровней.

При изменении размера стеганоконтейнера до размеров, указанных в таблице, или его среднечастотной фильтрации (фильтр размером 3×3) достигается приблизительно одинаковая устойчивость при использовании ДПА и ДКП. Наихудшие показатели при данных атаках получаем при использовании ДПХ первого уровня. ДПХ второго уровня увеличивает устойчивость ЦВЗ к данным атакам, однако при этом увеличивается и вычислительная сложность алгоритма цифрового маркирования.

При увеличении яркости стеганоконтейнера на 20% и 30%, а также при эквализации его гистограммы использование ДПА обеспечивает увеличение устойчивости ЦВЗ по сравнению с использованием ДКП и ДПХ.

Заключение

В работе проведено исследование эффективности ортогональных преобразований в частотных алгоритмах цифрового маркирования неподвижных изображений. Результаты вычислительного эксперимента показали, что выбор использованных в эксперименте ортогональных преобразований не влияет на незаметность внедренного ЦВЗ при их использовании в одном и том же алгоритме внедрения.

Применение дискретного косинусного преобразования и дискретного преобразования Адамара для внедрения ЦВЗ обеспечивают схожую устойчивость к наиболее распространенным операциям обработки изображения (сжатие JPEG, зашумление, среднечастотная фильтрация, эквализация гистограммы, изменение размера и яркости). При этом дискретное преобразования Адамара повышает устойчивость встроенного ЦВЗ к изменению яркости и эквализации гистограммы стеганоконтейнера.

Поскольку дискретное преобразование Адамара обладает наименьшей вычислительной сложностью, по сравнению с ДКП и ДПХ, а также простотой аппаратной реализацией, оно является эффективным в качестве основы для реализации частотных алгоритмов цифрового маркирования.

Литература

1. Грибунин В.Г., Оков И.Н., Туринцев В.И. Цифровая стеганография. М.: СОЛОН-Пресс, 2002. 272 с.
2. Su Q., Niu Y., Zhao Y., Pang S., Liu X. A dual color images watermarking scheme based on the optimized compensation of singular value decomposition // AEU – International Journal of Electronics and Communications. 2013. V. 67. N 8. P. 652–664.
3. Wu X., Sun W. Robust copyright protection scheme for digital images using overlapping DCT and SVD // Applied Soft Computing Journal. 2013. V. 67. N 2. P. 1170–1182.
4. Patra J.C., Phua J.E., Bornand C. A novel DCT domain CRT-based watermarking scheme for image authentication surviving JPEG compression // Digital Signal Processing: A Review Journal. 2010. V. 20. N 6. P. 1597–1611.
5. Bhatnagar G., Jonathan Wu Q.M. A new logo watermarking based on redundant fractional wavelet transform // Mathematical and Computer Modelling. 2013. V. 58. N 1–2. P. 204–218.
6. Bhatnagar G., Jonathan Wu Q.M., Raman B. A new robust adjustable logo watermarking scheme // Computers and Security. 2012. V. 31. N 1. P. 40–58.
7. Maheswari S., Rameshwaran K. A robust blind image watermarking based on double Haar wavelet transform (DHWT) // Journal of Scientific and Industrial Research. 2012. V. 71. P. 324–329.
8. Maity S.P., Kundu M.K. DHT domain digital watermarking with low loss in image informations // AEU – International Journal of Electronics and Communications. 2010. V. 64. N 3. P. 243–257.
9. Ho A.T.S., Shen J., Chow A.K.K., Woon J. Robust digital image-in-image watermarking algorithm using fast Hadamard transform // Proc. IEEE International Symposium on Circuits and Systems. 2003. V. 3. P. III826–III829.
10. Maity S.P., Kundu M.K. Perceptually adaptive spread transform image watermarking scheme using Hadamard transform // Information Sciences. 2011. V. 181. N 3. P. 450–465.
11. Shabanali Fami E., Samavi S., Rezaee Kaviani H., Molaei Radani Z. Adaptive watermarking in Hadamard transform coefficients of textured image blocks // 16th International Symposium on Artificial Intelligence and Signal Processing. Shiraz, Iran, 2012. V. 2012. Art. 6313799. P. 503–507.
12. Saryazdi S., Nezamabadi-pour H. A blind digital watermark in Hadamard domain // International Journal of Computer, Information, Systems and Control Engineering. 2007. V. 1. N 3. P. 784–787.
13. Разинков Е.В., Латыпов Р.Х. Встраивание цифрового водяного знака в изображение с использованием комплексного преобразования Адамара // Материалы II международной научной конференции по проблемам безопасности и противодействия терроризму. М.: МЦНМО, 2007. С. 509–514.
14. Bhatnagar G., Raman B. Robust watermarking in multiresolution Walsh-Hadamard transform // IEEE International Advance Computing Conference (IACC 2009). Patiala, Infia, 2009. Art. 4809134. P. 894–899.
15. Sarker I.H., Iqbal S. Content-based image retrieval using Haar wavelet transform and color moment // Smart Computing Review. 2013. V. 3. N 3. P. 155–165.
16. Ho A.T.S., Shen J., Tan S.H. A character-embedded watermarking algorithm using the fast Hadamard transform for satellite images // Proceedings of SPIE – The International Society for Optical Engineering. 2002. V. 4793. P. 156–167.
17. Балонин Ю.Н., Сергеев М.Б. Алгоритм и программа поиска и исследования М-матриц // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 3 (85). С. 82–86.

18. Востриков А.А., Чернышев С.А. Об оценке устойчивости к искажениям изображений, маскированных М-матрицами // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 5 (87). С. 99–103.
19. Sayood K. Introduction to Data Compression. Morgan Kaufmann Publ., 1996. 491 p.

Батура Владимир Александрович – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, batu-vladimir@yandex.ru

Тропченко Александр Ювенальевич – доктор технических наук, профессор, профессор, Университет ИТМО, 197101, Российская Федерация, tau@d1.ifmo.ru

Vladimir A. Batura – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, batu-vladimir@yandex.ru

Alexander Yu. Tropchenko – D.Sc., Professor, Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, tau@d1.ifmo.ru

Принято к печати 14.05.14
Accepted 14.05.14

УДК 004.6

ФИЗИЧЕСКИЕ РЕСУРСЫ ИНФОРМАЦИОННЫХ ПРОЦЕССОВ И ТЕХНОЛОГИЙ

М.О. Колбанёв[a], Т.М.Татарникова[b]

[a] Санкт-Петербургский государственный экономический университет, Санкт-Петербург, 191023, Российская Федерация
[b] Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация, tm-tatarn@yandex.ru

Аннотация.

Предмет исследования. Рассмотрены базовые информационные технологии, автоматизирующие информационные процессы сохранения, распространения и обработки данных, с точки зрения требуемых им физических ресурсов. Показано, что изучение этих процессов с такими традиционными для современной информатики целями, как способность передавать знания, степень автоматизации, обеспечение информационной безопасности, кодирование, надежность и других, уже недостаточно. Причиной этого являются, с одной стороны, увеличение объемов и интенсивности информационного взаимодействия в ходе предметной деятельности людей и, с другой стороны, приближение к пределу производительности информационных систем, основанных на полупроводниковых технологиях. Актуальной проблемой современных инженерных разработок стало создание таких технических средств, которые не просто обеспечивают поддержку информационного взаимодействия, но и потребляют рациональные объемы физических ресурсов. Таким образом, объектом исследования являются базовые информационные технологии, обеспечивающие сохранение, распространение и обработку данных для поддержки информационного взаимодействия людей, а предметом исследования – физические временные, пространственные и энергетические ресурсы, необходимые для реализации этих технологий.

Используемые подходы. Предпринимается попытка за счет учета в явном виде объемов физических ресурсов, необходимых для изменения состояний носителей информации, расширить возможности традиционной методологии кибернетики, которая заменяет рассмотрение материальной составляющей информации перебором состояний информационных объектов.

Цель работы. Выработка общего подхода к сравнению и последующему выбору базовых информационных технологий сохранения, распространения и обработки данных с учетом не только требований к качеству информационного взаимодействия в определенной предметной области и степени использования технических средств, но и объемов потребляемых при этом физических ресурсов.

Основные результаты работы. Предложена классификация ресурсов, потребляемых базовыми информационными технологиями, по их физической природе на пространственные, временные и энергетические. Показано, что основными пространственными ресурсами применительно к базовым информационным технологиям являются плотность записи данных, распределение пользователей в зоне охвата и размер техпроцесса, временными – время гарантированного сохранения, время доставки данных и производительность обработчика, энергетическими – уровни энергетического барьера и сигнала и энергопотребление. Выделены ключевые физические ресурсы для базовых информационных технологий сохранения, распространения и обработки данных, к которым отнесены соответственно плотность записи, время доставки и энергопотребление. На примере технологий сохранения данных предложен подход к выбору такой информационной технологии, которая удовлетворяет требованиям пользователей к качеству информационного взаимодействия при рациональном потреблении физических ресурсов.

Практическая значимость. Результаты работы могут быть полезны специалистам, занимающимся проектированием и эксплуатацией высокопроизводительных систем вычислений, хранения и распространения данных; разработкой способов повышения эффективности существующих коммуникаций, включая мобильную и оптическую связь, методов и алгоритмов для сбора, хранения и интеллектуального анализа больших объемов данных; внедрением новых информационных технологий.

Ключевые слова: базовые информационные процессы, информационные технологии, сохранение данных, распространение данных, обработка данных, пространственные, временные и энергетические ресурсы информационных технологий, принцип Г. Мура, принцип Р. Ландауэра, точка Т. Стерлинга.

PHYSICAL RESOURCES OF INFORMATION PROCESSES AND TECHNOLOGIES

M.O. Kolbanev[a], T.M. Tatarnikova[b]

[a] Saint Petersburg State University of Economics, Saint Petersburg, 191023, Russian Federation
[b] Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation, tm-tatarn@yandex.ru

Subject of study. The paper describes basic information technologies for automating of information processes of data storage, distribution and processing in terms of required physical resources. It is shown that the study of these processes with such traditional objectives of modern computer science, as the ability to transfer knowledge, degree of automation, information security, coding, reliability, and others, is not enough. The reasons are: on the one hand, the increase in the volume and intensity of information exchange in the subject of human activity and, on the other hand, drawing near to the limit of information systems efficiency based on semiconductor technologies. Creation of such technologies, which not only provide support for information interaction, but also consume a rational amount of physical resources, has become an actual problem of modern engineering development. Thus, basic information technologies for storage, distribution and processing of information to support the interaction between people are the object of study, and physical temporal, spatial and energy resources required for implementation of these technologies are the subject of study.

Approaches. An attempt is made to enlarge the possibilities of traditional cybernetics methodology, which replaces the consideration of material information component by states search for information objects. It is done by taking explicitly into account the amount of physical resources required for changes in the states of information media.

Purpose of study. The paper deals with working out of a common approach to the comparison and subsequent selection of basic information technologies for storage, distribution and processing of data, taking into account not only the requirements for the quality of information exchange in particular subject area and the degree of technology application, but also the amounts of consumed physical resources.
Main findings. Classification of resources consumed by the basic information technologies is suggested according to their physical nature. They are: spatial, temporal and energy resources. It is shown that the main spatial resources for basic information technologies are: data recording density, the users' distribution in the coverage area and size of engineering process; temporal resources are: time of guaranteed saving, data delivery time and the handler efficiency; energy resources include: the barrier and the signal energy levels and power consumption. Key physical resources are highlighted for basic information technologies of data storage, distribution and processing that include, respectively, recording density, delivery time and power consumption. We suggest an approach to the selection of such information technology that meets the users' needs to the quality of information exchange with the rational consumption of natural resources. An example of data storage technology is given.
Practical relevance. The results can be useful for specialists involved in the design and operation of high-performance computing, storage and distribution of data, developing the ways of improvement for the effectiveness of existing communications, including mobile and optical communications, methods and algorithms for collecting, storing and smart analysis of large amounts of data, introduction of new information technologies.

Keywords: basic information processes, information technologies, data storage, data distribution, data processing, spatial, temporal and energy resources of information technologies, Moore's Law, R. Landauer's principle, T. Sterling's point.

Введение

Проблема снижения ресурсоемкости различных видов производств уже давно стоит перед всем миром. Для многих государств, регионов, отраслей промышленности и отдельных предприятий экономия ресурсов становится приоритетной задачей. Решение этой задачи в соответствии с Федеральным законом № 261-ФЗ об энергосбережении является обязанностью не только промышленности, но и муниципальных учреждений, государственных органов. Не является исключением и отрасль информационных технологий, где задача снижения ресурсоемкости стала приоритетной. Многие исследователи, изучая информацию, отмечали, что любые ее преобразования основаны на физических законах. Например, А.А. Ляпунов [1] указывал на ограничения пространства, времени и энергии при выполнении информационных технологий, поскольку невозможны концентрация слишком большой массы знаков в ограниченном объеме пространства, получение новых знаков и их передача в новый носитель за слишком маленькое время и регистрация новых знаков слишком маленькой энергией. Н.Н. Моисеев считал, что за исключением потребности изучения целенаправленных действий в живой природе и обществе можно обойтись без термина «информация» и протекающие процессы описывать с помощью законов физики и химии [2]. Р. Ландауэр [3] ставит знак равенства между информационными и физическими процессами, поскольку «информация физична». Определение понятия информации, которое следует из работ Н. Винера [4], явно связывает информацию с ее физическими свойствами: «Информация – это обозначение содержания, полученное нами из внешнего мира в процессе приспосабливания к нему нас и наших чувств».

Несмотря на такое понимание, физические свойства информации все же находились на втором плане исследований информатиков всю вторую половину XX века. Этому есть объяснение. Согласно закону Г. Мура, объемные характеристики информационных систем росли экспоненциально [5, 6]. Все возрастающие вычислительные возможности полупроводниковых технологий давали возможность рассматривать поведение кибернетических систем исключительно как нематериальное, а информацию как нематериальную субстанцию, которая, тем не менее, переводит системы из одного состояния в другое и самым существенным образом влияет на принятие решений.

Методология кибернетики основывается на трех базовых составляющих: системном подходе, прикладной математике и цифровых информационных технологиях. Такие прикладные математические теории, как теория информации, теория принятия решений, теория массового обслуживания, теория управления, моделирование систем, математическая и формальная логики, теории алгоритмов и автоматов, теории формальных языков и грамматик, социальная информатика, исследование операций и другие и сегодня составляют основу «информационного» образования. Общее у этих теорий – это «переборный» или «цифровой» метод:

1. сначала интеллектуал должен сформулировать цель исследования;
2. затем для достижения этой цели необходимо или выбрать некоторые состояния системы, или перебрать состояния, или упорядочить состояния, или исключить некоторые состояния, или синтезировать новые состояния, и т.п.

Весь смысл исследования прячется в цели, а вся «физика» – в умении сократить перебор, который для сложных систем является достаточно большим.

Цифровая информационная технология при таком подходе призвана методами прикладной математики реализовать алгоритмы перебора состояний системы, описанных цифровыми массивами данных.

Кибернетика учитывает смысловую составляющую информации только через цель, которая формулируется вне системы, а материальную оставляющую рассматривает как изменение состояний объектов, не связанное с их физической природой.

В последние годы стало очевидным существование некоторого предела возможностей полупроводниковых технологий, и это обстоятельство заставляет вернуться к физическим основаниям информационных преобразований. Главным системным ограничением для суперхранителей, суперпереносчиков и суперобработчиков данных нашего времени является энергопотребление. Уже сегодня крупные центры обработки данных, системы коммутации и маршрутизации, суперкомпьютеры в процессе своей работы потребляют десятки мегаватт электроэнергии. Один маршрутизатор операторского класса, например, каждый год потребляет столько энергии, сколько выделяется при сжигании десятков тонн угля.

Особенностью современных информационных технологий [7, 8], использующих принцип фон Неймана, является необходимость многократного сохранения, распространения и обработки данных. Это означает, что объемы энергии, потребляемые каждым информационным битом за время его жизненного цикла, увеличиваются многократно.

У многих информационных технологий можно проследить взаимную зависимость уровня энергопотребления с другими физическими ресурсами, описывающими пространственные и временные параметры систем [9].

Возможности преобразования информационных битов при их сохранении, распространении и обработке зависит сегодня не только и не столько от существования того или иного программного обеспечения, перебирающего состояния систем. Первостепенное значение приобретает наличие физических ресурсов, поскольку именно использование физических ресурсов обеспечивает перемещение данных как материальных объектов во времени, в пространстве и изменение формы представления данных [10, 11].

Рассмотрим с общих позиций те физические ресурсы, которые необходимы для выполнения функций базовыми информационными технологиями сохранения, распространения и обработки данных [12]. Очевидно, что физические ресурсы самым существенным образом влияют на технологические возможности реализации базовых информационных процессов [13]. Целью настоящей работы является выработка общего подхода к сравнению и последующему выбору базовых информационных технологий сохранения, распространения и обработки данных с учетом не только требований к качеству информационного взаимодействия в определенной предметной области и степени использования технических средств, но и объемов потребляемых при этом физических ресурсов.

Пространственные ресурсы

Любая информационная технология требует пространства. Пространственные ресурсы измеряются в единицах длины и расстояния, описывают способы размещения информационных объектов и должны эффективно использоваться при реализации информационных процессов. К числу основных пространственных ресурсов информационных технологий можно отнести следующие:

- для технологий сохранения – размер запоминающих устройств для записи и плотность записи данных;
- для технологий распространения – территория, в пределах которой организуется информационное взаимодействие пользователей (зона охвата) и распределение (плотность) пользователей на этой территории;
- для технологий обработки – размер техпроцесса и количество транзисторов, размещаемых в одном чипе.

Кроме того, любая технология характеризуется объемом технологических помещений и допустимой плотностью размещения в них оборудования.

Плотность записи данных – это количество бит, которое размещается на единице площади (или объема) запоминающего устройства (ЗУ). Очевидно, что плотность обратно пропорционально зависит от размера физического элемента, сохраняющего бит. Пока размер атома – это нижний теоретический предел увеличения плотности записи. Дальнейшее уменьшение размера 1 бита связано с расщеплением атома и переходом на квантовые технологии.

Для создания энергонезависимой памяти сегодня наиболее широко используются законы супермагнетизма. Плотность записи современного жесткого диска – менее 700 ГБ на дюйм^2. Максимальная теоретическая плотность в случае использования технологии HAMR (магнитной записи с подогревом) составляет 5–20 ТБ на дюйм^2 и может быть достигнута в скором будущем. Размер минимальной единицы хранения при этом должен быть порядка 10 нм. Уже реализована возможность сохранения данных в ячейке памяти, состоящей из 12 магнитных атомов, в то время как обычный жесткий диск для хранения одного бита данных использует сотни тысяч атомов [http://habrahabr.ru/post/136414/]. Повышение плотности позволит создавать компактные, быстрые и энергетически эффективные устройства. Полупроводниковая энергонезависимая память позволяет сократить энергозатраты при записи, хранении и чтении дан-

ных и обеспечивает плотность записи в зависимости от используемого техпроцесса. Техпроцесс 10 нм может быть освоен в ближайшие годы.

Зоной охвата современных информационных сетей является вся Земля. Это стало возможно благодаря созданию Интернет, в основе которого лежат:

- стандартизированные правила взаимодействия;
- единое адресное пространство;
- совместимость внутренних и внешних данных для всех сетей и добровольность объединения.

Отдельные сети имеют собственную зону охвата, управляются собственными администрациями и характеризуются используемыми информационными технологиями, составом пользователей, их количеством, интенсивностью взаимодействия, мобильностью, распределением по территории и др.

Применительно к телефонным сетям процедура связи вне зоны охвата «домашней» сети называется роумингом. Она требует предварительной взаимной договоренности между операторами, поскольку предполагает согласованное использование ресурсов нескольких сетей. Особый вид роуминга позволяет пользоваться мобильной связью на морском и воздушном транспорте, а также получить доступ к спутниковым сетям.

Размер техпроцесса определяет плотность транзисторов на одном кристалле. В соответствии с законом Г. Мура производительность кремниевых интегральных микросхем и количество транзисторов на одном кремниевом кристалле удваивается каждые 18 месяцев, а их стоимость при этом уменьшается на 50%. Рост количества транзисторов в одном чипе означает уменьшение и размеров единичного транзистора, и ширины контактных дорожек. Уровень техпроцесса 2011–2012 г.г. – это 22 нм, что соответствует размещению около 1,5 млрд транзисторов на 160 мм^2.

Уменьшение размера техпроцесса позволяет не только увеличивать плотность хранения данных на полупроводниках, но и создавать более сложные и эффективные архитектуры процессоров, в частности, имеющие несколько вычислительных ядер и уровней кэш-памяти. Кроме того уменьшение техпроцесса позволяет сократить энергопотребление за счет перехода на новые типы транзисторов, уменьшения напряжение питания, отключения в режиме бездействия отдельных ядер, кэш-памяти или участков интегрированного графического ядра и др.

Плотность размещения оборудования оценивается при помощи целой группы параметров. Это и количество вычислительных операций (вычислительная плотность), и объем потребляемой энергии (энергетическая плотность), и скорость информационных каналов (сетевая плотность) на единицу площади оборудования и др.

Рекорд вычислительной плотности 2013 г. – это 1 Пфлоп/с на одну стойку площадью 1 м^2. Стойка состоит из 1024 вычислительных узлов, имеет совокупную емкость локального файлового хранилища узлов 0,5 ПБ и обеспечивает отвод более 0,4 МВт тепловой мощности за счет использования прямого жидкостного охлаждения [http://www.rscgroup.ru]. Новой технологией, позволяющей сократить издержки на создание ИТ-инфраструктуры, являются модульные центры обработки данных. Производственное помещение строится из сэндвич-панелей, снабжается необходимым количеством серверов и прочего инфраструктурного оборудования и может располагаться в любом месте пространства при наличии доступа к сетевым и энергетическим мощностям [14].

В табл. 1 сведены ключевые физические и технологические пространственные ресурсы, характеризующие базовые информационные технологии.

Пространственные ресурсы	Базовые информационные технологии		
	Сохранение	Распространение	Обработка
Физические	Площадь (объем) ЗУ	Зона охвата	Размер техпроцесса
Технологические	Плотность (объем) записи	Плотность пользователей	Плотность транзисторов

Таблица 1. Пространственные ресурсы базовых информационных технологий

Временные ресурсы

Особенность времени как ресурса заключается в том, что его нельзя запасти впрок, оно расходуется непрерывно и необратимо. Управлять временем можно лишь планируя продолжительность тех или иных операций, в том числе с учетом случайных факторов.

Временные ресурсы информационных технологий – это время, необходимое для выполнения информационных процессов или их отдельных этапов и фаз. Эти ресурсы могут быть разделены на две группы, характеризующие, во-первых, время предоставления услуг (обслуживания) при сохранении, распространении и обработке данных и, во-вторых, время доступа к информационным услугам.

Для каждого из базовых информационных процессов время обслуживания имеет свое название, отражающее специфику процесса. Для сохранения – это время гарантированного сохранения, для распространения – время доставки данных, для обработки – производительность обработчика.

Время гарантированного сохранения – это период времени, который начинается в момент записи данных на ЗУ и продолжается до тех пор, пока данные могут быть найдены на ЗУ, считаны и интерпретированы пользователем. Это время зависит от времени «жизни» минимальных единиц хранения, т.е. времени, в течение которого они сохраняют установленное состояние.

Примерами современных долговечных хранителей данных являются диски типа M-Disc, которые записывают данные на слое минерального материала, подобного камню, и гарантируют сохранность файлов на протяжении 1000 лет [http://www.mdisc.com].

Еще более выносливым является стеклянный диск. Он не имеет минерального слоя, устойчив к природным катастрофам, пожарам и излучениям, выдерживает условия открытого космоса, температуры, близкие к абсолютному нулю, и излучение Солнца. Одна из компаний дает гарантию в 100 лет на накопители, созданные на базе флэш-памяти с антикоррозийной защитой. Электроны в плавающем затворе транзисторов сохраняются тем дольше, чем ниже температура хранения [15]. Согласно новому открытию можно синтезировать частицу ДНК и записать в нее экзабайты данных. Затем в лиофилизированной форме ДНК можно сохранять теоретически тысячи лет.

Время доставки данных – это период времени, который начинается в момент поступления сигнала в канал связи и заканчивается по достижению данными заданной точки пространства (адресата).

Время доставки по сети связи включает время передачи данных от источника информации в канал связи, время перемещения сигнала по каналу между сетевыми центрами и время управления движением сигнала в сетевых центрах, таких как маршрутизаторы, серверы или телефонные станции. И в электрических, и в оптических сетях собственно время перемещения сигнала по каналу связи равно скорости света. Задержки передачи сигналов связаны с необходимостью обрабатывать адресную и другую управляющую информацию, сопровождающую данные при использовании коммутируемых сетей [16].

Пропускная способность канала – это наибольшая скорость передачи данных, измеряемая в бит/с, т.е. количество данных, которые сеть может передать за единицу времени между двумя оконечными устройствами. Она достигается при использовании оптимальных для данного канала настроек источника информации, когда на каждом такте работы канала каждый символ переносит максимально возможное количество бит данных.

Производительность (показатель, обратный времени обработки данных) – это количество операций обработки в секунду.

Основной задачей процесса обработки данных является получение нового массива данных из исходного при помощи некоторых алгоритмов. Для решения этой задачи в архитектуре фон Неймана задействованы вычислительные элементы и память, объединенные коммутационной сетью (интерконнектом). Вычислительные элементы – это процессоры, каждый из которых содержит несколько вычислительных ядер, память – это иерархически организованная система хранения программ и данных, включающая регистры, кэши, основную и внешнюю памяти [17].

В сложной архитектуре компьютеров скорость счета зависит не столько от свойств элементной базы, сколько от способов объединения процессоров, памяти и интерконнекта. Это обстоятельство подтверждают данные, приведенные в табл. 2.

Год	1997	2011	Изменение
Техпроцесс	250 нм	22 нм	↓ в 10 раз
Тактовая частота процессоров	1 ГГц	1–4 ГГц	↑ в 2,5 раза
Время переключения транзисторов	$250 \cdot 10^{-15}$ с	$3 \cdot 10^{-15}$ с	↓ в 100 раз
Максимальная производительность суперкомпьютера	1 Тфлоп/с	33 Пфлоп/с	↑ в 33000 раз

Таблица 2. Динамика изменения вычислительных характеристик компьютера

Дальнейшее увеличение производительности суперкомпьютеров требует увеличения числа вычислительных узлов и развития методов программирования параллельных вычислений, однако главным резервом увеличения производительности является уменьшение времени доступа при обращении к памяти и к интерконнекту для перемещения данных между вычислительными узлами.

Появление суперкомпьютеров производительностью до 1 Эфлоп/с (10^{18} флоп/с) ожидается до 2020 г. Обсуждается возможность приближения суперкомпьютеров к зетта-масштабу (10^{21} флоп/с) до 2030 г.

Время доступа – это интервал времени между моментами поступления заявки на предоставление информационной услуги до момента начала ее реализации. Оно зависит от способа использования ресурсов информационных технологий, таких как объем запоминающих устройств, каналов и процессоров или энергии [18]. Если за некоторым пользователем заранее закреплен достаточный физический и технологический ресурс, то время доступа будет малой величиной, которой можно пренебречь. Однако, как прави-

ло, информационные системы организуют доступ многих пользователей к ограниченному количеству ресурсов. При этом возникают коллизии, и пользователи вынуждены ожидать освобождения нужных им ресурсов, если они уже используются другими пользователями. Если количество ресурсов системы рассчитано таким образом, что время доступа не превышает согласованной с пользователем величины, то систему называют системой реального времени [19].

В табл. 3 сведены ключевые физические и технологические временные ресурсы, характеризующие базовые информационные технологии.

Временные ресурсы	Базовые информационные технологии		
	Сохранение	Распространение	Обработка
Физические	Время «жизни» минимальной единицы хранения	Время доставки знака	Время переключения транзисторов
Технологические	Гарантированное время хранения	Пропускная способность	Производительность

Таблица 3. Временные ресурсы базовых информационных технологий

Энергетические ресурсы

Вслед за увеличением объема и интенсивности информационных потоков и охватываемой ими территории малая энергия, требуемая для управления малыми информационными потоками, перерастает в большую. В результате информационные системы потребляют сегодня колоссальное количество электроэнергии.

По всему миру на снабжение информационного и телекоммуникационного оборудования сейчас тратится около 160 ГВт, что составляет 8% от всей вырабатываемой электроэнергии, и эти показатели продолжают быстро расти [20]. По различным оценкам, к 2020 г. потребность оборудования информационных систем в электроэнергии увеличится более чем в два раза и достигнет 400 ГВт. Основными потребителями являются оконечные устройства, центры обработки данных и оборудование сетей.

Бит как единица оценки количества данных уже недостаточен для сравнения возможностей и эффективности информационных систем. Имеют значение и физический размер бита, и время его гарантированного сохранения, и энергия, необходимая для сохранения, передачи и обработки бита.

Эффективность информационных систем связана сегодня с фактическим потреблением ими физических ресурсов (в первую очередь электроэнергии**)** и оценивается, например:

- объемом энергии, потребляемой в расчете на единицу информационных услуг;
- стоимостью транзакций в киловатт-часах или объеме выбросов углерода;
- объемами выбросов углерода в пересчете на один сервер или на группу пользователей;
- соотношением энергопотребления информационного оборудования и инженерных систем, поддерживающим его работу;
- энергопотреблением на 1 м^2 площади технических помещений и т.д.

Р. Ландауэр в 1961 г. показал [7], что расход энергии в процессе вычислений связан с уничтожением битов данных, и сформулировал следующий принцип: «Независимо от физики и технологии вычислительного процесса при потере 1 бита данных в процессе вычисления как минимум выделяется энергия, равная $k_{\text{В}}T\ln 2$, Дж», где $k_{\text{В}}$ – постоянная Больцмана, определяющая связь между температурой и энергией (порядка $1{,}3807\cdot10^{-23}$ Дж/К); T – температура, при которой ведутся вычисления (300 К = 26,85°С). Остальные операции (копирование, установка, перенос и др.) требуют сколь угодно мало энергии при достаточно малой скорости протекания. В 2012 г. были представлены результаты экспериментов, подтвердивших этот результат.

Объяснение этого принципа очевидно. Обработка бита является операцией над двумя битами. Поскольку биты материальны и имеют размер, то и для перехода в состояние 0 или 1 они должны получить энергию. При выполнении операции один из входных битов превращается в результат операции на выходе, а другой теряется, и его энергия выделяется в виде тепла и излучений.

Количество тепла, которое выделяется при стирании 1 бита, очень мало. Но в архитектуре фон Неймана значения битов в памяти переписывается с огромной частотой, и выделяемую энергию уже нельзя не учитывать. Тем более, что уровень энергетических затрат на обработку одного бита при технологии 22 нм лежит в пределах $\left(k_{\text{В}}\cdot T\cdot 10^{5}\right)-\left(k_{\text{В}}\cdot T\cdot 10^{6}\right)$, т.е. в миллион раз больше, чем минимально возможный, а с учетом сопутствующих потерь на 10 Пфлоп/с сегодня тратится порядка 10 МВт электроэнергии.

То, что на языке программистов называется изменением данных в памяти, на физическом уровне означает рассеивание в пространстве тепла и излучений. Таким образом, энергопотребление – это главное системное ограничение для будущих информационных технологий.

При сохранении минимальные единицы хранения данных должны быть отделены друг от друга и от среды достаточно сильными энергетическими барьерами. Вероятность их искажения зависит от многих физических и химических факторов и определяется законом Аррениуса. Считается, что для сохранения минимальных единиц хранения в течение миллиона лет нужен энергетический барьер 60–70 КВт.

При распространении данных сигналы подвергаются воздействию помех. Для достоверной доставки следует или увеличивать уровень сигнала на этапе физического кодирования символов сообщений, или применять алгоритмы помехоустойчивого кодирования, фактически заменяя один символ группой символов. И в одном, и в другом случае требуется энергия. Теплотехнический консорциум (сообщество инженеров-теплотехников, занятых в области производства компьютерной техники) исследовал технические характеристики компьютерного оборудования и проанализировал тенденции развития новых средств вычислительной техники. Результаты исследований показали, что наибольшую удельную тепловую нагрузку создает телекоммуникационное оборудование.

Обработка данных. Т. Стерлинг в 2009 г. на конференции по суперкомпьютерам в Гамбурге предположил, что экзафлопсный рубеж окажется пределом развития современных суперкомпьютеров. Точка Стерлинга – это условное ограничение производительности суперкомпьютера, построенного на доступных сегодня технологиях [20]. Это ограничение следует из принципа Р. Ландауэра.

Пусть задано энергопотребление суперкомпьютера в 20 МВт = $20 \cdot 10^6$ Вт, что близко к пределу при электронных технологиях. Если разделить эту величину на $150\, k_{\text{в}} T$, где 150 – это эмпирический коэффициент минимального уровня энергии на обработку одного бита для надежной работы компьютера, то при комнатной температуре получим около $4 \cdot 10^{26}$ операций/с. Выполнение одной операции над 64-разрядным числом с плавающей точкой требует 20 тыс. однобитовых операций, поэтому в пересчете на операции с плавающей точкой получаем $2 \cdot 10^{22}$ флоп/с. Последнюю оценку надо уменьшить более чем на 2 порядка, учитывая недостатки материалов и технологий производства. В результате приходим к выводу Т. Стерлинга: максимальная производительность суперкомпьютера при современных технологиях находиться в пределах 32–128 Эфлоп/с и, вероятно, никогда не превзойдет величины 64 Эфлоп/с. Это означает, что закон Г. Мура перестанет действовать в близком будущем.

В табл. 4 сведены ключевые физические и технологические энергетические ресурсы, характеризующие базовые информационные технологии.

Энергетические ресурсы	Базовые информационные технологии		
	Сохранение	Распространение	Обработка
Физические	Энергетический барьер	Уровень сигнала	Энергозатраты на обработку бита
Технологические	Энергопотребление при сохранении	Энергопотребление при распространении	Энергопотребление при обработке

Таблица 4. Энергетические ресурсы базовых информационных технологий

Ресурсная модель базовых информационных технологий

Для описания ресурсного обеспечения базовых информационных технологий может быть использован параллелепипед (рис. 1), грани которого отображают нижние и верхние границы пространства, времени и энергии, необходимые информационным технологиям на некотором этапе их развития. К соответствующим значениям следует стремиться при выборе информационной техники [19]. В этой модели ось ординат (S) отображает пространственные, ось аппликат (F) – энергетические, ось абсцисс (T) – временные параметры технологии. Точки, лежащие в пределах объема параллелепипеда и имеющие координаты $0 \le S \le S^{\max}$, $0 \le T \le T^{\max}$, $0 \le F \le F^{\max}$, соответствуют некоторым уже реализованным или еще разрабатываемым технологиям.

В качестве примера рассмотрим технологии сохранения больших данных, которым также требуются все большие объемы физических ресурсов. При этом необходимо учитывать требования, в том числе к:

- плотности записи, поскольку от нее зависят размеры и количество носителей;
- величине гарантированного времени сохранения данных;
- величине энергозатрат, которые необходимы для записи/считывания данных на носитель и для защиты носителя от внешних воздействий между моментами записи и считывания.

Перечисленные пространственные, временные и энергетические параметры цифровых технологий сохранения зависят друг от друга. Улучшение любого из них может быть, как правило, достигнуто только за счет ухудшения других. Это хорошо демонстрируют, например, технологии полупроводниковой памяти SLC, MLC, TLC, в которых увеличение плотности записи достигается за счет снижения гарантированного времени хранения. То же самое можно отнести и к магнитной памяти, где увеличение плотности (уменьшение площади доменов) ведет к появлению взаимного влияния магнитных полей и, следовательно, снижает гарантированное время хранения и увеличивает энергетические затраты для защиты (восста-

новления) данных в процессе хранения. Полупроводниковая память характеризуется малым энергопотреблением по сравнению с магнитной, но, заменяя магнитную память полупроводниковой, следует учитывать, что за экономию энергии придется заплатить меньшим временем гарантированного хранения. Уменьшение энергопотребления дисковыми массивами возможно за счет остановки или уменьшения скорости вращения дисков, но это увеличивает время доступа к данным, и т.д.

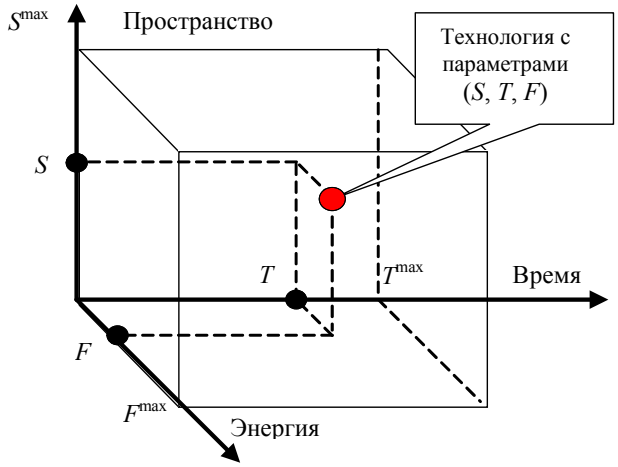


Рис. 1. Модель физических ресурсов информационных процессов

В общем случае эффективность технологии сохранения тем выше, чем больше плотность и гарантированное время хранения данных и меньше затрачиваемая при этом энергия, поэтому требованиям пользователей к реализации процесса сохранения соответствует параллелепипед с гранями S^{cox}, T^{cox}, F^{cox} на рис. 2.

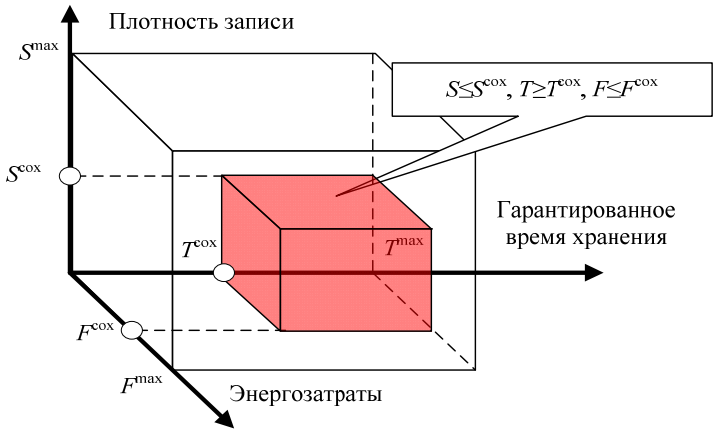


Рис. 2. Допустимая область параметров для технологии сохранения данных

В качестве базовых технологий для долговременного сохранения данных могут использоваться магнитные диски и ленты, полупроводниковая память, компакт и М-диски и др. В каждом из этих технологических сегментов существует множество альтернативных реализаций, в том числе и у разных производителей. Проведенные исследования показали, что каждая технология характеризуется собственными физическими параметрами:

- магнитные диски обеспечивают высокую плотность, но малое время хранения;
- магнитные ленты гарантируют длительное время хранения, но характеризуются и сравнительно значительным временем доступа;
- полупроводниковая память позволяет экономить энергию, но не способна на долговременное хранение;
- компакт-диски гарантируют длительное время хранения, но имеют ограниченную плотность записи;
- М-диски наиболее долговечны, но энергозатратны в процессе записи и т.д.

Примерами технических решений, принятых и реализованных при сохранении больших данных, могут служить следующие системы.

Большой адронный коллайдер является источником огромного количества научных данных о результатах столкновений элементарных частиц. Оцифрованные данные только о тех столкновениях, которые дали интересные с точки зрения физики результаты, поступают со скоростью около 50 событий в

секунду. В главном центре хранения и обработки эти данные записываются на магнитную ленту. К настоящему времени на десятках тысяч картриджей уже собрано более 100 ПБ данных. Доступ к картриджам автоматизирован: они хранятся в специальных подвальных помещениях на полках, откуда их достает робот. Энергопотребление системы хранения составляет 3,5 МВт. Фрагменты данных копируются на перекрывающий кэш диска для доступа и распределения между исследовательскими центрами по всему миру. Особенности данной системы – это отсутствие жестких ограничений на время обработки данных, что позволяет уменьшить энергозатраты.

Другой пример – это система долговременного хранения данных ColdStorage компании Facebook. Она располагается в отдельном здании и оптимизирована с точки зрения уменьшения энергопотребления и увеличения плотности размещения данных, а не производительности и доступности. Для этого используются магнитные диски, которые не рассчитаны на постоянную эксплуатацию, но позволяют менять скорость вращения, увеличивать количество дисков в одной стойке и уменьшать количество одновременно вращающихся дисков. В результате выбора именно такой технологии хранения для энергопитания массива дисков емкостью 1 ЭБ требуется примерно 0,375 МВт вместо 1,5 МВт.

Еще одно решение от Facebook – это экспериментальное хранилище, которое состоит из трехсот тысяч оптических дисков, где хранятся 30 ПБ данных. Нужный диск с требуемыми файлами находит робот. В перспективе система оптического хранения сможет сохранять до 150 ПБ. Эта система заметно увеличивает время доступа к затребованным файлам, но позволяет увеличить гарантированное время сохранения и на 80% снизить энергопотребление.

Выбор того или иного решения должен основываться на использовании обобщенного критерия или сравнении доступных технологий друг с другом по всем физическим характеристикам, которые могут быть отображены в виде параллелепипеда, подобного представленному на рис. 2. В процессе такого сравнения следует, используя принцип Парето, удалить из рассматриваемого набора такие технологии, которые заведомо хуже других по пространственным, временным и энергетическим характеристикам в совокупности. Если технологий, превосходящих другие хотя бы по одному показателю, останется достаточно много, то следует применить один из методов многокритериального выбора.

Набор параметров задачи оптимизации зависит при этом от вида базовой технологии сохранения данных. Оптимизируемыми параметрами могут являться, например, скорость вращения и количество одновременно вращающихся дисков, плотность дисков на стойке, пространственные, временные и энергетические характеристики робототизированных систем, объемы сохраняемых данных и др. В качестве критериев эффективности можно выбирать потребляемые физические ресурсы системы, а в качестве ограничений – вероятностно-временные характеристики производительности и доступа, размеры производственных помещений, возможности силовых агрегатов и т.п.

Предложенный подход развивает традиционные схемы оптимизации производительности информационных систем за счет выбора числа обслуживающих устройств, их быстродействия и надежности без учета объемов потребляемых при этом физических ресурсов и применим к широкому кругу базовых информационных процессов и технологий.

Заключение

Выявленный тренд развития базовых информационных технологий показывает, что учет затрат на обеспечение информационных технологий физическими ресурсами становится существенным при проектировании мощных информационных систем, таких как, например, системы хранения данных, центры обработки данных или суперкомпьютеры. Их эффективность зависит не только от возможностей информационных технологий в части объемов хранения, скорости передачи и обработки данных, но и от объемов занимаемого пространства, временных параметров физических процессов и потребляемой энергии. Исходя из этого, при проектировании подобных информационных систем следует:

- рассматривать весь спектр доступных информационных технологий сохранения, распространения и обработки;
- учитывать в процессе выбора физические параметры технологий;
- согласовывать физические параметры технологий с допустимыми параметрами задач пользователя;
- учитывать взаимную зависимость пространственных, временных и энергетических характеристик технологий.

Литература

1. Ляпунов А.А**.** Проблемы теоретической и прикладной кибернетики. М.: Наука, 1980. 335 с.
2. Моисеев Н.Н. Человек и ноосфера. М.: Молодая гвардия, 1990. 351 с.
3. Landauer R. Information is physical // Physics Today. 1991. V. 44. N 5. P. 23–29.
4. Винер Н. Кибернетика, или управление и связь в животном и машине. М.: Советское радио, 1958. 215 с.
5. Moore G.E. Cramming more components onto integrated circuits // Proceedings of the IEEE. 1998. V. 86. N 1. P. 82–85.

6. Kish L.B. Moore's law and the energy requirement of computing versus performance // IEE Proceedings: Circuits, Devices and Systems. 2004. V. 151. N 2. P. 190–194.
7. Landauer R. Irreversibility and heat generation in the computing process // IBM Journal of Research and Development. 2000. V. 44. N 1. P. 261–269.
8. Советов Б.Я., Колбанёв М.О., Татарникова Т.М. Технологии инфокоммуникации и их роль в обеспечении информационной безопасности // Геополитика и безопасность. 2014. № 1 (25). С. 69–77.
9. Bean J., Dunlap K. Energy-efficient data centers: a close-coupled row solution // ASHRAE Journal. 2008. V. 50. N 10. P. 34-36+38+40-42.
10. Schmidt R., Beaty D., Dietrich J. Increasing energy efficiency in data centers // ASHRAE Journal. 2007. V. 49. N 12. P. 18–21+24.
11. Gea-Banacloche J., Kish L.B. Future directions in electronic computing and information processing // Proceedings of the IEEE. 2005. V. 93. N 10. P. 1858–1863.
12. Советов Б.Я., Колбанёв М.О., Татарникова Т.М. Модель физических характеристик сигналов // Материалы VIII Санкт-Петербургской межрегиональной конференции «Информационная безопасность регионов России (ИБРР-2013)». Санкт-Петербург, 2013. С. 64–65.
13. Kish L.B., Granqvist C.G. Does information have mass? // Proceedings of the IEEE. 2013. V. 101. N 9. P. 1895–1899.
14. Belady C.L., Beaty D. Roadmap for datacom cooling // ASHRAE Journal. 2005. V. 47. N 12. P. 52–55.
15. Тысячелетний накопитель. Новейшие разработки в области хранения информации // Chip. 2012. № 6. С. 114–119.
16. Колбанёв М.О., Татарникова Т.М., Воробьёв А.И. Модель обработки клиентских запросов // Телекоммуникации. 2013. № 9. С. 42–47.
17. Tatarnikova T., Kolbanev M. Statement of a task corporate information networks interface centers structural synthesis // IEEE EUROCON 2009. 2009. Art. 5167903. P. 1883–1887.
18. Советов Б.Я., Колбанёв М.О., Татарникова Т.М. Оценка вероятности эрланговского старения информации // Информационно-управляющие системы. 2013. № 6 (67). С. 25–28.
19. Богатырев В.А., Богатырев А.В. Функциональная надежность систем реального времени // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 4 (86). С. 150–151.
20. Лёвшин И. Многоточие Стерлинга // Суперкомпьютеры. 2010. № 3 (15). С. 6–8.

Колбанёв Михаил Олегович – доктор технических наук, профессор, профессор, Санкт-Петербургский государственный экономический университет, Санкт-Петербург, 191023, Российская Федерация, mokolbanez@mail.ru

Татарникова Татьяна Михайловна – доктор технических наук, доцент, профессор, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация, tm-tatarn@yandex.ru

Mikhail O. Kolbanev – D.Sc., Professor, Professor, Saint Petersburg State University of Economics, Saint Petersburg, 191023, Russian Federation, mokolbanez@mail.ru

Tatiana M. Tatarnikova – D.Sc., Associate professor, Professor, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation, tm-tatarn@yandex.ru

Принято к печати 11.05.14
Accepted 11.05.14

УДК 621. 391. 037. 372

ПОИСК ЛЮДЕЙ ПО ФОТОРОБОТАМ: СОСТОЯНИЕ ПРОБЛЕМЫ И ТЕХНОЛОГИИ

Г.А. Кухарев[a], Ю.Н. Матвеев[b,c], Н.Л. Щеголева[d]

[a]Западно-Поморский технологический университет в Щецине, Щецин, 70-310, Польша
[b] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[c] ООО «Центр речевых технологий», Санкт-Петербург, 196084, Российская Федерация, matveev@mail.ifmo.ru, matveev@speechpro.com
[d] Санкт-Петербургский государственный электротехнический университет (ЛЭТИ), Санкт-Петербург, 197376, Российская Федерация

Аннотация. Обсуждается проблема поиска людей по фотороботам, составленным по словесному портрету. Приводится обзор состояния этой проблемы от исходных понятий и используемой терминологии до современных технологий создания фотороботов, реальных сценариев и результатов поиска. Представлены история развития систем формирования композиционных портретов (фотороботов и скетчей) и идеи, реализованные в этих системах. Обсуждается задача автоматического сравнения фотороботов с оригинальными фотографиями, вскрываются причины недостижимости устойчивого поиска фотопортретов-оригиналов по фотороботам в реальных сценариях.
Формулируются требования к базам фотороботов в дополнение к существующим бенчмарковым базам изображений лиц, а также способы реализации таких баз. В рамках этих способов обсуждаются методы генерации популяции фотороботов из исходного фоторобота для повышения результативности поиска по нему фотопортрета-оригинала. Представлен метод повышения индекса подобия в паре фоторобот/фотопортрет, основанный на вычислении среднего фоторобота из сформированной популяции. Показано, что такие фотороботы более подобны портретам-оригиналам и их использование в обсуждаемой проблеме поиска может привести к высоким результатам. При этом сформированные фотороботы отвечают требованиям правдивого сценария, поскольку учитывают возможность неполной информации в словесных портретах. Обсуждаются результаты применения этих методов для баз CUHK Face Sketch database и CUHK Face Sketch FERET database, а также опубликованных в открытой печати фотороботов и соответствующих им фотографий.
Ключевые слова: изображения лиц, фоторобот, скетч.
Благодарности. Данное исследование проводится при частичной финансовой поддержке Правительства Российской Федерации (грант № 074-U01).

PEOPLE RETRIEVAL BY MEANS OF COMPOSITE PICTURES: PROBLEM STATE-OF-THE-ART AND TECHNOLOGIES

G.A. Kukharev[a], Yu.N. Matveev[b, c], N.L. Shchegoleva[d]

[a]West Pomeranian University of Technology, Szczecin, 70-310, Poland
[b] ITMO University, Saint Petersburg, 197101, Russian Federation
[c] STC Ltd., Saint Petersburg, 196084, Russian Federation, matveev@mail.ifmo.ru, matveev@speechpro.com
[d] Saint Petersburg Electrotechnical University (LETI), Saint Petersburg, 197376, Russian Federation

Abstract. We discuss the problem of people retrieval by means of composite pictures constructed according to descriptive portrait. An overview of the problem state-of-the-art is provided beginning from the basic concepts and terminology to a modern technology for composite picture creation, real-world scenarios and search results. The development history of systems for forming composite portraits (photo robots and sketches) and the ideas implemented in these systems are provided. The problem of automatic comparison of composite pictures with the original ones is discussed, and the reasons for unattainability of stable retrieval of originals by a composite picture in real-world scenarios are revealed.
Requirements to composite pictures databases in addition to the existing benchmark databases of facial images and also methods for implementation of such databases are formulated. Approaches for generation of sketches population from an initial one that increase effectiveness of identikit-based photo image retrieval systems are proposed. The method of similarity index increasing in the couple identikit-photograph based on computation of an average identikit from the created population is provided. It is shown that such composite pictures are more similar to original portraits and their use in the discussed search problem can lead to good results. Thus the created identikits meet the requirements of the truthful scenario as take into account the possibility of incomplete information in descriptions. Results of experiments on CUHK Face Sketch and CUHK Face Sketch FERET databases and also open access identikits and corresponding photos are discussed.
Keywords: face images, composite picture, sketch.
Acknowledgements. The work is partially financially supported by the Government of the Russian Federation (grant 074-U01).

Введение в проблему синтеза фотороботов

В течение почти двадцати лет (начиная с работ [1, 2]) не утихает интерес к проблеме автоматического сравнения субъективного портрета, составленного по показаниям свидетелей некоторого криминального события, и оригинальными фотопортретами подозреваемых. Исходной при этом является информация, содержащаяся в показаниях свидетелей, и записанная с их слов (что определяется как словесный портрет подозреваемого человека). По форме исполнения субъективный портрет может быть представлен как рисованный портрет и как композиционный портрет. Рисованный портрет – это штриховой или полутоновый рисунок всей области лица, выполненный художником или криминалистом по словес-

ному портрету [3]. Композиционный портрет представляет собой изображение лица, составленное из его отдельных примитивов (например, бровей, глаз, носа, губ), а также сопутствующих элементов. Последние включают головные уборы, очки, сережки, банты, заколки и т.д. [3]. В свою очередь примитивы могут быть как рисованными (заранее подготовленными), так и состоять из фрагментов фотоизображений лиц. В этих случаях получается рисовано-композиционный или фотокомпозиционный портрет. Метод изготовления фотокомпозиционных портретов из фрагментов фотографий по словесному портрету был предложен в середине прошлого века французским криминалистом П. Шабо [3–6]. Эти портреты П. Шабо назвал фотороботами, и с тех пор все субъективные портреты, независимо от техники их создания и формы представления, стали называть фотороботами.

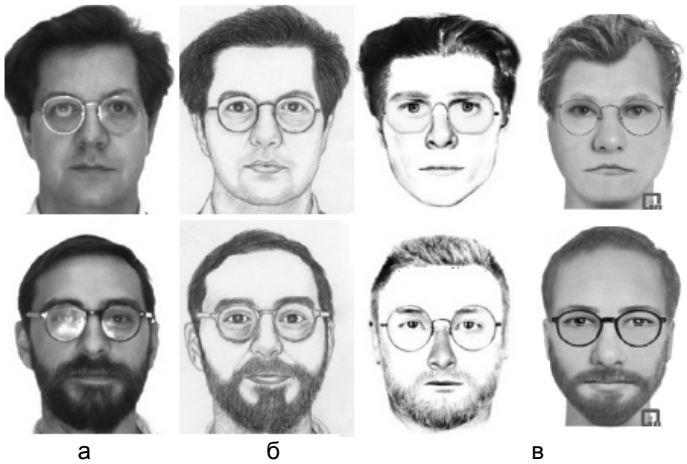
В зарубежной научно-технической литературе используется другая терминология для фотороботов, основу которой составляет слово скетч (Sketch [7–11]), что в переводе с английского означает «эскиз или набросок». При этом используются следующие основные формы таких фотороботов: Viewed Sketch; Artist Sketch; Composite Sketch; Composite Forensic Sketch.

Viewed Sketch – рисунок, выполненный художником по фотографии или непосредственно по лицу человека, которого видит художник. Часто под формой Viewed Sketch понимают также компьютерный рисунок, автоматически полученный из исходного цифрового изображения лица. Примеры Viewed Sketch-рисунков, выполненных художником по фотографии, показаны на рис. 1 в колонке «б». Дополнительно доработанный художником компьютерный рисунок (Viewed Sketch) определяется как Artist Sketch.

Forensic Sketch – рисунок, выполненный художником-криминалистом по словесному портрету со слов свидетеля. Если для составления скетча используется библиотека примитивов лица, то результат определяется как «Composite Sketch». Примеры Composite Sketch показаны на рис. 1 в колонке «в». Если при этом Composite Sketch составлен криминалистом по словесному портрету, то он определяется как Composite Forensic Sketch.

В настоящее время все фотороботы выполняются с помощью специальных компьютерных программ, позволяющих составлять композиционные портреты по примитивам, выбираемым из встроенной библиотеки этих программ. Первыми и наиболее простыми программами были «IdentiKit», «PhotoFit» и EFIT, более совершенными являются программы «Mac-a-Mug», «Портрет», «Облик» и «FACES», «IdentiKit 2000» [3–6, 8–12]. Основная идея, реализованная в этих программах, – механическая сборка (коллаж) оператором программы области лица из отдельных примитивов, содержащихся в базе. При этом выбранные в общий фоторобот примитивы могут принадлежать формально разным по генотипу людям. Однако поскольку склеивание примитивов и их подгонка по границам и текстуре области лица выполняется методами компьютерной графики, то это создает эффект хорошо выполненного фотопортрета.

Несмотря на представительные базы примитивов лиц и совершенную технику их склеивания в компьютерных программах в общий портрет, а также развитый интерфейс этих программ, качество получаемых фотороботов существенно зависит от опыта специалиста, обслуживающего саму программу, и субъективизма человека, составившего словесный портрет. Например, на рис. 1 видно, что исходное фото и Viewed Sketch практически точно совпадают между собой, что определено способом получения Viewed Sketch. Однако Composite Sketches (фотороботы) менее подобны исходному фото (хотя выполнены именно по нему), а также различаются и между собой, что обусловлено разными характеристиками использованных программ синтеза фотороботов.



а б в

Рис. 1. Исходные изображения (фотопортреты) (а), Viewed Sketches (б) и два варианта Composite Sketches [9] (в)

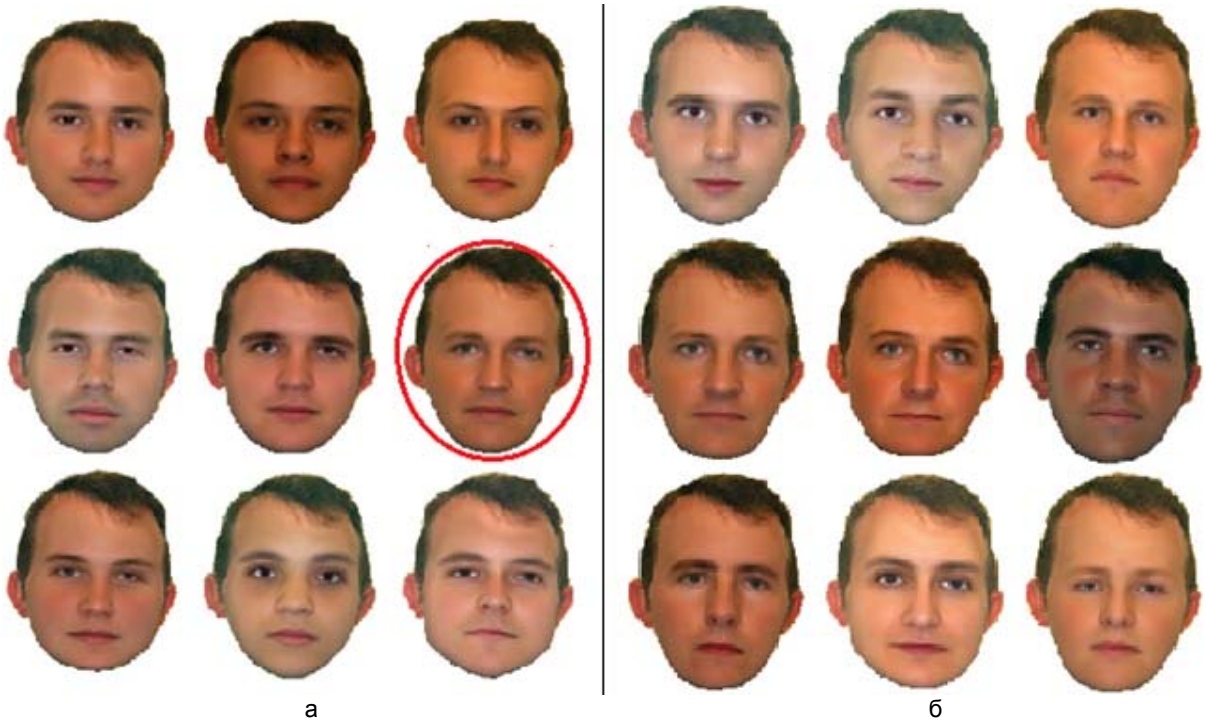
Можно также отметить, что этот факт «неподобия» наблюдается практически при любых способах механической сборки фотороботов по отдельным примитивам. И мера этого неподобия растет, если фо-

тороботы формируются не по исходному изображению, а непосредственно по показаниям свидетелей или составленных ими словесных портретов. Ситуация эта еще более усугубляется, если исходный словесный портрет составляется свидетелем по памяти через несколько дней после контакта с подозреваемым (или преступником), когда в памяти свидетеля сохранился уже не полный состав исходных примитивов лица подозреваемого.

Эти факты «неподобия» вызвали новый интерес к созданию более совершенных методов и компьютерных систем построения фотороботов, в результате чего появились системы композиции лиц на основе эволюционных алгоритмов (ЭА) и интерактивных стратегий. В этом случае исходные лица для фотороботов не составляются по отдельным примитивам, а выбираются из базы систем с учетом фенотипа (или особенностей лица, вытекающих из словесного портрета) и рассматриваются как единое целое, а все изменения лиц осуществляются в рамках ЭА и корректируются свидетелем в интерактивном режиме. Наиболее простыми исходными данными по фенотипу могут быть, например, форма головы и (или) особенности отдельных примитивов лица, а также, например, национальные особенности лиц, пол и возраст человека, представленные в словесном портрете.

Первой из систем на основе ЭА и интерактивных стратегий была представлена система К. Соломона «EFITV-V – **E**igen **FIT** version **V**», а затем и система Чарльза Фровда «EvoFIT – **Evo**lutionary **F**acial **I**maging **T**echnique for Creating Composites» [13, 14]. Обе системы используют представление изображений лиц на основе модели формы (Active Shape Model – ASM) и модели внешнего вида (Active Apperance Model – AAM). При этом ASM определяет контур всего лица, а AAM – его текстуру. Параметры этих моделей представлены 50-ю признаками в собственном подпространстве, основанном на методах анализа главных компонент (Principal Component Analysis – PCA) и преобразовании Карунена–Лоэва (KLT – Karhunen-Loeve Transform). И именно эти параметры моделей изменяются в рамках ЭА на основе процедур клонирования и случайной мутации 50-ти исходных признаков.

В этих системах уже на первой пробе синтеза фоторобота генерируется «популяция» из нескольких лиц, отвечающих как индивидуальности искомого лица (фенотипу), так и случайно измененным его параметрам. Пример такой исходной популяции для системы «EFITV-V» из работы [15] показан на рис. 2, а. Далее свидетелем в интерактивном режиме выбирается тот результат из исходной популяции изображений лиц, который наиболее точно соответствует описанию исходного словесного портрета или отдельным примитивам словесного портрета (например, на рис. 2, а, выбранный результат отмечен кругом). Опираясь на этот результат как на текущую модель фоторобота, система генерирует новую популяцию фотороботов по измененным на основе ЭА параметрам текущей модели. Полученный при этом результат – новая популяция лиц – показан на рис. 2, б (пример также взят из работы [15]).



а б

Рис. 2. Примеры популяции из 9 лиц в системе «EFIT-V» [13]: популяция нескольких изображений лиц, отвечающих как индивидуальности искомого лица (фенотипу), так и случайно измененным его параметрам, где выбранная свидетелем модель фоторобота отмечена кругом (а); новая популяция изображений лиц, сгенерированная на основе выбранной свидетелем модели фоторобота (б)

Если сравнить обе популяции изображений лиц, представленных на рис. 2, то можно отметить два факта. Факт первый: популяция на рис. 2, а, представляет разные лица, имеющие подобие только по форме волос на голове, но заметные различия по форме головы и базовым примитивам лица. Факт второй: популяция рис. 2, б, действительно представляет генотип лица одного и того же человека с небольшими вариациями его основных особенностей (примитивов). Наконец, теперь уже из этой популяции (рис. 2, б) выбирается то лицо, которое более точно соответствует описаниям исходного словесного портрета или отдельным (ранее не учтенным) его примитивам. Процесс продолжается далее, до полного утверждения (или согласования) свидетелем полученного результата.

Подход, реализованный в системах «EFIT-V» и «EvoFIT», основан на идеях ЭА и интерфейса человек–компьютер, поскольку промежуточные решения ЭА корректируются самим автором исходного словесного портрета. Окончательное решение также определяется автором, что приводит к быстрому и более точному получению окончательного фоторобота. Именно поэтому эти идеи были определены их авторами как стратегия интерактивной эволюции в создании фотореалистичных композиционных лиц (фотороботов). В ряде выполненных экспериментов было показано, что получаемые в рамках данной стратегии фотороботы известных людей настолько близки своим оригиналам, что относительно легко были идентифицированы экспертами (психологами и криминалистами) [15, 16].

Подводя итог настоящему разделу, отметим, что эффект высокого уровня подобия между фотопортретом-оригиналом и синтезированным фотороботом определяется следующими характеристиками стратегии интерактивной эволюции:

1. отказом от механизма «сборки фотороботов» по отдельным примитивам;
2. выбором в качестве исходного фоторобота целостного изображения лица с учетом данных по фенотипу;
3. использованием параметрических моделей изображения лица в собственном пространстве признаков;
4. изменением параметров модели фоторобота на основе ЭА;
5. выбором наилучших решений в рамках интерактивного общения со свидетелем.

Проблемы автоматического сравнения фотороботов с оригинальными фотографиями

Возвращаясь к исходной проблеме, отметим, что построение фотороботов является только первым шагом в решении более общей задачи – задачи автоматического (т.е. без участия человека) поиска соответствий между фотороботом и оригинальным изображением лица. Эта задача появляется, например, при поиске оригинального изображения лица в большой базе данных по заданному фотороботу или при поиске соответствий между фотороботами и лицами людей в системах видеонаблюдений, а также при решении задач взаимного распознавания оригинального фото и скетча.

Как было показано выше, интерактив со свидетелем является чрезвычайно важным (можно сказать, даже принципиальным) в системах «EFIT-V» и «EvoFIT», поскольку только свидетель принимает окончательное решение о подобии фоторобота субъективному портрету. Однако даже фотороботы, построенные в рамках стратегии интерактивной эволюции, не решают проблемы автоматического сравнения фотороботов с оригинальными фотографиями.

Как показали результаты поиска, представленные в работах [10, 11], стабильное распознавание Composite Sketches и Composite Forensic Scetches и устойчивый поиск соответствующих им фотопортретов в специальных криминалистических базах являются на сегодня практически недостижимыми.

Этот печальный факт является следствием трех основных причин:

1. низким качеством скетчей, отображающих реальный словесный портрет;
2. несовершенством методов взаимного распознавания пары скетч-фотопортрет;
3. отсутствием необходимых для этих случаев баз фотопортретов и соответствующих им фотороботов.

Первая причина определяется технологией перевода имеющегося словесного портрета в соответствующее изображение. И здесь мы сталкиваемся с субъективизмом свидетелей, как правило, случайных людей, не связанных с криминалистикой, и особенностями их внимания и описания подозреваемого, сложенным с участием свидетеля в процессе перевода словесного портрета в скетч и высокой ответственностью свидетеля за ошибки. Очевидно, что лучшее решение здесь можно получить при использовании систем «EFIT-V» и «EvoFIT», хотя это и не всегда возможно на практике. В этом случае либо отсутствует возможность использования подобного класса систем, либо отсутствует возможность интерактивного общения со свидетелем, либо, наконец, интерактивное общение проходит так поздно (через несколько дней), что свидетель утрачивает (забывает) часть важной информации о подозреваемом.

Вторая причина связана с неразвитостью методологии (идей, методов, подходов и практических решений) сравнения изображений скетч-фотопортрет, если скетчи выходят за рамки классов Viewed Sketch и Artist Sketch, а также отсутствием опыта такого сравнения ввиду недоступности информации по объектам сравнения (например, личностей криминального характера).

Третья причина является следствием неприспособленности «старых баз фотопортретов» преступников к современным компьютерным технологиям обработки изображений в лицевой биометрии. Кроме того, практически отсутствуют доступные специальные оценочные (бенчмарковые) базы данных, в которых каждый класс содержит оригинальный фотопортрет и различные варианты (или вариации) скетчей для него. Следует отметить, что из перечисленных причин третья является основополагающей: именно она не позволяет развивать методологию сравнения изображений скетч–фотопортрет, поскольку отсутствует объект исследований, что ограничивает или сдерживает выполнение задач сравнения на практике.

Именно поэтому исследователи обратились к задачам создания баз скетчей в дополнение к известным бенчмарковым базам изображений лиц [17, 18], к разработке методов сравнения скетчей с соответствующими фотопортретами и к моделированию задач поиска фотопортретов по заданным скетчам [6–9, 17–21]. В результате этого появились первые базы скетчей, наиболее популярные среди которых – **CU**HK **F**ace **S**ketch database (CUFS) и **CU**HK **F**ace **S**ketch **F**ERET database (CUFSF), содержащие фото и полученные по ним скетчи классов Viewed Sketch и Artist Sketch [15, 16]. Кроме того, в работах [19–23] предложены новые идеи автоматического построения скетчей из исходных фотопортретов людей и методы их сравнения. При этом исследования выполнены на базах CUHK и CUFSF, а результатом являлись также скетчи класса Viewed Sketch. Если же говорить о методах сравнения фото и скетчей, то можно отметить следующее: использование достаточно сложных методов обработки в приложении к скетчам из баз CUHK и CUFSF не является оправданным. Возможно, большую роль здесь сыграла «мода» на перечисленные ниже методы, а не обоснованность их выбора. Среди них: скрытые марковские модели (HMM – Hidden Markov Models), методы на основе CITE (coupled information-theoretic encoding) и CITP (coupled information-theoretic projection), методы SIFT (Shift Invariant Feature Transform) и LBP (Local Binary Pattern), гистограммы BPH (Binary Pattern Histogram) и др., использованные в работах [8, 9, 17–23]. Заметим также, что при представлении результатов взаимного распознавания фото↔скетч в рамках баз CUFS и CUFSF в этих работах недостаточно ясно изложены не только методы обработки, но и параметры обучающих и тестовых выборок изображений, что не позволяет однозначно представить модель выполненных экспериментов. А это существенно затрудняет оценку представленных результатов, не позволяет проверить реализованную модель выполненных экспериментов и использовать полученные результаты в рамках мета-анализа.

Состояние дел по этим проблемам (подходы, решения, результаты и их анализ) было представлено в работах [24, 25] и в настоящей работе дальше не анализируется.

Какие базы скетчей нужны сегодня?

База CUHK содержит скетчи, сгенерированные автоматически из исходных фото и дорисованные художниками, и включает 188 пар фото-скетч. База CUFSF содержит скетчи, передающие основные особенности портретов людей из базы FERET и особенности их лиц, но с элементами изменений (небольшого преувеличения с элементами карикатуризма), внесенными художником. Обе эти базы фактически составлены из скетчей, которые были определены выше как Artist Sketch.

Различие способов получения скетчей базы CUFS и базы CUFSF привело к тому, что первые из них распознаются простыми методами [24, 25] с большей эффективностью (почти до 100%), в то время как вторые также распознаются простыми методами, но только при очень точном согласовании размеров и ориентации области лица в плоскости *XY*. К сожалению, на сайте [18] доступны только выделенные изображения (cropped sketches) скетчей крайне низкого качества (по разрешению, размерам и текстуре), что практически не позволяет проводить с ними репрезентативные исследования.

Пример согласования выделенных областей лиц базы FERET и соответствующих им скетчей базы CUFSF [18] приведен на рис. 3.

В сравнении с выделенными скетчами (cropped sketches) из базы [18], на скетчах рис. 3, б, можно увидеть дополнительную ретушь областей лиц и элементов на ней (лба, носа, рта…). И, как видно, эти операции выполнены «вручную», т.е. не в автоматическом режиме обработки изображений. Кроме того, текстура на области лиц «выглажена» низкочастотной фильтрацией. И, наконец, выполнено согласование основных антропометрических параметров (линии глаз, расстояния между центрами глаз и т.д.).

Возможно, что для введения в задачи исследования методов сравнения скетчей и фото и их взаимного распознавания такое выравнивание необходимо. Однако, если исходить из реальных задач и, например, криминальных событий и сценариев (где актуальной становится задача поиска подозреваемого по заданному скетчу), то «условие выравнивания лиц по антропометрике» практически недостижимо. И это связано с тем, что заранее неизвестно, как выглядит оригинальное фото подозреваемого, а, следовательно, неизвестны и параметры лица на фото, и в какой степени будет оно соответствовать имеющемуся словесному портрету (и сгенерированному по нему скетчу).

Обратимся к рис. 4, где показаны пары фото/фоторобот (Composite Forensic Sketches), и отметим, что между фото и фотороботом нет такого высокого подобия, которое наблюдается, например, в соответствующих парах на рис. 1 или рис. 3.

а

б

Рис. 3. Примеры согласования областей лиц на фото (а) и скетчах [18] (б)

На рис. 4 на всех четырех парах фото/скетч заметно различие размеров (высоты и ширины) областей лиц, размеров и положения примитивов на лицах, а также нарушение симметрии областей лиц и различие в ориентации взгляда. Естественно, что сравнение таких скетчей с фотооригиналами, ориентированными, как правило, анфас и нормализованными по стандарту, будет весьма затруднительным или практически невозможным.

При этом возникает новый вопрос: как же тогда сравнивать между собой скетчи и фотооригиналы? И здесь напрашивается аналогия с теми стратегиями, которые реализованы в системах «EFIT-V» и «EvoFIT». Ответ получаем такой: каждый исходный скетч должен быть несколько раз модифицирован и представлен с новыми параметрами по геометрии областей лиц (размеру, симметрии, сдвигу) с целью создания «новой популяции» таких скетчей. Данная модификация как бы сымитировала получение $K>1$ скетчей от «группы из K свидетелей». И, уже в рамках этой популяции, можно решить задачу сравнения скетчей с фотооригиналом. Собственно само сравнение может выполняться со средним (для всей популяции) скетчем или с каждым скетчем из полученной популяции на основе мажоритарных механизмов или, например, на основе смеси экспертов (Mixtures of experts). И только в таком случае можно надеяться на хороший результат сравнения!

а

б

Рис. 4. Фотографии (а) и соответствующие им скетчи, составленные по описаниям свидетелей (данные с сайта http://prikol.i.ua/view/474038) (б)

Именно эти факты послужили отправной точкой для предложенных в работах [22, 23] подходов по генерации нового состава (популяции) скетчей из любого заданного исходного скетча. Причем механизм получения этих своеобразных популяций скетчей реализуется достаточно просто, что показано и о чем свидетельствуют результаты экспериментов на примере баз CUFS и CUFSF, представленные в работах

[24, 25]. Справедливости ради отметим, что подобные же результаты можно было бы получить и на основе подходов для генерации «популяции карикатур», изложенных в работе [26]. Однако решения в [26] основаны на точных моделях форм (ASM) и моделях внешнего вида (AAM) изображений лиц, линейных преобразованиях между двумя изображениями (основанных на PCA и KLT) и различных вариантах перекомпоновки и «преувеличения» компактных областей лиц или примитивов лиц. С учетом этого отметим, что подход [26] проигрывает подходам [24, 25] по сложности алгоритмов генерации требуемых популяций лиц. А высокая точность моделей внешнего вида, достигаемая в подходе [26], совершенно не обязательна в задаче генерации популяции скетчей, поскольку, по определению, мы не знаем того фотопортрета, к которому принадлежит заданный нам скетч.

В следующем разделе, в отличие и в дополнение к [24, 25], приводится полное описание процесса генерации «популяции скетчей» и исследуются их характеристики на примере Composite Forensic Sketches. Хотя вместо скетчей можно генерировать и популяции новых фотопортретов из некоторого заданного фотопортрета (например, из портрета mug-shot, часто используемого в криминалистике [10, 11]).

Алгоритм генерации «популяции скетчей»

Блок-схема алгоритма генерации «популяции скетчей» приведена на рис. 5. По этому алгоритму из исходного скетча (поданного на вход 1) формируются $K>2$ новых скетчей с небольшими геометрическими изменениями области лица. На вход 2 подаются параметры изменений скетчей.

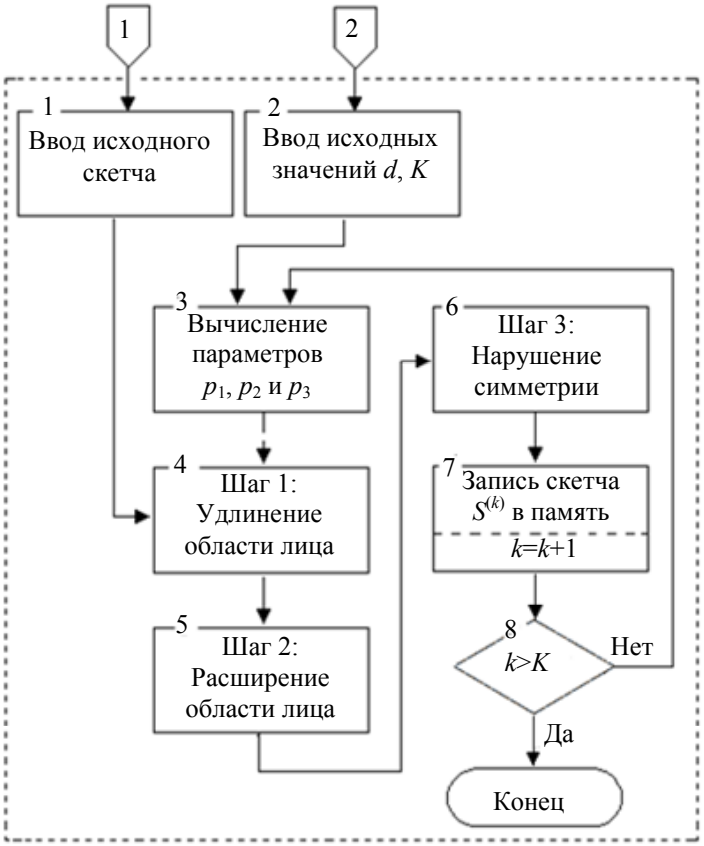


Рис. 5. Алгоритм генерации популяции скетчей, где 1 и 2 – входы для данных и параметров

Пусть задана матрица **S** размером $M \times N$, представляющая исходное изображение скетча в формате GRAY. При этом примем, что область лица на скетче занимает не менее 80% всего исходного изображения, что определено, например, биометрическими стандартами.

Для каждого значения $k = 1, 2, ..., K$ сформируем три параметра p_1, p_2 и p_3 с использованием генератора случайных чисел и отмасштабируем их значения в диапазоне $\pm d$ так, что

$$p_i = sign(R_n^{(i)})\, fix(dR_u^{(i)}), \text{ для } i = 1, 2, 3, \quad (1)$$

где p_i – параметр; d – выбранное значение изменения параметров, причем $1 < d \le \sigma$; σ – максимальное значение границ изменения параметров; $R_n^{(i)}$ и $R_u^{(i)}$ – нормально и равномерно распределенные случайные числа; $sign(R_n^{(i)})$ – знак числа.

При этом параметры p_1 и p_2 связаны с изменением длины и ширины лица на исходном изображении, что приводит к изменению положения линии глаз и расстоянию между центрами глаз. Параметр p_3

связан с изменением положения линии симметрии на области лица. А параметр σ может быть, например, связан с числом пикселей, на которые изменяется (увеличивается или уменьшается) центральная область лица, положение линии глаз и расстояния между центрами глаз или оси симметрии лица. Далее алгоритм геометрических изменений области лица складывается из трех шагов, на каждом из которых выполняется одна операция изменения геометрии области лица.

Шаг 1. Если $p_1 > 0$, то удаляем первые (p_1–1) строк из матрицы **S**. Если $p_1 < 0$, то удлиняем матрицу **S**, дописывая ее первые (p_1–1) строк сверху матрицы **S**. Указанные действия можно записать следующим образом:

$$\mathbf{S}^{(1)}(var \times N) = \begin{cases} S(p_1 : M, :), & \text{if } p_1 > 0 \\ [S(1 : abs(p_1); \quad S], & \text{if } p_1 < 0 \end{cases}, \tag{2}$$

где $\mathbf{S}^{(1)}(var \times N)$ – матрица с уменьшенным или увеличенным числом строк, что определено параметром «*var*».

Далее выполняем перемасштабирование матрицы $\mathbf{S}^{(1)}$ до ее первоначального размера $M \times N$ так, что

$$\mathbf{S}^{(1)}(var \times N) \rightarrow S^{(1)}(M \times N). \tag{3}$$

При этом длина лица в матрице-результате $\mathbf{S}^{(1)}$ в (3) увеличивается, если выполнялось условие $p_1 > 0$, или уменьшается, если выполнялось условие $p_1 < 0$. При этом линия глаз и линия рта смещаются вверх или вниз. Заметим, что интересующее нас изменение области лица скетча будет находиться в пределах текущих значений *d* от исходной длины лица.

Шаг 2. Далее, если $p_2 > 0$, то удаляем первые (p_2–1) столбцы из матрицы $\mathbf{S}^{(1)}$. Если $p_2 < 0$, то удаляем последние (p_2–1) столбцы матрицы $\mathbf{S}^{(1)}$. Указанные действия можно записать следующим образом:

$$\mathbf{S}^{(2)}(M \times var) = \begin{cases} S^{(1)}(:, p_2 : N), & \text{if } p_2 > 0 \\ S^{(1)}(:, 1 : N - abs(p_2)), & \text{if } p_2 < 0 \end{cases}. \tag{4}$$

При этом матрица-результат $\mathbf{S}^{(2)}$ в формуле (4) становится на $abs(p_2–1)$ столбцов меньше при любом значении p_2.

Далее опять выполним перемасштабирование матрицы $\mathbf{S}^{(2)}$ до первоначального размера $M \times N$:

$$\mathbf{S}^{(2)}(M \times var) \rightarrow \mathbf{S}^{(2)}(M \times N), \tag{5}$$

что неминуемо приведет к расширению области лица в поле изображения и к асимметрии области лица. При этом без учета изменения длины области лица в (3) расширение области лица по (5) определится величиной, близкой к значению *d*.

Шаг 3. Выполним циклический сдвиг матрицы $\mathbf{S}^{(2)}$ влево на (p_3–1) столбцов, если $p_3 > 0$, или вправо, если $p_3 < 0$. Этот шаг запишем следующим образом:

$$\mathbf{S}^{(3)}(M \times N) = \begin{cases} [\mathbf{S}^{(2)}(:, p_3 + 1 : N) \quad \mathbf{S}^{(2)}(:, 1 : p_3)], & \text{if } p_3 > 0; \\ [\mathbf{S}^{(2)}(:, 1 : N - abs(p_3) + 1 : N) \quad \mathbf{S}^{(2)}(:, 1 : N - abs(p_3))], & \text{if } p_3 < 0, \end{cases} \tag{6}$$

что приводит к циклическому сдвигу всего изображения скетча, нарушающему его симметрию относительно центральной линии симметрии лица.

Теперь перепишем формулу (6) в новой форме:

$$\mathbf{S}^{(k)} = \mathbf{S}^{(3)}, \tag{7}$$

где матрица-результат (7) представляет собой новый скетч.

Запишем результат $\mathbf{S}^{(k)}$ в память и перейдем снова к формированию параметров p_1, p_2 и p_3 и далее к шагам 1–3, вплоть до формирования нового скетча $\mathbf{S}^{(k+1)}$ в соответствии с (1)–(7) и т.д.

Теперь ответим на следующие вопросы: как и чем оценить меру подобия между исходным фото и соответствующим ему скетчем? Должно ли их внешнее сходство (видимое человеку) соответствовать некоторому формальному показателю? И что нам важнее – видимое человеком внешнее сходство фото и скетча или формальная оценка этого сходства? Эти вопросы особенно актуальны в связках «свидетель→словесный портрет→скетч» и «скетч→найденное по нему фото». Некоторые ответы на эти вопросы можно найти в работах [24–28]. Ниже дополним эти ответы новым анализом результатов и обзором по ним.

Метод повышения индекса подобия в паре скетч-фото

В работах [24, 25] было показано, что, если скетчи получаются непосредственно из исходных фотопортретов (Viewed Sketch или Artist Sketch), то они распознаются простыми системами со 100% результатом. Однако, как только эти скетчи пройдут этап локальных изменений области лица, результативность их распознавания существенно снижается. Отмеченные факты вполне объяснимы, поскольку мы изменили локальную структуру скетчей, добиваясь приближения их к ситуации, отвечающей правдивому сценарию. Подобные ситуации возникают при композиции скетчей по словесному портрету или при автоматическом получении скетчей для неточно заданных (или неполных) параметров «исходных данных». Если

же рассуждать на уровне «субъективного и формального подобия», в рамках полученной популяции скетчей мы сохранили их субъективное внешнее сходство с исходным скетчем, но существенно уменьшили значение формального параметра подобия – индекса структурного подобия (Structural SIMilarity Index – SSIM [27, 28]). Индекс SSIM (ISSIM) позволяет оценить степень подобия (искажения) двух изображений как комбинацию трех факторов: яркостных изменений, изменений контраста и потери корреляции между изображениями. При этом, поскольку результативность распознавания скетчей линейно связана со значением ISSIM, нашей следующей задачей является проведение нового (дополнительного) этапа обработки, сохраняющего субъективное внешнее сходство скетчей, но существенно увеличивающее значение индекса ISSIM.

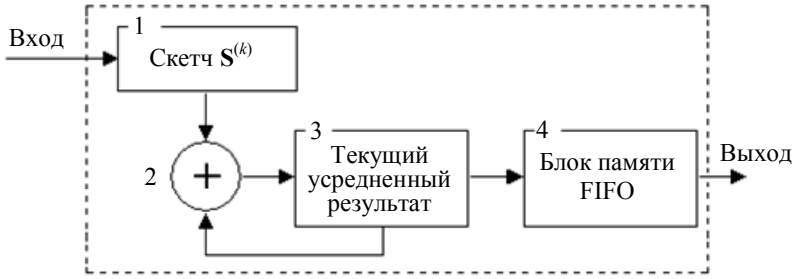


Рис. 6. Блок-схема алгоритма создания новой популяции скетчей: блок 1 – ввод исходного скетча; блок 2 – кумулятивное суммирование всех поступающих скетчей; блок 3 – вычисление текущего среднего скетча; блок 4 – запись результата в память

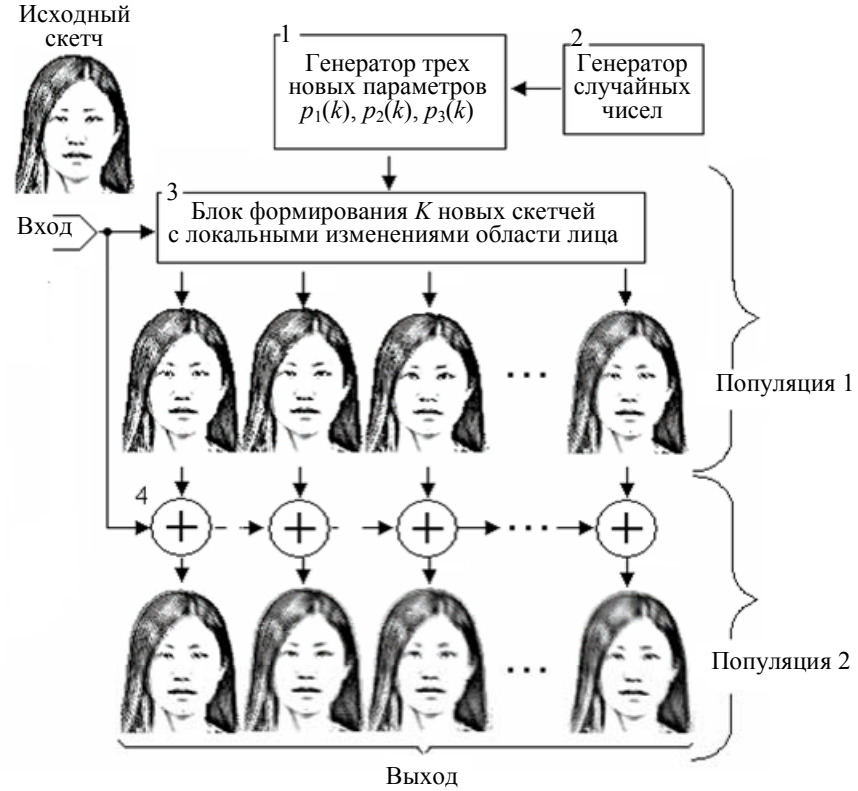


Рис. 7. Структура метода формирования популяции скетчей: блок 1 – генератор параметров $p_1(k)$, $p_2(k)$, $p_3(k)$; блок 2 – генератор случайных чисел; блок 3 – формирование K новых скетчей с локальными изменениями области лица; блок 4 – суммирование скетчей

Однако это не может быть операция, обратная той, что была описана выше, поскольку мы и дальше исходим из предположения «неполного знания параметров исходного фотопортрета» или «полного незнания этих параметров». Это будет этап создания новой популяции скетчей, предложенный в [24, 25], идеи которого представляет рис. 6. Здесь на вход 1 последовательно поступают K скетчей $\mathbf{S}^{(k)}$, представленных матрицами данных размером $M \times N$. На выходе также получаем K новых скетчей $\tilde{\mathbf{S}}^{(k)}$, что достигается кумулятивным суммированием всех поступающих скетчей и вычислением текущего среднего скетча $\tilde{\mathbf{S}}^{(k)}$:

$$\tilde{\mathbf{S}}^{(k)} = \left(\sum_{j=1}^{k} \mathbf{S}^{(j)}\right)/k, \text{ для } k = 1, 2, \ldots, K. \quad (8)$$

Операция кумулятивного суммирования (8) реализуется в блоках 2 и 3. Результат записывается в память (на промежуточном этапе это может быть память типа FIFO – блок 4), что показано на рис. 6. На рис. 7 показан фрагмент взаимодействия блоков 3 и 4, где под номером 4 показана линейка сумматоров.

Анализ результатов

Теперь покажем, что скетчи популяции 2 имеют более высокий индекс подобия с исходным изображением, чем скетчи, полученные в популяции 1.

Рассмотрим рис. 8, на котором показаны исходный фотопортрет (рис. 8, б) и соответствующие ему скетчи из [17]: Viewed Sketch (рис. 8, а) и Artist Sketch (рис. 8, в). На рис. 8, г, показаны значения индекса SSIM между исходным фотопортретом и Viewed Sketch из популяции 1 (нижняя кривая П1) и популяции 2 (верхняя кривая П2). На рис. 8, д, показаны значения индекса SSIM между исходным фотопортретом и Artist Sketch из популяции 1 (нижняя кривая П1) и популяции 2 (верхняя кривая П2). Кривые значений ISSIM приведены для 9 скетчей из этих популяций. Прямые горизонтальные линии отмечают пороги 0,62 и 0,495 как значения индекса подобия между фотопортретом (рис. 8, б) и исходными скетчами (рис. 8, а, в).

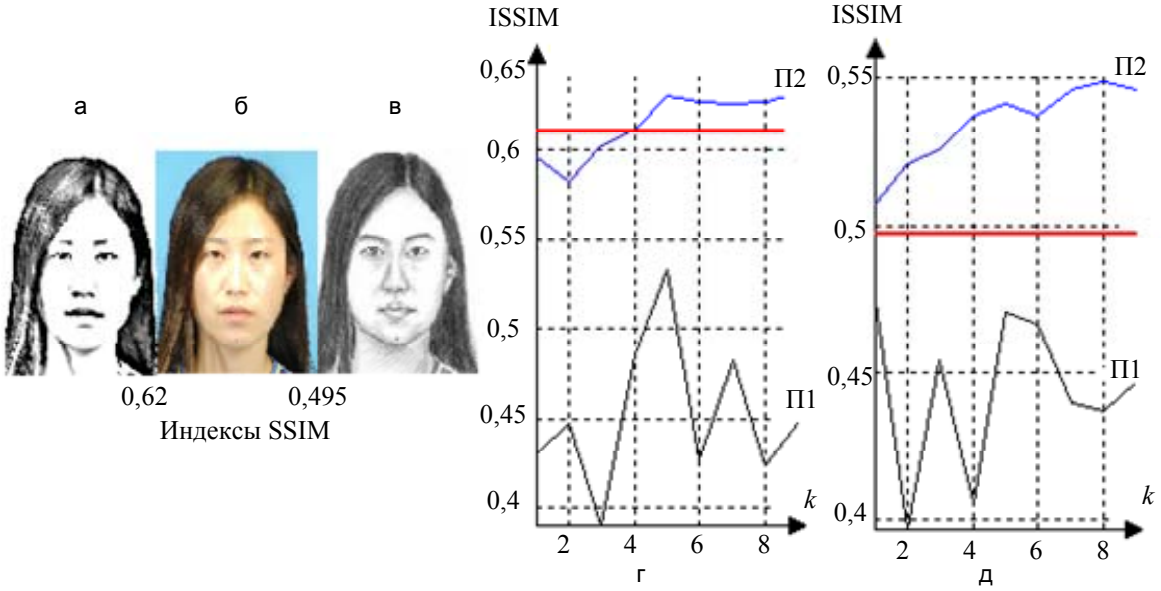


Рис. 8. Изменения индекса SSIM для исходных скетчей и сгенерированных популяций скетчей: Sketch (а); исходный фотопортрет (б); Artist Sketch (в); значения индекса SSIM между исходным фотопортретом и Viewed Sketch из популяции 1 (нижняя кривая П1) и популяции 2 (верхняя кривая П2) (г); значения индекса SSIM между исходным фотопортретом и Artist Sketch из популяции 1 (нижняя кривая П1) и популяции 2 (верхняя кривая П2) (д)

Как видно из приведенных результатов, значения индекса SSIM для скетчей, сформированных в популяции 2, выше порога сравнения в обоих случаях. Таким образом, можно утверждать, что скетчи из популяции 2 ближе к портретам-оригиналам и, следовательно, качество этих скетчей стало выше. При этом сформированные скетчи отвечают требованиям правдивого сценария, поскольку учитывают возможность неполной информации о портретах-оригиналах – неточно заданных или неполных параметрах «исходного фотопортрета».

А теперь покажем, что предлагаемый способ повышения качества скетчей может быть применен также и для фотороботов, применяемых в криминалистической практике. Используем фоторобот из работы [10], который интересен тем, что был распознан при ранге 72, т.е. портрет-оригинал для него находился на 72 месте в последовательности найденных по нему результатов.

На рис. 9 приведены: оригинальное фото (рис. 9, а) и соответствующий ему фоторобот (рис. 9, б) из работы [10], а также значения индексов SSIM между ними в сформированных популяциях. В верхнем ряду показаны: портрет-оригинал, вариант этого портрета, полученного в популяции 1 (цифры над ним – значения параметров p_1, p_2 и p_3); портрет, полученный в популяции 2 для значения $k = 10$. В нижнем ряду показаны: фоторобот, вариант фоторобота, полученного в популяции 1 (цифры над ним – значения параметров p_1, p_2 и p_3); фоторобот, полученный в популяции 2 для значения $k = 10$. Уже из этих результатов видно, что индекс SSIM для фотороботов популяции 2 выше, чем индекс SSIM популяции 1. Этот вывод подтверждается динамикой изменений индекса SSIM, что показано на графиках рис. 10.

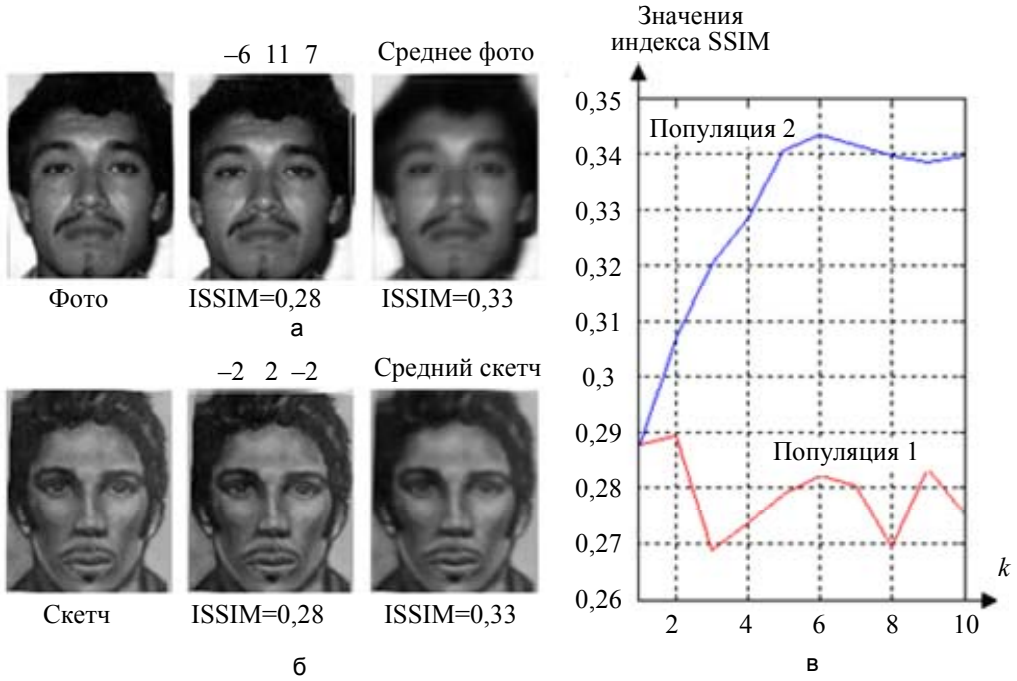


Рис. 9. Портрет-оригинал (фото) (а); фоторобот (скетч) (б) и значения индекса SSIM для них (в)

На рис. 10 показаны варианты Composite Forensic Sketches (см. исходные скетчи на рис. 4) для популяции 1 и 2, а на рис. 11 – значения индекса SSIM для них. Здесь также видно, что индекс SSIM для скетчей популяции 2 выше, чем индекс SSIM популяции 1.

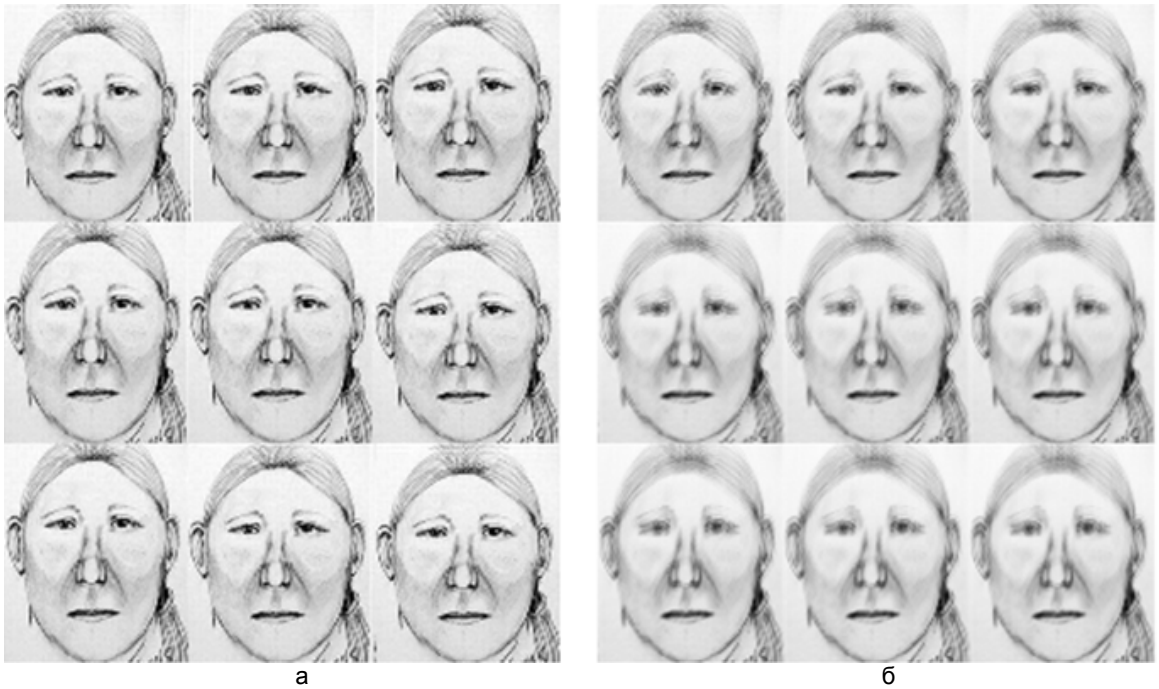


Рис. 10. Варианты Composite Forensic Sketches для популяций: 1 (а); 2 (б)

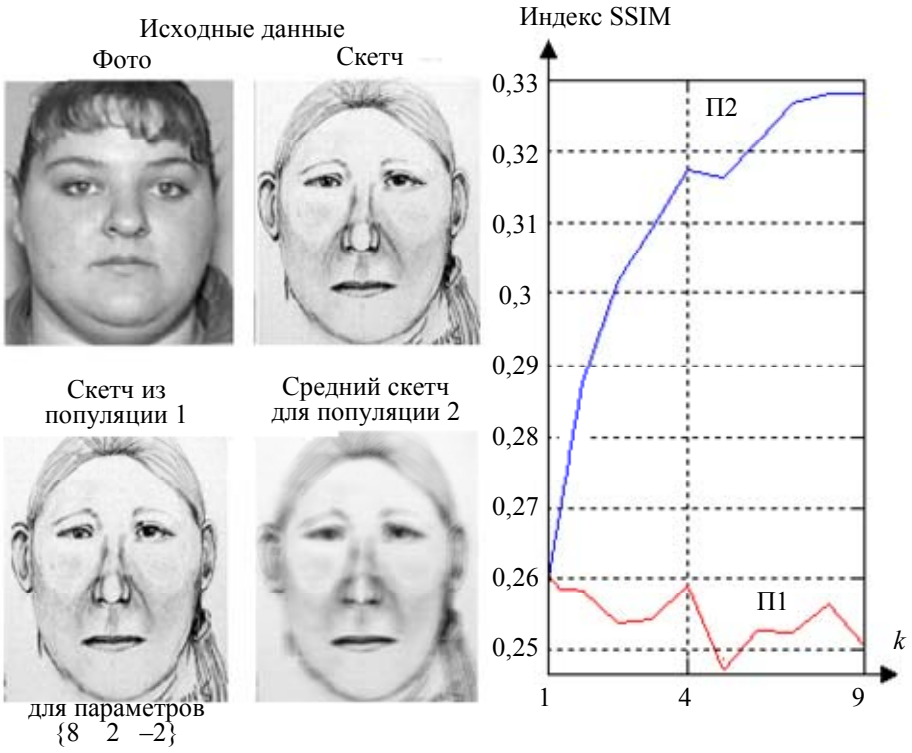


Рис. 11. Характеристики фотооригинала и Composite Forensic Sketch (а) и значения индекса SSIM для популяций 1 и 2 (б)

Подводя итог, отметим следующее.

1. Модификация исходных данных (фотороботов/скетчей) и их представление в форме популяции 1 имитирует получение новых данных от группы из K свидетелей. Этот эффект можно воспринимать как создание более объективного представления для фоторобота по имеющимся исходным данным. В таких предпосылках, уже в рамках популяции 1, можно достаточно эффективно решать задачу сравнения новых данных с фотооригиналом.
2. При этом само сравнение может выполняться со средним (для всей популяции) фотороботом или с каждым фотороботом из популяции на основе мажоритарных механизмов или, например, на основе смеси экспертов (Mixtures of experts [29]).
3. Модификация результата популяции 1 в результат популяции 2 улучшает подобие между парами фоторобот–оригинальный фотопортрет. С учетом отмеченного эффекта и в совокупности с механизмами, перечисленными в пункте 2, это создает новые условия для еще более эффективного сравнения фотороботов с фотопортретами-оригиналами.
4. Для оценки подобия между парами фоторобот–оригинальный фотопортрет можно использовать индекс SSIM, поскольку он оценивает корреляцию и текстуру локальных областей между исходными данными [27, 28].

Заключение

В настоящей работе выполнен обзор задач, связанных с проблемой поиска людей по фотороботам, а также систематизирован опыт и результаты, накопленные за последние два десятилетия по этой проблеме. Представлены исходные понятия, используемая терминология, идеи и современные технологии создания фотороботов, а также показаны трудности и причины неудач, возникающих в реальных сценариях поиска. Представлена история развития систем формирования композиционных портретов (фотороботов и скетчей) и идеи, реализованные в этих системах. Выполнен анализ задач автоматического сравнения фотороботов с оригинальными фотографиями, вскрыты причины недостижимости устойчивого поиска фотопортретов-оригиналов по фотороботам в реальных сценариях.

Сформулированы требования к базам фотороботов в дополнение к существующим базам изображений лиц, а также способы реализации таких баз. Как один из возможных вариантов рассмотрены методы генерации популяции фотороботов из исходного фоторобота для повышения результативности поиска по нему фотопортрета-оригинала. Представлен метод повышения индекса подобия в паре фоторобот-фотопортрет (оригинал), основанный на вычислении среднего фоторобота из сформированной популяции. Показано, что сформированные таким образом фотороботы более подобны портретам-оригиналам, и

их использование в обсуждаемой проблеме поиска может привести к высоким результатам. При этом сформированные фотороботы отвечают требованиям правдивого сценария, поскольку учитывают возможность неполной информации в словесных портретах. Обсуждаются результаты применения этих методов для двух популярных баз фото–скетчи, а также опубликованных в открытой печати фотороботов и соответствующих им фотопортретов. Эти примеры демонстрируют, что выявленный способ генерации скетчей обладает характеристикой универсальности за счет возможности независимого его использования (для скетчей и фотороботов) и любых доступных баз исходных данных.

По результатам выполненного обзора можно прогнозировать, что методы сравнения скетчей с соответствующими фотопортретами должны быть основаны на подходах, ориентированных на конкретные сценарии. Исходя из этого, дальнейшие исследования необходимо связывать с анализом различных сценариев, взятых из реальных ситуаций. Именно для этих случаев необходимо искать и создавать новые варианты синтеза скетчей и методы их распознавания.

Литература

1. Uhl R.J. Jr., Lobo N.V. A framework for recognizing a facial image from a police sketch // Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 1996. P. 586–593.
2. Konen W. Comparing facial line drawings with gray-level images: a case study on PHANTOMAS // Lecture Notes in Computer Science. 1996. V. 1112 LNCS. P. 727–734.
3. Криминалистика / Под ред. Н.П. Яблокова. 3-е изд. М.: Юристъ, 2005. 781 с.
4. Азбука криминалистики [Электронный ресурс]. Режим доступа: http://www.expert.aaanet.ru/arhiv/identif.htm, свободный. Яз. рус. (дата обращения 11.08.2014).
5. Identi-Kit Solutions [Электронный ресурс]. Режим доступа: http://www.identikit.net, свободный. Яз. англ. (дата обращения 11.08.2014).
6. IQ Biometrics. Faces Software [Электронный ресурс]. Режим доступа: http://www.iqbiometrix.com, свободный. Яз. англ. (дата обращения 11.08.2014).
7. Yuen P.C., Man C.H. Human face image searching system using sketch // IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans. 2007. V. 37. N 4. P. 493–504.
8. Tang X., Wang X. Face photo-sketch synthesis and recognition // Proc. 9th IEEE International Conference on Computer Vision. Nice, France, 2003. V. 1. P. 687–694.
9. Wang X., Tang X. Face photo-sketch synthesis and recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009. V. 31. N 11. P. 1955–1967.
10. Klare B., Li Z, Jain A.K. Matching forensic sketches to mug shot photos // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011. V. 33. N 3. P. 639–646.
11. Han H., Klare B.F., Bonnen K., Jain A.K. Matching composite sketches to face photos: a component-based approach // IEEE Transactions on Information Forensics and Security. 2013. V. 8. N 3. P. 191–204.
12. Davies G.M., Valentine T. Facial composites: forensic utility and psychological research. In: Handbook of Eyewitness Psychology. Mahwah: LEA, 2007. V. 2. P. 59–83.
13. Gibson S.J, Solomon C.J, Pallares-Bejarano A. Synthesis of photographic quality facial composites using evolutionary algorithms // Proc. British Machine Vision Conference. Norwich, UK, 2003. P. 221–230.
14. Frowd C.B., Hancock P.J.B., Carson D. EvoFIT: a holistic, evolutionary facial imaging technique for creating composites // ACM Transactions on Applied Psychology. 2004. V. 1. P. 1–21.
15. George B., Gibson S.J., Maylin M.I.S., Solomon C.J. EFIT-V - Interactive evolutionary strategy for the construction of photo-realistic facial composites // Proc. 10th Annual Conference on Genetic and Evolutionary Computation. Atlanta, USA, 2008. P. 1485–1490.
16. Frowd C.D., Pitchford M., Skelton F., Petkovic A., Prosser C., Coates B. Catching even more offenders with EvoFIT facial composites // Proc. 3rd International Conference on Emerging Seurity Technologies, EST-2012. Lisbon, Portugal, 2012. P. 20–26.
17. CUHK Face Sketch Database [Электронный ресурс]. Режим доступа: http://mmlab.ie.cuhk.edu.hk/facesketch.html, свободный. Яз. англ. (дата обращения 11.08.2014).
18. CUHK Face Sketch FERET Database [Электронный ресурс]. Режим доступа: http://mmlab.ie.cuhk.edu.hk/cufsf/, свободный. Яз. англ. (дата обращения 11.08.2014).
19. Zhang W., Wang X., Tang X. Coupled information-theoretic encoding for face photo-sketch recognition // Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011. P. 513–520.
20. Li X., Cao X. A simple framework for face photo-sketch synthesis // Mathematical Problems in Engineering. 2012. V. 2012. Art. 910719.
21. Kiani Galoogahi H., Sim T. Face photo retrieval by sketch example // Proc. 20th ACM International Conference Multimedia. Nara, Japan, 2012. P. 949–952.

22. Sharma A., Jacobs D.W. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch // Proc. 24th IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011. P. 593–600.
23. Chang L., Zhou M., Han Y., Deng X. Face sketch synthesis via sparse representation // Proc. 20th International Conference on Pattern Recognition. Istanbul, Turkey, 2010. P. 2146–2149.
24. Kukharev G.A., Buda K., Shchegoleva N.L. Methods of face photo-sketch comparison // Pattern Recognition and Image Analysis. 2014. V. 24. N 1. P. 102–113.
25. Kukharev G., Buda K., Shchegoleva N. Sketch generation from photo to create test databases // Przegląd Elektrotechniczny (Electrical Review). 2014. V. 90. N 2. P. 97–100.
26. Yu H., Zhang J.J. Mean value coordinates–based caricature and expression synthesis // Signal, Image and Video Processing. 2013. V. 7. N 5. P. 899–910.
27. Wang Z., Bovik A.C. A universal image quality index // IEEE Signal Processing Letters. 2002. V. 9. N 3. P. 81–84.
28. Wang Z., Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: from error visibility to structural similarity // IEEE Transactions on Image Processing. 2004. V. 13. N 4. P. 600–612.
29. Masoudnia S., Ebrahimpour R. Mixture of experts: a literature survey // Artificial Intelligence Review. 2014. V. 42. N 2. P. 275–293.

Кухарев Георгий Александрович – доктор технических наук, профессор, профессор, Западно-Поморский технологический университет в Щецине, Щецин, 70-310, Польша, gkukharev@wi.zut.edu.pl

Матвеев Юрий Николаевич – доктор технических наук, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; главный научный сотрудник, ООО «Центр речевых технологий», Санкт-Петербург, 196084, Российская Федерация, matveev@mail.ifmo.ru, matveev@speechpro.com

Щеголева Надежда Львовна – кандидат технических наук, доцент, Санкт-Петербургский государственный электротехнический университет (ЛЭТИ), Санкт-Петербург, 197376, Россия, NLSchegoleva@etu.ru

Georgy A. Kukharev – D.Sc., Full Professor, Westpomeranian University of Technology, Szczecin, 70-310, Poland, gkukharev@wi.zut.edu.pl

Yuri N. Matveev – D.Sc., Professor, ITMO University, Saint Petersburg, 197101, Russian Federation; Chief scientist, STC Ltd., Saint Petersburg, 196084, Russian Federation, matveev@mail.ifmo.ru, matveev@speechpro.com

Nadezhda L. Shchegoleva – PhD, Associate professor, Saint Petersburg Electrotechnical University (LETI), Saint Petersburg, 197376, Russian Federation, NLSchegoleva@etu.ru

Принято к печати 25.09.14
Accepted 25.09.14

6 ТЕХНИЧЕСКИЕ СИСТЕМЫ И ТЕХНОЛОГИИ ENGINEERING AND TECHNOLOGIES

UDC 004.9

EXTENDED SPEECH EMOTION RECOGNITION AND PREDICTION

T. Anagnostopoulos[a], S.E. Khoruzhnikov[a], V.A. Grudinin[a], C. Skourlas[b]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, thanag@mail.ifmo.ru
[b] Technological Educational Institute of Athens, Athens, 12243, Greece, cskourlas@teiath.gr

Abstract. Humans are considered to reason and act rationally and that is believed to be their fundamental difference from the rest of the living entities. Furthermore, modern approaches in the science of psychology underline that humans as a thinking creatures are also sentimental and emotional organisms. There are fifteen universal extended emotions plus neutral emotion: hot anger, cold anger, panic, fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt and neutral position. The scope of the current research is to understand the emotional state of a human being by capturing the speech utterances that one uses during a common conversation. It is proved that having enough acoustic evidence available the emotional state of a person can be classified by a set of majority voting classifiers. The proposed set of classifiers is based on three main classifiers: *k*NN, C4.5 and SVM RBF Kernel. This set achieves better performance than each basic classifier taken separately. It is compared with two other sets of classifiers: one-against-all (OAA) multiclass SVM with Hybrid kernels and the set of classifiers which consists of the following two basic classifiers: C5.0 and Neural Network. The proposed variant achieves better performance than the other two sets of classifiers. The paper deals with emotion classification by a set of majority voting classifiers that combines three certain types of basic classifiers with low computational complexity. The basic classifiers stem from different theoretical background in order to avoid bias and redundancy which gives the proposed set of classifiers the ability to generalize in the emotion domain space.

Keywords: speech emotion recognition, affective computing, machine learning.

Acknowledgements. The research was carried out with the financial support of the Ministry of Education and Science of the Russian Federation under grant agreement №14.575.21.0058.

РАСПОЗНАВАНИЕ И ПРОГНОЗИРОВАНИЕ ДЛИТЕЛЬНЫХ ЭМОЦИЙ В РЕЧИ

Т. Анагностопулос[a], С.Э. Хоружников[a], В.А. Грудинин[a], К. Скоурлас[b]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, vlad@digiton.ru
[b] Технический образовательный институт Афин, Афины, 12243, Греция, cskourlas@teiath.gr

Аннотация. Люди действуют рационально, и это их фундаментальное отличие от других видов жизни. Кроме того, в современной психологии подчеркивается, что люди как разумные создания отличаются чувствами и эмоциями. Существует пятнадцать видов универсальных длительных эмоций, плюс нейтральное эмоциональное состояние, такие как гнев, злость, паника, страх, тревога, отчаяние, грусть, восторг, радость, интерес, скука, стыд, гордость, отвращение, презрение и нейтральное отношение. В данном исследовании рассматривается понимание эмоционального состояния человека по анализу речи в процессе общения. Доказано, что на основе достаточного объема акустических данных эмоциональное состояние человека может быть классифицировано набором мажоритарных классификаторов. Предложенный набор классификаторов построен на основе трех базовых классификаторов: kNN, C4.5 и SVMRBFKernel. Этот набор обеспечивает лучшую обработку классификаций эмоций, чем каждый из базовых классификаторов в отдельности. Он сравнивается с двумя другими наборами классификаторов: один-против-всех (OAA) мультиклассовый SVM с гибридными ядрами и с набором классификаторов, состоящим из двух базовых классификаторов C5.0, и нейронная сеть (NeuralNetwork). Предложенный вариант достигает лучшего результата, чем два других набора классификаторов. В настоящей статье осуществляется классификация эмоций набором мажоритарных классификаторов, который состоит из трех определённых базовых классификаторов, имеющих низкую вычислительную сложность. Базовые классификаторы базируются на различных теоретических данных с целью избегания отклонений и избыточности, что дает предложенному набору классификаторов возможность обобщиться в пространство определений эмоций.

Ключевые слова: распознавание эмоций в речи, расчет эмоций, машинное обучение

Благодарности. Исследования проводились при финансовой поддержке Министерства образования и науки Российской Федерации в рамках Соглашения о предоставлении субсидии №14.575.21.0058.

Introduction

Humans are considered to reason and act rationally and that is believed to be their fundamental factor that differentiates them from the rest of living entities. Although, modern approaches in the science of psychology underline that human except of thinking creatures are also sentimental and emotional organisms. The field of psychology that studies this aspect of human nature is Emotion Intelligence [1]. Emotion is a subjective, conscious experience characterized primarily by psycho physiological expressions, biological reactions and mental states. It is often associated and considered reciprocally influential with mood, temperament, personality, disposition and motivation [2]. Emotion is often the driving force behind motivation, positive or negative [3]. There are fifteen universal extended emotions plus neutral emotion, that is: hot anger, cold anger, panic, fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt and neutral [4].

The experience of emotion is referenced as affect and it is a key part of the process of an organism's interaction with stimuli [5]. Affect also refers to affect display [6], which is the facial, speech or gestural behavior that serves as an indicator of affect. Affective computing is the study and development of systems and devices that can recognize, interpret, process and simulate human affects. It is an interdisciplinary field spanning from informatics, psychology and cognitive science [7]. One field of informatics that could be used in order to classify affects and exploit their fundamental emotional state is machine learning. Thus we expand the wide area of the affective computing with this of machine learning algorithms and classification models [8].

In the current paper the problem of speech emotion recognition will be treated as an ensemble classification and prediction issue. First, a number of base classifiers are going to be used for speech emotion classification. Then an ensemble majority voting classifier will expand the dynamics of the base classifiers in order to create a concrete classification model. The proposed model is evaluated with other state-of-the-art models and it is proven that it achieves higher classification scores over the other models.

The paper is organized as follows. In Section "Related Work", the related work of the state-of-the-art speech classification models is presented. In Section "Data Model", the data model which is used in the current study is described. In Section "Ensemble Classification", it is described how an ensemble classifier is built from a set of base classifiers. In Section "Emotion Prediction", it is presented how speech emotion can be predicted. In Section "Performance Evaluation", the evaluation of the proposed model with the other state-of-the-art models is performed. In Section "Discussion and Conclusion", a discussion is done in order to explain the effect of the proposed classification model. The paper concludes with Section "References", where future work and trends are outlined.

Related Work

A vast amount of work has been done in the area of speech emotion recognition. Among all we can distinguish [9] where an automatic feature selector which combined the random forest RF2TREE ensemble algorithm and the simple decision tree C4.5 algorithm is developed. In [10] a Hidden Markov Model (HMM) is proposed for joint speech and emotion recognition in order to include multiple versions of each emotion. Then emotion classification was performed using ensemble majority voting between emotion labels. The authors in [11] demonstrate commonly used k Nearest Neighbors (kNN) classifier for segment-based speech emotion recognition and classification. In [12] frame-wise emotion classification is used based on vector quantization techniques. Within this scheme in order to classify an input utterance an emotion was classified using an ensemble majority voting scheme between frame-level emotion labels. The authors in [13] used Fuzzy Logic classification in order to combine categorical and primitives-based speech emotion recognition.

The authors in [14] implemented a real-time system for discriminating between neutral and angry speech which used Gaussian Mixture Models (GMMs) for Mel-Frequency CepstralCoefficients (MFCC) features in combination with a prosody-based classifier. In [15] it is demonstrated that emotion can be better differentiated by specific phonemes than others using phoneme-specific GMM. The authors in [16] investigate combination of features at different levels of granularity by integrating GMM log-likelihood score with commonly-used suprasegmental prosody-based emotion classifiers. In [17] GMMs are applied to emotion recognition using a combined feature set which was obtained by concatenating MFCC and prosodic features.

The authors in [18] demonstrated Support Vector Machine (SVM) classification with manifold learning methods using covariance matrices of prosodic and spectral measures evaluated over the entire utterance. In [19] an SVM Multiple Kernel Learning (MKL) is proposed, where the decision rule is a weighted linear combination of multiple single kernel function outputs. The authors in [20] introduce SVM classification with Radial Basis Function (RBF) kernels and MFCC statistics over phoneme type classes in the utterance. In [21] the authors use an ensemble mixture model of base SVM classifiers where the outputs of the classifiers are normalized and combined using a thresholding fusion function in order to classify the speech emotion. The authors in [22] use an ensemble mixture model of which combines C5.0 and Neural Network (NN) base classifiers in order to achieve speech emotion classification.

Finally, in [23] the authors use an ensemble majority voting classifier which combines kNN, C4.5 and SVM Polynomial Kernel. The results were apparently better than the previous approaches. However, the emotions classified were limited to the six basic emotions with regards to the Ekman's emotion taxonomy [24]. In this paper we propose a model which is designed to extend the classification to sixteen emotions. The proposed model is compared with the model in [21] and the model in [22], given the same speech emotion database [25]. The results show that the proposed model achieves better performance than those models.

Data Model

HUMAINE [25] database is used in order to perform emotion classification from speech utterances. The speech utterances ranged from positive to negative emotions. We used fifteen universal extended emotions plus neutral emotion, that is: hot anger, cold anger, panic, fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt and neutral. A set of acoustic parameters which are related to the aforementioned emotional [4] are employed.

The acoustic parameters are:
– F0:
1. Perturbation,
2. Mean,
3. Range,
4. Variability,
5. Contour,
6. Shift Regularity.

– Formants:
7. F1 Mean,
8. F2 Mean,
9. F1 Bandwidth,
10. Formant Precision.

– Intensity:
11. Mean,
12. Range,
13. Variability.

– Spectral Parameters:
14. Frequency range,
15. High-frequency energy,
16. Spectral noise.

– Duration:
17. Speech rate,
18. Transition time.

We perform z-transformation [26] to these eighteen acoustic parameters and we feed them to the base classifiers, as it is discussed in the next section.

Ensemble Classification

The proposed model is based on ensemble classification majority voting scheme over certain types of base classifiers which are of low computational complexity [27]. Three base classifiers are used from different theoretical background in order to avoid bias and redundancy [8].

The three base classifiers are:
1. *k*NN, which is a nonparametric classifier,
2. C4.5, which is a nonmetric classifier, and
3. SVM with RBF Kernel, which is a linear discriminant function classifier.

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**0.9**	0	0	0	0.1	0	0	0	0	0	0	0	0
	E5	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0
	E6	0	0	0	0	0	**0.8**	0	0	0.2	0	0	0	0	0	0	0
	E7	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0
	E8	0	0	0	0	0	0	0	**0.9**	0	0	0	0	0.1	0	0	0
	E9	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0	0	**0.8**	0.2	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0
	E12	0	0	0	0	0	0	0	0.1	0	0	0	**0.9**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0
	E15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 1. Confusion matrix of *k*NN for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.95$

Classification problem

A number of observation pairs (x_i, y_i) $i = 1, \dots, n$ where $x \in X \subset \mathbb{R}^p$ and $y \in Y = \{$hot anger, cold anger, panic, fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt, neutral$\}$ is observed. $X$is known as the predictor space (or attributes) and Y is the response space (or class). In this case the

number of attributes is eighteen like the number of the acoustic parameters. The objective is to use these observations in order to estimate the relationship between $X$and Y, thus predict Y from X. Usually the relationship is denoted as a classification rule,

$$h_j(X) = \arg\max P(y|X, \theta_j) \quad \text{(Eq. 1)}$$

where, $j = 1, \ldots, 3$, and $P(.,.)$ is the probability distribution of the observed pairs, θ is the parameter vector for each base classifier, and j is the number of the base classifiers. In this case, we have three classification rules, one for each base classifier.

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0
	E5	0	0	0	0	**0.9**	0	0	0	0	0	0	0	0	0	0.1	0
	E6	0	0	0	0	0	**0.8**	0	0	0	0	0.2	0	0	0	0	0
	E7	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0
	E8	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0
	E9	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0	0	**0.9**	0.1	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0
	E12	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0.1	0	0	0	**0.9**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0
	E15	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	**0.9**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 2. Confusion matrix of C4.5 for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.96$

Ensemble majority voting classification

Each of the three base classifiers is an expert in a different region of the predictor space because they treat the attribute space under different theoretical basis [28]. The three classifiers could be combined in such a way in order to produce an ensemble majority voting classifier that is superior to any of the individual rules. A popular way to combine these three base classification rules is to let an ensemble classifier,

$$C(X) = mode\,\{h_1(X), h_2(X), h_3(X)\} \quad \text{(Eq. 2)}$$

to classify X to the class that receives the largest number of classifications (or votes) [29]. In the next section the three base classifiers and the ensemble classifier are built. It is shown that the ensemble majority voting classifier achieves better accuracy as it is analyzed in the relative confusion matrices.

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0.1	0
	E5	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0
	E6	0	0	0	0	0	**0.9**	0	0	0	0.1	0	0	0	0	0	0
	E7	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0
	E8	0	0.1	0	0	0	0	0	**0.9**	0	0	0	0	0	0	0	0
	E9	0	0	0	0	0	0	0	0	**0.9**	0	0.1	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0.1	0	**0.9**	0	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0
	E12	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0	0.1	0	0	**0.9**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0
	E15	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	**0.9**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 3. Confusion matrix of SVM RBF Kernel for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.95$

Prediction accuracy

In order to measure the accuracy of the classification, we define the metric of classification accuracy. In the case of a separate emotional class, we define the emotion class accuracy,

$$e_k = \frac{tp_k + tn_k}{tp_k + tn_k + fp_k + fn_k} \quad \text{(Eq. 3)}$$

here, $k = 1, \dots, 16$, denotes the number of the emotional classes, and tp_k, tn_k, fp_k, fn_k denote the emotion class true positive, true negative, false positive and false negative classified utterances, respectively. In the case of all emotional classes in average, we define the prediction accuracy,

$$p = \frac{\sum_{k=1}^{16} e_k}{16} \quad \text{(Eq. 4)}$$

which denotes the overall accuracy of a classifier given a specific observed number of observation pairs (x_i, y_i) $i = 1, \dots, n$ for the fifteen universal extended emotions plus neutral emotion.

Emotion Prediction

10-*fold-cross-validation* technique [30] is used, provided by WEKA data mining open source workbench [31], in order to measure the emotion class accuracy e_k and the prediction accuracy p for the proposed classification scheme. HUMAINE [25] database is used in order to perform emotion classification from the same speech utterances. Specifically, English language speech information of 48 persons (26 males and 22 females) is exploited. Every person has expressed speech utterances of the fifteen universal extended emotions plus neutral emotion, thus the total number of the observed pairs is $n = 768$. Because of space limitations in visualizing the results in tabular format a label is assigned to each emotion, that is E1: hot anger, E2: cold anger, E3: panic, E4: fear, E5: anxiety, E6: despair, E7: sadness, E8: elation, E9: happiness, E10: interest, E11: boredom, E12: shame, E13: pride, E14: disgust, E15: contempt and E16: neutral.

The confusion matrix [32] is presented in Table 1 for the emotion class accuracy e_k and the prediction accuracy p of the *k*NN nonparametric classifier. Table 2 presents the confusion matrix for the emotion class accuracy e_k and the prediction accuracy p of the C4.5 nonmetric classifier. Table 3 presents the confusion matrix for the emotion class accuracy e_k and the prediction accuracy p of the SVM RBF Kernel classifier. The confusion matrix for the emotion class accuracy e_k and the prediction accuracy p of the proposed ensemble majority voting classifier are presented in Table 4.

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**0.9**	0	0	0	0.1	0	0	0	0	0	0	0	0
	E5	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0
	E6	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0
	E7	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0
	E8	0	0.1	0	0	0	0	0	**0.9**	0	0	0	0	0	0	0	0
	E9	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0	0	**0.9**	0.1	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0
	E12	0	0	0	0	0	0	0	0.1	0	0	0	**0.9**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0
	E15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 4. Confusion matrix of Ensemble Majority Voting Classifier for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.97$

Confusion matrices were obtained from WEKA. Specifically, 10 fold cross validation is used where the sample is divided into 10 equal length parts. There is no resampling in the classification process. For 10 consecutive repetitions the classifier is trained with 9 parts and tested with the remaining 1 part. During the repetitions each of the 10 parts is considered only one time for testing. For each repetition classification results are computed which are summarized to true positives, true negatives, false positives and false negatives. After the 10 repetitions the classification results are averaged and presented in the confusion matrices. Confusion matrices have more information than the presented classification schema in (Eq. 2) because expect of the true positives and true negatives, described in (Eq. 2), they also incorporate the false positives and false negatives as well. In WEKA, confidence interval for each acoustic parameter value is set to 95 percent.

Table 5 depicts the overall emotion class accuracy e_k for the three base classifiers and the proposed ensemble majority voting classifier. Table 6 depicts the overall prediction accuracy p for the three base classifiers and the proposed ensemble majority voting classifier. As it is proved, the emotion class accuracy e_k and the prediction accuracy p of the ensemble majority voting classifier is greater than these of the three base classifiers. In the discussion Section "Discussion and Conclusion" it is explained why these experimental results are observed.

Performance Evaluation

The proposed model is compared with other two classification models [21] and [22], in literature and it is proved to achieve better results by means of emotion class accuracy e_k and prediction accuracy p, given the same speech emotion HUMAINE database [25]. The same experimental setup is used as in Section "Emotion Prediction". 10-*fold-cross-validation* technique is used in order to measure the emotion class accuracy e_k and the prediction accuracy p for the compared classification schemes. The model in [21] uses a one-against-all (OAA) multiclass SVM classification scheme with Hybrid kernel functions, which constitutes an ensemble classifier. The core of OAA for multiclass SVM classifiers, as it is introduced in [33], is that the observed pair (x_i, y_i) $i = 1, \dots, n$ can be classified only if one of the SVM classes accepts the observed pair while all other SVMs reject it at the same time, thus making a unanimous decision. The model in [22] uses an ensemble classifier which constitutes of a combination of C5.0 and NN base classifiers. The core of the combined ensemble classifier is that it classifies an observed pair (x_i, y_i) $i = 1, \dots, n$ to the class with the higher probability density function (PDF) among the two base classifiers.

		Emotion Class Accuracy e_k															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Classifier	*k*NN	1.0	0.9	1.0	0.9	1.0	0.8	1.0	0.9	1.0	0.8	1.0	0.9	1.0	1.0	1.0	1.0
	C4.5	1.0	1.0	1.0	1.0	0.9	0.8	1.0	1.0	1.0	0.9	1.0	1.0	0.9	1.0	0.9	1.0
	SVM RBF	1.0	0.9	1.0	0.9	1.0	0.9	1.0	0.9	0.9	0.9	1.0	1.0	0.9	1.0	0.9	1.0
	Majority	**1.0**	**1.0**	**1.0**	**0.9**	**1.0**	**1.0**	**1.0**	**0.9**	**1.0**	**0.9**	**1.0**	**0.9**	**1.0**	**1.0**	**1.0**	**1.0**

Table 5. Emotion Class Accuracy e_k for the three base classifiers and the proposed Ensemble Majority Voting Classifier

		Prediction Accuracy p
Classifier	*k*NN	0.95
	C4.5	0.96
	SVM RBF	0.95
	Majority	**0.97**

Table 6. Prediction Accuracy p for the three base classifiers and the proposed Ensemble Majority Voting Classifier

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0.1	0
	E5	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0
	E6	0	0	0	0	0	**0.9**	0	0	0	0.1	0	0	0	0	0	0
	E7	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0
	E8	0	0.1	0	0	0	0	0	**0.9**	0	0	0	0	0	0	0	0
	E9	0	0	0	0	0	0	0	0	**0.9**	0	0.1	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0.1	0	**0.9**	0	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**0.9**	0	0	0	0.1	0
	E12	0	0	0	0	0	0	0	0.1	0	0	0	**0.9**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0	0.1	0	0	**0.9**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0
	E15	0	0	0	0	0	0	0	0.2	0	0	0	0	0	0	**0.8**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 7. Confusion matrix of OAA multiclass SVM classifier with Hybrid kernel functions for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.93$

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0.1	0
	E5	0	0	0.1	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0
	E6	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0
	E7	0.1	0	0	0	0.1	0	**0.8**	0	0	0	0	0	0	0	0	0
	E8	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0
	E9	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0	0	**0.9**	0.1	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0
	E12	0	0	0	0	0	0	0	0.2	0	0	0	**0.8**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0.1	**0.9**	0	0
	E15	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	**0.9**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 8. Confusion matrix of Combined C5.0 and NN classifier for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.94$

Table 7 presents the confusion matrix for the emotion class accuracy e_k and the prediction accuracy p of the OAA multiclass SVM with hybrid kernel functions. Table 8 presents the confusion matrix for the emotion class accuracy e_k and the prediction accuracy p of the combined C5.0 and NN classifier. Table 9 depicts the overall emotion class accuracy e_k for these two models and our ensemble majority voting classifier. Table 10 depicts the overall prediction accuracy p for these two models and our ensemble majority voting classifier.

As it is proved the emotion class accuracy e_k and the prediction accuracy p of the ensemble majority voting classifier is greater than these of the other two compared classifiers. In the discussion Section "Discussion and Conclusion" it is explained why it is observed these experimental results.

		Emotion Class Accuracy e_k															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Classifier	SVM Hybrid	1.0	0.9	1.0	0.9	1.0	0.9	1.0	0.9	0.9	0.9	0.9	0.9	0.9	1.0	0.8	1.0
	C5.0 – NN	1.0	1.0	1.0	0.9	0.9	1.0	0.8	1.0	1.0	0.9	1.0	0.8	1.0	0.9	0.9	1.0
	Majority	**1.0**	**1.0**	**1.0**	**0.9**	**1.0**	**1.0**	**1.0**	**0.9**	**1.0**	**0.9**	**1.0**	**0.9**	**1.0**	**1.0**	**1.0**	**1.0**

Table 9. Emotion Class Accuracy e_k for the two compared classifiers and the proposed Ensemble Majority Voting Classifier

Discussion and Conclusion

A discussion is performed in order to explain why these experimental results are observed in the two previous Sections "Emotion Prediction" and "Performance Evaluation". In Section "Emotion Prediction" it is proved that the ensemble majority voting classifier achieves better scores than the three base classifiers. This is explained because each base classifier is biased in a specific domain of the emotion classification problem, thus the advantages of one classifier might be disadvantages for the other two classifiers and vice versa. The overall superiority of the ensemble classifier is its ability to combine the redundant information of the base classifiers in order to create a more sound classification scheme.

		Prediction Accuracy p
Classifier	SVM Hybrid	0.93
	C5.0 – NN	0.94
	Majority	**0.97**

Table 10. Prediction Accuracy p for the two compared classifiers and the proposed Ensemble Majority Voting Classifier

In Section "Performance Evaluation" it is also proved that the ensemble majority voting classifier achieves better scores than the two compared ensemble classifiers. In the case of [21] model this is explained because the base classifiers are of the same general SVM linear discriminant functions bias. In the case of [22] model this is explained because the base classifiers where too few (i.e., only two) in order their union not to be able to generalize to the whole set of pairs $(x_i, y_i)\ i = 1, \ldots, n$. Both [21] and [22] models do not take into

consideration the majority votes, of the whole set of the classifiers (see Eq. 2, Section "Ensemble Classification"), which is used by the proposed model.

It is proved that the proposed ensemble majority voting classifier achieves better performance in classifying the fifteen universal extended emotions plus neutral emotion than the three base classifiers and the other two compared ensemble classifiers. Future work is intended by exploiting other context (i.e., facial expressions) in order to design multimodal models.

References

1. Matthews G., Zeidner M., Roberts R.D. *Emotional Intelligence: Science and Myth*. Cambridge, MIT Press, 2003, 697 p.
2. Schacter D.L. *Psychology*. 2nd ed. NY, Worth Publishers, 2011, 624 p.
3. Gaulin S.J.C., McBurney D.H. *Psychology: An Evolutionary Approach*. Upper Saddle River, Prentice Hall, 2003.
4. Scherer K.R. Vocal communication of emotion: a review of research paradigms. *Speech Communication*, 2003, vol. 40, no. 1–2, pp. 227–256. doi: 10.1016/S0167-6393(02)00084-5
5. Thompson E.R. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-Cultural Psychology*, 2007, vol. 38, no. 2, pp. 227–242. doi: 10.1177/0022022106297301
6. Parkinson B., Simons G. Worry spreads: interpersonal transfer of problem-related anxiety. *Cognition and Emotion*, 2012, vol. 26, no. 3, pp. 462–479. doi: 10.1080/02699931.2011.651101
7. Picard R.W. *Affective Computing*. Cambridge, MIT Press, 2000, 304 p.
8. Duda R.O., Hart P.E., Stork D.G. *Pattern Classification*. NY, John Wiley and Sons, 2000, 735 p.
9. Rong J., Chen Y.-P.P. Chowdhury M., Li G. Acoustic features extraction for emotion recognition. *Proc. 6th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2007*, 2007, art. 4276418, pp. 419–424. doi: 10.1109/ICIS.2007.48
10. Meng H., Pittermann J., Pittermann A., Minker W. Combined speech-emotion recognition for spoken human-computer interfaces. *Proc. IEEE International Conference on Signal Processing and Communications*, 2007, art. 4728535, pp. 1179–1182. doi: 10.1109/ICSPC.2007.4728535
11. Shami M.T., Kamel M.S. Segment-based approach to the recognition of emotions in speech. *Proc. IEEE International Conference on Multimedia and Expo, ICME 2005,* 2005, vol. 2005, art. 1521436, pp. 366–369. doi: 10.1109/ICME.2005.1521436
12. Sato N., Obuchi Y. Emotion recognition using mel-frequency cepstral coefficients. *Journal of Natural Language Processing*, 2007, vol. 14, no. 4, pp. 83–96. doi: 10.5715/jnlp.14.4_83
13. Grimm M., Mower E. Kroschel K., Narayanan S. Combining categorical and primitives-based emotion recognition. *Proc. 14th European Signal Processing Conference*. Florence, Italy, 2006, pp. 345–357.
14. Kim S., Georgiou P.G., Lee S., Narayanan S. Real-time emotion detection system using speech: multi-modal fusion of different timescale features. *Proc. 9th IEEE International Workshop on Multimedia Signal Processing, MMSP 2007*. Chania, Crete, 2007, art. 4412815, pp. 48–51. doi: 10.1109/MMSP.2007.4412815
15. Sethu V., Ambikairaja E., Epps J. Phonetic and speaker variations in automatic emotion classification. *Proc. Annual Conference of the International Speech Communication Association, Interspeech*. Brisbane, Australia, 2008, pp. 617–620.
16. Vlasenko B., Schuller B., Wendemuth A., Rigoll G. Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. *Affective Computing and Intelligent Interaction*, 2007, vol. 4738 LNCS, pp. 139–147.
17. Vondra M., Vich R. Recognition of emotions in german speech using gaussian mixture models. *Multimodal Signals: Cognitive and Algorithmic Issues*, 2009, vol. 5398 LNAI, pp. 256–263. doi: 10.1007/978-3-642-00525-1_26
18. Ye C., Liu J., Chen C., Song M., Bu J. Speech emotion classification on a riemannian manifold. *Advances in Multimedia Information Processing – PCM 2008*, 2008, vol. 5353 LNCS, pp. 61–69. doi: 10.1007/978-3-540-89796-5_7
19. Gonen M., Alpaydin E. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 2011, vol. 12, pp. 2211–2268.
20. Bitouk D., Verma R., Nenkova A. Class-level spectral features for emotion recognition. *Speech Communication*, 2010, vol. 52, no. 7–8, pp. 613–625. doi: 10.1016/j.specom.2010.02.010
21. Yang N., Muraleedharan R., Kohl J., Demirkol I., Heinzelman W., Sturge-Apple M. Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion. *Proc. 4th IEEE Workshop on Spoken Language Technology, SLT 2012*. Miami, Florida, 2012, art. 6424267, pp. 455–460. doi: 10.1109/SLT.2012.6424267

22. Javidi M.M., Roshan E.F. Speech emotion recognition by using combinations of C5.0, neural network (NN), and support vectors machines (SVM) classification methods. *Journal of Mathematics and Computer Science*, 2013, vol. 6, no. 3, pp. 191–200.
23. Anagnostopoulos T., Skourlas C. Ensemble majority voting classifier for speech emotion recognition and prediction. *Journal of Systems and Information Technology*, 2014, vol. 16, no. 3, pp. 222–232. doi: 10.1108/JSIT-01-2014-0009
24. Ekman P. An argument for basic emotions. *Cognition and Emotion*, 1992, pp. 169–200.
25. Douglas-Cowie E., Cowie R., Sneddon I., Cox C., Lowry O., McRorie M., Martin J.-C., Devillers L., Abrilian S., Batliner A., Amir N., Karpouzis K. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. *Proc. 2nd International Conference on Affective Computing and Intelligent Interaction, ASCII 2007*. Lisbon, Portugal, 2007, vol. 4738 LNCS, pp. 488–500.
26. Jury E.I. *Theory and Application of the Z-Transform Method*. Malabar, Krieger Pub Co, 1973, 330 p.
27. Friedman J., Hastie T., Tibshirani R. *The Elements of Statistical Learning*. NY, Springer, 2001, 524 p.
28. Alpaydin E. *Introduction to Machine Learning*. 2nd ed. Cambridge, MIT Press, 2010, 581 p.
29. Basu S., Dasgupta A. The mean, median, and mode of unimodal distributions: a characterization. *Theory of Probability and its Applications*, 1997, vol. 41, no. 2, pp. 210–223. doi: 10.1137/S0040585X97975447
30. Seymour G. *Predictive Inference*. NY, Chapman and Hall, 1993, 240 p.
31. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H. The WEKA data mining software: an update. *SIGKDD Explorations*, 2009, vol. 11, no. 1, pp. 10–18. doi: 10.1145/1656274.1656278
32. Stehman S.V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 1997, vol. 62, no. 1, pp. 77–89. doi: 10.1016/S0034-4257(97)00083-7
33. Vapnik V.N. *The Nature of Statistical Learning Theory*. 2nd ed. NY, Springer, 2000, 314 p.

Theodoros Anagnostopoulos – Lead Research Associate, Department of Infocommunication Technologies, ITMO University, Saint Petersburg, 197101, Russian Federation, thanag@mail.ifmo.ru

Sergei E. Khoruzhnicov – Dean of the Faculty, Department of Infocommunication Technologies, ITMO University, Saint Petersburg, 197101, Russian Federation, xse@mail.ifmo.ru

Vladimir A. Grudinin – Head of Department, Department of Infocommunication Technologies, ITMO University, Saint Petersburg, 197101, Russian Federation, grudinin@mail.ifmo.ru

Christos Skourlas – PhD, Professor, Professor, Department of Informatics, Technological Educational Institute of Athens, Athens, 12243, Greece, cskourlas@teiath.gr

Анагностопулос Теодорос – PhD, ведущий научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, thanag@mail.ifmo.ru

Хоружников Сергей Эдуардович – кандидат физ.-мат. наук, доцент, декан, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, xse@mail.ifmo.ru

Грудинин Владимир Алексеевич – кандидат технических наук, доцент, заведующий кафедрой, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, vlad@digiton.ru

Скоурлас Кристос – PhD, профессор, профессор в департаменте информатики, Технический образовательный институт Афин, Афины, 12243, Греция, cskourlas@teiath.gr

Принято к печати 01.09.14
Accepted 01.09.14

УДК 621.039.564

РАЗРАБОТКА РЕЗЕРВИРОВАННОГО БЛОКА УПРАВЛЕНИЯ ЭЛЕКТРОПРИВОДОМ НА ОСНОВЕ АВТОМАТНОГО ПОДХОДА

Ю.Ю. Янкин[a], А.А. Шалыто[b]

[a] ОАО «Концерн «НПО «Аврора», Санкт-Петербург, 194021, Российская Федерация, yankinyy@gmail.com
[b] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

Аннотация. Рассматривается реализация резервированной аппаратуры управления электроприводом, выполненной на основе микросхем с программируемой структурой. Резервирование выполнено способом постоянного общего резервирования. В отличие от раздельного резервирования замещением с использованием ненагруженного резерва, общего резервирования замещением, такой способ обеспечивает сохранение всех функций аппаратуры при переходе на резерв, а также непрерывный контроль исправности основного и резервного каналов. Приведен пример реализации такой аппаратуры. Структурная схема канала управления электроприводом содержит два блока управления – основной и резервный, а также четыре источника питания. Программирование аппаратуры осуществлялось с использованием автоматного подхода. Разработана модель канала управления электроприводом, обеспечивающая совместное моделирование управляющей машины состояний и силового преобразователя. Благодаря наглядности и иерархичности конечных автоматов было сокращено время отладки по сравнению с традиционным программированием. Для синтеза управляющей машины состояний в системе проектирования производителя микросхем необходимо описать ее на языке описания аппаратуры. Такое описание формировалось автоматически средствами пакета MATLAB. Для проверки результатов были разработаны и изготовлены два образца блока управления, два образца источников вторичного электропитания и макет прибора. Блоки устанавливались в макет прибора. Образцы блоков были выполнены в соответствии с требованиями, предъявляемыми к поставляемой аппаратуре. Приведены результаты моделирования и испытаний канала управления в исправном состоянии, а также при имитации неисправности основного блока управления. Автоматный подход позволил наблюдать и отлаживать переходы в управляющей машине состояний при моделировании переходных процессов, протекающих при имитации неисправностей. Результаты работы могут быть использованы при создании отказоустойчивых каналов управления электроприводом.

Ключевые слова: электропривод, резервирование, автоматное программирование, конечный автомат, микросхемы с программируемой структурой, программируемые логические интегральные схемы.

REDUNDANT ELECTRIC MOTOR DRIVE CONTROL UNIT DESIGN USING AUTOMATA-BASED APPROACH

Y. Y. Yankin[a], A. A. Shalyto

[a] Concern "AURORA", Scientific and Production Association, Joint Stock Company ("Aurora", JSC), Saint Petersburg, 194021, Russian Federation, yankinyy@gmail.com
[b] ITMO University, Saint Petersburg, 197101, Russian Federation

Abstract. Implementation of redundant unit for motor drive control based on programmable logic devices is discussed. Continuous redundancy method is used. As compared to segregated standby redundancy and whole system standby redundancy, such method provides preservation of all unit functions in case of redundancy and gives the possibility for continuous monitoring of major and redundant elements. Example of that unit is given. Electric motor drive control channel block diagram contains two control units – the major and redundant; it also contains four power supply units. Control units programming was carried out using automata-based approach. Electric motor drive control channel model was developed; it provides complex simulation of control state-machine and power converter. Through visibility and hierarchy of finite state machines debug time was shortened as compared to traditional programming. Control state-machine description using hardware description language is required for its synthesis with FPGA-devices vendor design software. This description was generated automatically by MATLAB software package. To verify results two prototype control units, two prototype power supply units, and device mock-up were developed and manufactured. Units were installed in the device mock-up. Prototype units were created in accordance with requirements claimed to deliverable hardware. Control channel simulation and tests results in the perfect state and during imitation of major element fault are presented. Automata-based approach made it possible to observe and debug control state-machine transitions during simulation of transient processes, occurring at imitation of faults. Results of this work can be used in development of fault tolerant electric motor drive control channels.

Keywords: electric drive, redundancy, automata-based programming, finite state machine, programmable logic device, FPGA.

Введение

Традиционно функция регулирования мощности ядерной энергетической установки (ЯЭУ) атомных ледоколов осуществляется с помощью электромеханического привода, выполненного на основе специально разработанного шагового двигателя (ШД). Согласно одному из требований ОПБ-К-98/05[1], единичные отказы средств управляющих систем безопасности не должны препятствовать дистанционному приведению в действие систем безопасности. Выполнение этого требования невозможно без использования резервирования. Среди работ, касающихся резервирования в электромеханических приводах, известны [1–8]. Работы [1, 2] посвящены исследованию отказоустойчивых силовых преобразователей для электроприводов переменного тока. В работах [3, 4, 8] рассматриваются вопросы проектирования электриче-

[1] ОПБ-К-98/05 «Общие положения обеспечения ядерной и радиационной безопасности корабельных ядерных энергетических установок». М., 2005. 36 с.

ских двигателей с резервированными обмотками. Авторы публикаций [5–7] рассматривают различные способы управления электрическими двигателями с резервированными обмотками. Все упомянутые работы касаются электромеханического привода на основе вентильных или асинхронных электродвигателей. В настоящей работе рассматриваются резервированные блоки управления электромеханическим приводом на основе специально разработанного шагового электродвигателя. Указанные блоки могут быть построены на основе аналоговых микросхем, цифровых микросхем низкой степени интеграции, цифровых микросхем высокой степени интеграции или микроконтроллеров. Резервирование таких блоков может быть выполнено способом раздельного резервирования замещением с использованием ненагруженного резерва [9], общего резервирования замещением, а также постоянного общего резервирования[1]. Раздельное резервирование замещением [9] обеспечивает только удержание привода в достигнутом положении на время замены блока управления и не обеспечивает непрерывный контроль исправности резервного канала. Общее резервирование замещением обеспечивает сохранение функции управления приводом в случае выхода из строя основного канала, но также не обеспечивает непрерывный контроль исправности резервного канала. В отличие от двух предыдущих способов, использование постоянного общего резервирования обеспечивает сохранение функции управления приводом в случае выхода из строя основного канала и непрерывный контроль исправности резервного канала.

Предложен автоматный подход, позволяющий создавать блоки управления электроприводом на основе программируемых логических интегральных схем (ПЛИС) с постоянным общим резервированием, при котором основной и резервный блоки управления работают одновременно на один ШД и электрическая нагрузка распределена между блоками.

Предлагаемый подход

Структурная схема канала управления электроприводом, реализующего постоянное общее резервирование, приведена на рис. 1. Электроснабжение систем безопасности ЯЭУ должно осуществляться не менее чем от двух основных (основного и резервного) и аварийного источников[2].

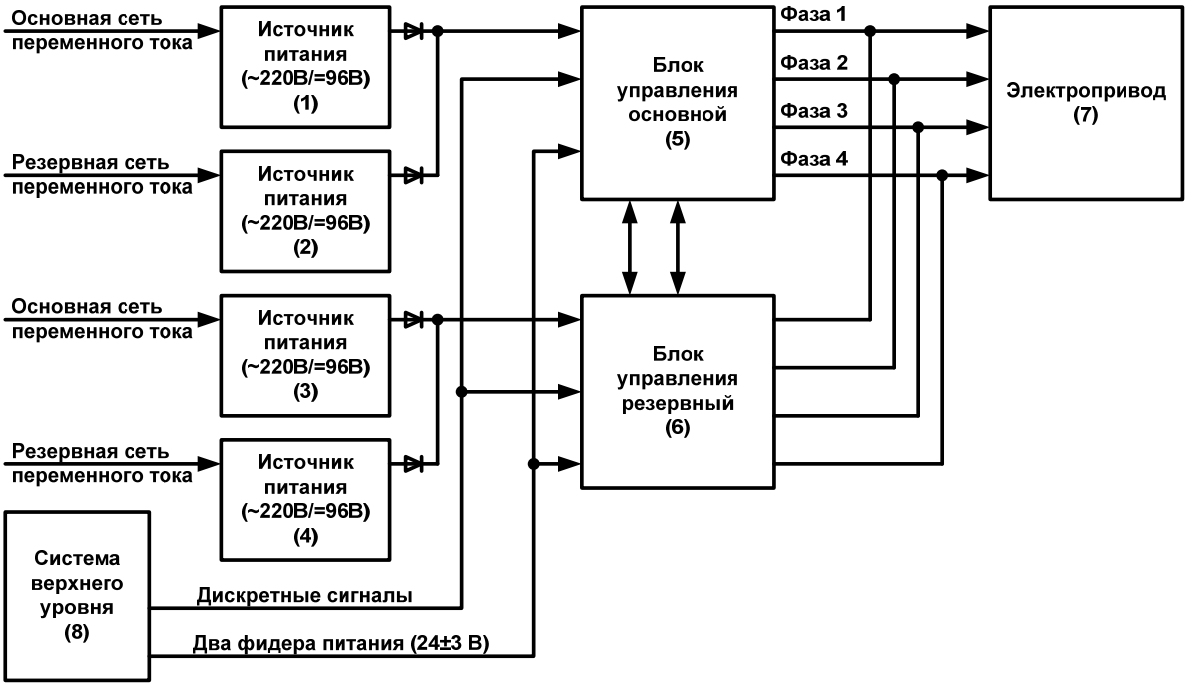


Рис. 1. Структурная схема канала управления

Источники питания (1)–(4) предназначены для питания силовых преобразователей блоков управления (5) и (6), предназначенных для формирования режимов работы электропривода (7) в соответствии с сигналами, поступающими из системы верхнего уровня (8), а также формирования токов фаз ШД и контроля исправности.

В настоящей работе блоки (1)–(4) не рассматриваются. В предложенной структурной схеме каждый источник питания должен обеспечивать возможность управления электроприводом при отсутствии остальных источников питания. Аналогично блоки управления (5) и (6) должны обеспечивать возможность управления при отсутствии соседнего блока.

[1] ГОСТ 27.002-89 Надежность в технике. Основные понятия. Термины и определения.
[2] ПБЯ В.08-88/05 «Правила ядерной безопасности корабельных ядерных энергетических установок». М., 2005. 80 с.

В отличие от структурной схемы, приведенной в [9], здесь содержится два типа блоков вместо пяти, обеспечивается сохранение возможности управления положением электропривода при выходе из строя любого одного блока, а также обеспечивается непрерывный контроль исправности основного и резервного блоков управления. Отметим, что замена вышедшего из строя блока может осуществляться без прерывания работы канала. Новый блок при этом определяет свое текущее состояние по входным дискретным сигналам и выходным сигналам соседнего блока, кроме числа пройденных шагов, что не влияет на функционирование, так как в системе присутствует датчик положения электропривода. Дискретные входные сигналы поступают из системы верхнего уровня одновременно на основной и резервный блоки управления.

Программирование ПЛИС осуществлялось с использованием подхода, предложенного в [9] и отличающегося от традиционного программирования с использованием блок-схем, языков программирования аппаратуры, языка описания конечных автоматов [10] тем, что алгоритм управления электроприводом задается в общем случае в виде системы графов переходов конечных автоматов, как это было предложено в работах [11, 12]. Текст на языке описания аппаратуры, необходимый для синтеза устройства в САПР производителя микросхем программируемой логики, формируется автоматически средствами пакета MATLAB. Упомянутый подход является развитием автоматного программирования [12] применительно к программированию аппаратуры. Вопрос о применении автоматов при создании программ управления электроприводом, но для микроконтроллеров, описан в работе [13].

Подход состоит из следующих этапов.

1. Создание схемы связей блока управления с объектом управления и системой верхнего уровня.
2. Разработка перечня и описания входных и выходных переменных.
3. Получение алгоритма работы от Заказчика (в виде словесного описания и временных диаграмм).
4. Эвристическое проектирование системы графов переходов конечных автоматов.
5. Отображение графов переходов с использованием пакета Stateflow, входящего в состав MATLAB, по методике [14].
6. Разработка модели объекта управления средствами MATLAB-Simulink.
7. Комплексное моделирование системы автоматов и объекта управления с получением временных диаграмм работы.
8. Сравнение результатов моделирования с требуемым алгоритмом работы. Если результаты моделирования не удовлетворяют требованиям Заказчика, то необходимо вернуться к этапу проектирования графов переходов.
9. Реализация системы графов переходов, представленных в Stateflow на языке описания аппаратуры (Verilog, VHDL), с помощью интерпретатора пакета HDL Coder, входящего в MATLAB.
10. Моделирование полученной программы с использованием САПР производителя ПЛИС или другой системы HDL-моделирования, например, ModelSim.
11. Компиляция и последующая загрузка в целевую аппаратуру.

Реализация канала управления электроприводом

Блоки управления (5), (6) представляют собой электронные блоки, структурная схема которых приведена на рис. 2. Они содержат по два функциональных узла, причем УФ предназначен для формирования режимов работы электропривода, ШИМ-сигналов управления силовыми преобразователями блоков (5), (6) и контроля исправности блока управления, узел усиления УУ – для усиления ШИМ-сигналов и формирования токов фаз ШД (здесь ШИМ – широтно-импульсная модуляция).

Структурная схема модели канала управления, содержащего блоки (5), (6) и нагрузку (7), приведена на рис. 3. Модель реализована в среде MATLAB-Simulink с использованием расширений SimPowerSystems, Stateflow, HDL Coder, так как средства задания графов переходов, предлагаемые производителями ПЛИС [15, 16], являются узкоспециализированными и малофункциональными. Она содержит модель, предназначенную для имитации входных сигналов блоков управления (а), модели управляющих машин состояний основного и резервного блоков (б) и (в), модель силовых преобразователей блоков управления и нагрузки (г), средства визуализации результатов моделирования (д), модель аналого-цифрового преобразователя (АЦП) сигналов датчиков токов фаз с возможностью имитации неисправности (е). Модель (г) содержит модели силовых преобразователей (СП) основного и резервного блоков СП1, СП2, модели датчиков токов (ДТ) фаз ДТ1, ДТ2, включенные последовательно.

Автоматы

Конечный автомат узла УФ содержит пять гиперсостояний UF.T1, UF.T2, UF.T3, UF.T4, UF.T5. При этом, например, гиперсостояние UF.T1 (рис. 4) содержит четыре вложенных автомата S3, S4, S9, S6, гиперсостояние UF.T2 – шесть вложенных автоматов, а гиперсостояния UF.T3, UF.T4, UF.T5 не содержат вложенных автоматов. Гиперсостояния UF.T1, UF.T3 предназначены для выбора режимов работы электропривода и формирования команд на движение, UF.T4, UF.T5 предназначены для контроля исправно-

сти блока и сравнения токов фаз с заданными уставками. Гиперсостояния UF.T1, UF.T3 управляют гиперсостоянием UF.T2, предназначенным для формирования команд на включение фаз и уставок токов фаз. Уставки поступают на пропорционально-интегральный регулятор тока, выход которого поступает на широтно-импульсный модулятор, его выход подается на коммутатор, определяющий порядок включения силовых транзисторных ключей, выполненный также в виде конечного автомата и содержащий два гиперсостояния. Таким образом, система состоит из 19 конечных автоматов.

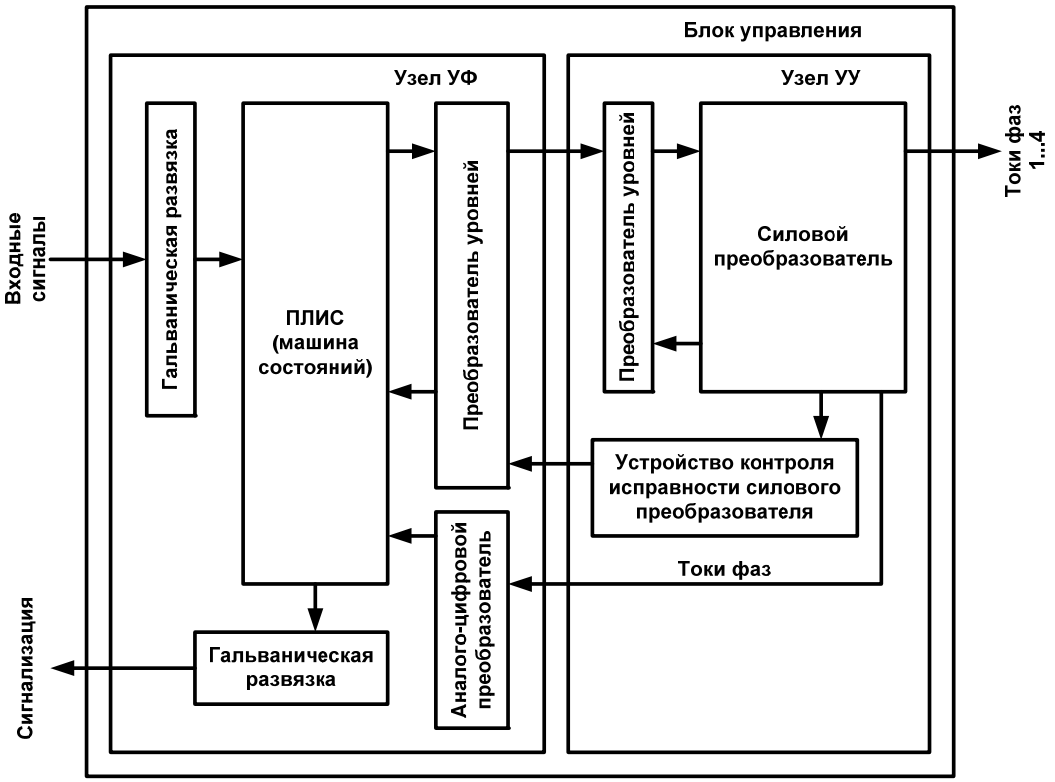


Рис. 2. Структура блока управления

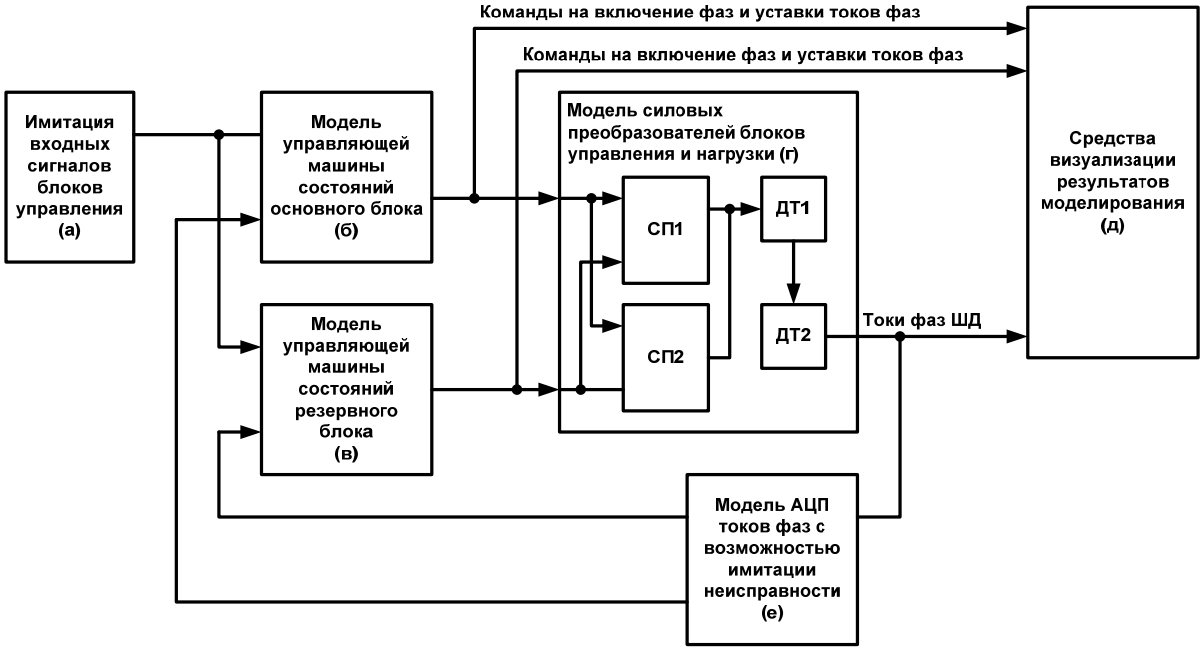


Рис. 3. Структурная схема модели канала управления

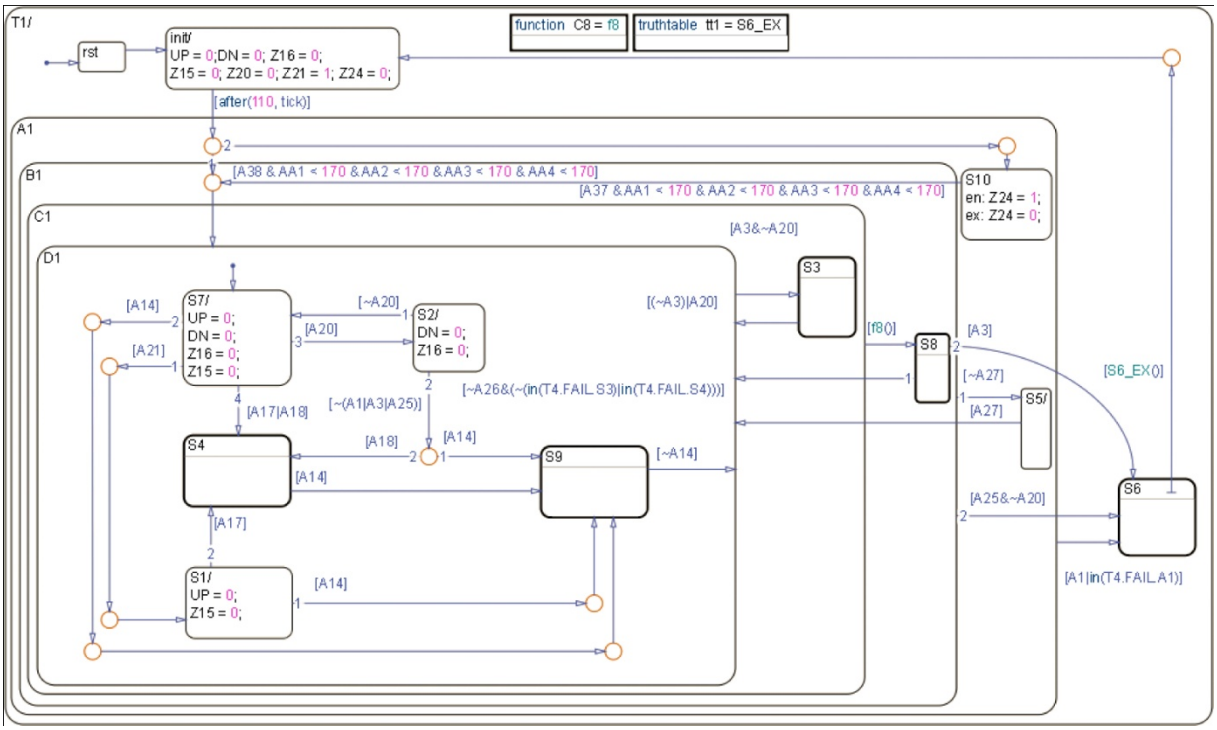
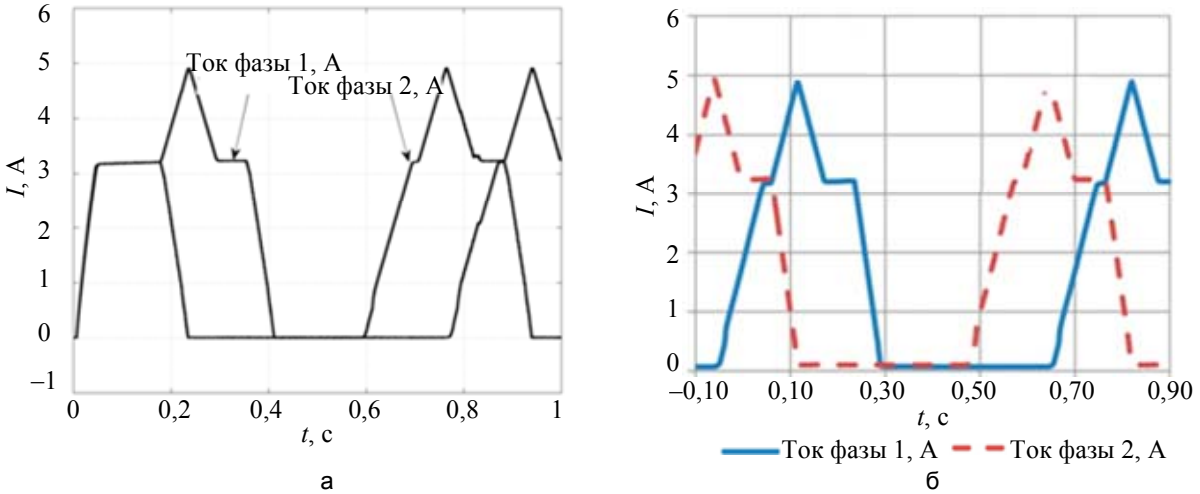


Рис. 4. Гиперсостояние UF1.T1

Результаты моделирования и испытаний опытных образцов в исправном состоянии

Результаты моделирования в режиме движения при автоматическом управлении – токи в фазах 1 и 2 ШД – приведены на рис. 5, а. Из их рассмотрения следует, что в начальный момент времени блоки были выключены и токи в фазах ШД отсутствовали, затем блоки вышли на токи удержания, равные примерно 3,2 А, в момент времени около 0,2 с блоки делают один шаг, затем на короткое время выходят в режим удержания, затем делают следующий шаг и переходят на фазы 3 и 4 (на рисунке не показаны). В момент времени, соответствующий 0,7 с, блоки делают очередной шаг и выходят на ток удержания, затем в момент времени около 0,9 с блоки делают еще один шаг. Такой режим работы определяется входными сигналами. По излому на начальном участке нарастания кривой силы тока (на уровне 0,5 А) можно видеть момент включения резервного блока.



а б

Рис. 5. Результаты моделирования (а) и испытаний (б)

Для проверки результатов были разработаны и изготовлены два образца блока управления, два образца источников вторичного электропитания и макет прибора. Образцы блоков были выполнены в соответствии с требованиями, предъявляемыми к поставляемой аппаратуре.

На рис. 5, б, показаны результаты испытаний макетных образцов блоков управления. Испытания проводились в условиях стенда, блоки были установлены в макет прибора, использовались макетные образцы штатных источников вторичного электропитания. В качестве нагрузки для блоков был использован

штатный электропривод с нагрузочным устройством, обеспечивающим номинальную нагрузку на валу двигателя. В процессе испытаний измерения силы токов фаз проводились с использованием двух токовых пробников типа Agilent N2782A и осциллографа Agilent MSO6032A.

Из рассмотрения результатов испытаний следует, что вначале блоки делали шаг с кратковременным выходом на ток удержания, затем в момент времени около 0,1 с блоки делают следующий шаг, затем еще один и переходят на фазы 3 и 4 ШД, затем в моменты времени около 0,6 с и 0,8 с блоки делают еще два шага. Из сравнения графиков на рис. 5, а, б, следует, что результаты моделирования и испытаний практически совпадают.

Моделирование и испытания блоков при имитации неисправности основного блока

На рис. 6, а, показаны кривые сил токов фаз 1 и 2 при моделировании имитации неисправности основного блока и исправном резервном блоке. В начальный момент времени блоки исправны и работают параллельно на один ШД. В момент времени 0,11 с происходит выключение ведущего блока и токи начинают спадать. В момент времени 0,135 с резервный блок регистрирует спад силы токов фаз ниже порогового значения, а также признак неисправности основного, переходит в роль основного и подает напряжение на обмотку ШД.

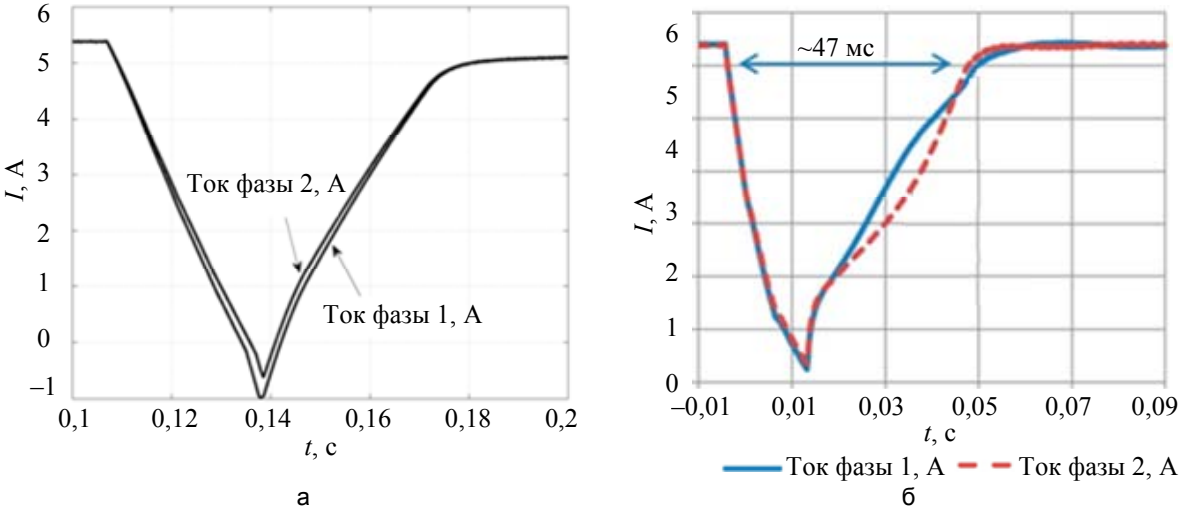


Рис. 6. Результаты моделирования (а) и испытаний (б) при имитации неисправности основного блока

Силы токов фаз начинают увеличиваться до уставки силы тока удержания и устанавливаются равными 3,06 А. Время переходного процесса от уровня 2,9 А при спаде силы токов до уровня 2,9 А при нарастании силы токов составляет примерно 65 мс.

На рис. 6, б, показаны результаты испытаний блоков управления при имитации неисправности основного блока. Из сравнения рис. 6, а, б, следует, что результаты моделирования и испытаний практически совпадают. Время переходного процесса на рис. 6, б, от уровня 2,9 А при спаде силы токов до уровня 2,9 А при нарастании силы токов составляет примерно 47 мс. Такая разность длительности переходного процесса объясняется отличием фактических параметров обмотки двигателя от параметров, использованных при моделировании.

Испытания подтвердили, что за время включения резервного блока привод не успевает сдвинуться с места более чем на один шаг.

Также проводились испытания с целью проверки параллельной работы блоков в режиме удержания, движения вверх или вниз; сохранения возможности управления при выходе из строя основного или резервного блоков; возможности замены неисправного блока без прерывания работы канала. Все испытания были успешно пройдены.

Заключение

Реализация постоянного общего резервирования позволила обеспечить резервирование функции регулирования положения компенсирующей группы, непрерывный контроль исправности основного и резервного блоков управления, облегченные режимы работы полупроводниковых приборов силовых преобразователей блоков.

Несмотря на большое число конечных автоматов в системе, применение автоматного подхода по сравнению с традиционным подходом позволило существенно сократить время отладки программного обеспечения блоков благодаря: наглядности и иерархичности графов переходов, возможности пошаговой отладки автомата, визуализации системы автоматов, визуализации переходов внутри автомата, переходов во вложенные состояния, возврата в состояние верхнего уровня, совместному моделированию управ-

ляющего автомата и силового полупроводникового преобразователя, реализации прямого цифрового управления силовыми транзисторными ключами автоматным способом.

Литература

1. Jahns T.M. Improved reliability in solid state AC drives by means of multiple independent phase drive units // IEEE Transactions on Industry Applications. 1980. V. 1 A-16. N 3. P. 321–331.
2. Welchko B.A., Lipo T.A., Jahns T.M., Schulz S.E. Fault tolerant three-phase AC motor drive topologies: a comparison of features, cost, and limitations // IEEE Transaction on Power Electronics. 2004. V. 19. N 4. P. 1108–1116.
3. Ertugrul N., Soong W., Dostal G., Saxon D. Fault tolerant motor drive system with redundancy for critical applications // PESC Record – IEEE Annual Power Electronics Specialists Conference. 2002. V. 3. P. 1457–1462.
4. Hopper T., Anders M., Stuckmann C. Building electric motors for space, with redundancy and high reliability // Proc. 14th European Space Mechanics and Tribology Symposium, ESMATS2011. Constance, Germany, 2011. P. 373–378.
5. Estima J.O., Cardoso A.J.M. Fast fault detection, isolation and reconfiguration in fault-tolerant permanent magnet synchronous motor drives // IEEE Energy Conversion Congress and Exposition, ECCE 2012. Raleigh, USA, 2012. Art. 6342310. P. 3617–3624.
6. Mecrow B.C., Jack A.G., Haylock J.A., Coles J. Fault-tolerant permanent magnet machine drives // IEE Proceedings: Electric Power Applications. 1996. V. 143. N 6. P. 437–442.
7. Heo H.-J., Im W.-S., Kim J.-M., Kim Y.-G., Oh J.-S. Fault tolerant control methods of dual type independent multi-phase BLDC motor under open-switch fault conditions // IEEE Applied Power Electronics Conference and Exposition - APEC. 2012. Art. 6166032. P. 1591–1596.
8. Hong G., Wei W., Wei X., Yanming L. Design of electrical/mechanical hybrid 4-redundancy brushless DC torque motor // Chinese Journal of Aeronautics. 2010. V. 23. N 2. P. 211–215.
9. Янкин Ю.Ю., Шалыто А.А. Автоматное программирование ПЛИС в задачах управления электроприводом // Информационно-управляющие системы. 2011. № 1. С. 50–56.
10. Пимкин А. Транслятор описания конечного автомата в исходный код на языке описания аппаратуры Verilog [Электронный ресурс]. Режим доступа: http://ded32.ru/abnl/?adsdata=AsJDvjngrvsPp;cI8YiPERzKnyXh4Nnpq3K948B5eKN;Yjt!my4de!4H0cMGF4x0LuVf94xtqBhsFDyZEsjHKJU4Yvu64u5r8nH2xz3Gjgoo, свободный. Яз. рус. (дата обращения 16.03.14).
11. Harel D. Statecharts: a visual formalism for complex systems // Science of Computer Programming. 1987. V. 8. N 3. P. 231–274.
12. Поликарпова Н.И., Шалыто А.А. Автоматное программирование. СПб: Питер, 2009. 176 с.
13. Козаченко В.Ф. Эффективный метод программной реализации дискретных управляющих автоматов во встроенных системах управления [Электронный ресурс]. Режим доступа: http://www.motorcontrol.ru/publications/state_mashine.pdf, свободный. Яз. рус. (дата обращения 10.08.2014).
14. Stateflow Getting Started Guide / R2014a MathWorks Documentation. [Электронный ресурс]. Режим доступа: http://www.mathworks.com/help/pdf_doc/stateflow/sf_gs.pdf, свободный. Яз. англ. (дата обращения 24.09.2014).
15. About the State Machine Editor/Quartus II Help [Электронный ресурс]. Режим доступа: http://quartushelp.altera.com/current/master.htm#mergedProjects/verify/rtl/rtl_view_sme.htm, свободный. Яз. англ. (дата обращения 14.03.14).
16. StateCad Help [Электронный ресурс]. Режим доступа: http://www.xilinx.com/support/documentation/sw_manuals/xilinx10/help/iseguide/mergedProjects/state/whnjs.htm, свободный. Яз. англ. (дата обращения 14.03.14).

Янкин Юрий Юрьевич – ведущий инженер, ОАО «Концерн «НПО «Аврора», Санкт-Петербург, 194021, Российская Федерация, yankinyy@gmail.com

Шалыто Анатолий Абрамович – доктор технических наук, профессор, заведующий кафедрой, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, shalyto@mail.ifmo.ru

Yuri Yu. Yankin – leading engineer, Concern "AURORA", Scientific and Production Association, Joint Stock Company ("Aurora", JSC), Saint Petersburg, 194021, Russian Federation, yankinyy@gmail.com

Anatoly A. Shalyto – D.Sc., Professor, Department head, ITMO University, Saint Petersburg, 197101, Russian Federation, shalyto@mail.ifmo.ru

Принято к печати 10.04.14
Accepted 10.04.14

7 МАТЕМАТИЧЕСКОЕ И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ
MODELING AND SIMULATION

УДК 532.529

ПРИМЕНЕНИЕ И РЕАЛИЗАЦИЯ РАЗНОСТНЫХ СХЕМ ВЫСОКОЙ РАЗРЕШАЮЩЕЙ СПОСОБНОСТИ ДЛЯ РЕШЕНИЯ ЗАДАЧ ГАЗОВОЙ ДИНАМИКИ НА НЕСТРУКТУРИРОВАННЫХ СЕТКАХ

К.Н. Волков[a,b]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[b] Университет Кингстона, Лондон, SW15 3DW, Великобритания, k.volkov@kingston.ac.uk

Аннотация. Разрабатывается подход к дискретизации нестационарных уравнений Навье–Стокса на неструктурированных сетках в рамках метода конечных объемов, обсуждаются его преимущества и перспективы развития. Рассматриваются особенности дискретизации невязких и вязких потоков, а также производных по времени. К преимуществам предлагаемого подхода относятся: возможность работы как на структурированных, так и неструктурированных сетках; использование разностных схем высокого порядка по времени и по пространственным координатам; выбор для дискретизации законов сохранения среднемедианного контрольного объема; применение соотношений для расчета градиента и псевдолапласиана, позволяющих получить более точные результаты на сильно растянутых сетках в пограничном слое; запись соотношений для расчета потоков через грани внутренних и граничных контрольных объемов в одинаковой форме, что обеспечивает более простую программную реализацию. Подход позволяет реализовать стратегию адаптации сетки в соответствии с особенностями конкретного течения, а также дает обширные возможности для параллелизации процессов вычислений. Возможности разработанного подхода демонстрируются на примере решения задачи, связанной с моделированием нестационарных течений в элементах газотурбинных двигателей.

Ключевые слова: газовая динамика, неструктурированная сетка, разностная схема, профиль.

APPLICATION AND IMPLEMENTATION OF HIGH-RESOLUTION DIFFERENCE SCHEMES FOR SOLUTION OF GAS DYNAMICS PROBLEMS ON UNSTRUCTURED MESHES

K.N. Volkov[a,b]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
[b] Kingston University, London, SW15 3DW, United Kingdom, k.volkov@kingston.ac.uk

Abstract. The paper deals with an approach to finite volume discretization of unsteady Navier-Stokes equations on unstructured meshes, and its advantages and development prospects are discussed. Features of inviscid and viscous flux discretization and temporal derivatives are considered. The advantages of the proposed approach include: the ability to operate on both structured and unstructured meshes; usage of high-order finite difference schemes in time and space; selection of median control volume for discretization of governing equations; application of expressions for calculation of the gradient and pseudo-laplasian making it possible to obtain more accurate results on highly stretched meshes in the boundary layer; writing of equations for the calculation of fluxes through the faces of interior and boundary control volumes in the same form, that simplify software implementation. This approach gives the possibility to implement a strategy of mesh adaptation taking into account the features of the certain flow and gives wide opportunities to parallelize computations. Possibilities of the developed approach are demonstrated on the example of the problem solution related to simulation of unsteady flows in the gas turbine engines.

Keywords: fluid dynamics, unstructured mesh, finite-difference scheme, aerofoil.

Введение

Развитие вычислительной газовой динамики и компьютерной техники делает возможным разработку и реализацию методов расчета нестационарных течений жидкости и газа в пространственных областях сложной конфигурации [1]. Традиционно при решении задач газовой динамики применяются регулярные сетки (структурированные сетки с четырехугольными ячейками на поверхности и шестигранными в пространстве). Сетка представляет собой упорядоченную по определенным правилам структуру данных с выраженными сеточными направлениями (в общем случае имеется криволинейная система координат). В преобразованном (вычислительном) пространстве ячейки сетки являются топологическими прямоугольниками (двумерные задачи) или параллелепипедами (трехмерные задачи).

Для структурированных сеток сравнительно легко реализуются вычислительные алгоритмы на основе современных монотонных методов высокого порядка точности. Однако диапазон геометрических объектов, описываемых структурированными сетками, ограничен. Как правило, невозможно построить единую сетку для всей расчетной области, в связи с чем производится разделение поля течения на подобласти, в каждой из которых генерируется своя сетка регулярной структуры. Блочный подход предоставляет широкие возможности для использования эффективных численных методов внутри отдельных бло-

ков. Основной его недостаток состоит в достаточно сложной процедуре сшивки решений, полученных в различных подобластях [2].

Характерной особенностью неструктурированных сеток является произвольное расположение узлов сетки в физической области (отсутствуют выраженные сеточные направления, нет структуры сетки, подобной регулярным сеткам). Число ячеек, содержащих каждый конкретный узел, может изменяться от узла к узлу. Узлы сетки объединяются в многоугольники (двумерный случай) или в многогранники (трехмерный случай). Как правило, на плоскости используются треугольные и четырехугольные ячейки, а в пространстве – тетраэдры и призмы. Основное преимущество неструктурированных сеток перед регулярными состоит в большей гибкости при дискретизации физической области сложной формы, а также в возможности полной автоматизации их построения. Для неструктурированных сеток легче реализуются локальные сгущения и адаптация сетки в зависимости от поведения решения.

В отличие от хорошо разработанных технологий метода конечных элементов, конечно-объемные технологии на неструктурированных сетках характеризуются отсутствием единых принципов, позволяющих провести дискретизацию конвективных и диффузионных потоков, источниковых членов, а также учет граничных условий [3]. Достаточно часто способы дискретизации, имеющие различные характеристики, объединяются [4].

В настоящей работе разрабатывается подход к дискретизации нестационарных уравнений Навье–Стокса на неструктурированных сетках в рамках метода конечных объемов. Расчетная сетка строится при помощи одного из коммерческих пакетов, таких как Gambit или ICEM CFD. Разработанные программные средства используют трансляцию сетки из формата сеточного генератора в формат общедоступной библиотеки ADF Software Library (Advanced Data Format), которая является частью библиотеки CGNS (CFD General Notation System), разработанной для внутреннего использования в корпорации Boeing и получившей широкое распространение в NASA и компании McDonnel Douglas Aerospace. Возможности разработанного подхода демонстрируются на примере решения задачи, связанной с обтеканием решетки профилей.

Метод конечных объемов

В консервативных переменных уравнение, описывающее нестационарное трехмерное течение вязкого сжимаемого газа, записывается как

$$\frac{\partial Q}{\partial t} + \nabla \cdot F(\mathbf{n}, Q, \nabla Q) = H(Q, \nabla Q). \quad (1)$$

Здесь $Q(\mathbf{x},t)$, $F(\mathbf{n}, Q, \nabla Q)$, $H(Q, \nabla Q)$ представляют собой вектор консервативных переменных в точке $\mathbf{x}$ в момент времени t, вектор потока через поверхность, ориентация которой задается внешней единичной нормалью $\mathbf{n}$, и источниковый член соответственно. При моделировании турбулентных течений уравнение (1) дополняется уравнениями модели турбулентности, а вместо молекулярных коэффициентов переноса используются их эффективные значения.

Вектор потока расщепляется на невязкую (индекс I) и вязкую (индекс V) составляющие:

$$F(\mathbf{n}, Q, \nabla Q) = F^I(\mathbf{n}, Q) + F^V(\mathbf{n}, Q, \nabla Q), \quad (2)$$

Введем вектор невязки

$$R(Q) = \nabla \cdot F(\mathbf{n}, Q, \nabla Q) - H(Q, \nabla Q).$$

Уравнение (2) примет следующий вид:

$$\frac{\partial Q}{\partial t} + R(Q) = 0. \quad (3)$$

Для дискретизации уравнений, записанных в виде (1), (2) или (3), используется метод конечных объемов на гибридной сетке.

Интегрируя уравнение (1) по контрольному объему V_i с границей ∂V_i, ориентация которой задается внешней единичной нормалью $\mathbf{n} = \{n_x, n_y, n_z\}$, и применяя теорему Гаусса–Остроградского, получим:

$$\frac{\partial}{\partial t}\int_{V_i} Q d\Omega + \oint_{\partial V_i} \left[F(\mathbf{n}, Q, \nabla Q) - (\mathbf{v}_b \cdot \mathbf{n}) Q \right] dS = \int_{V_i} H(Q, \nabla Q) d\Omega. \quad (4)$$

Преобразуем уравнение (4) к виду

$$\frac{\partial Q_i}{\partial t} + R_i(Q) = 0. \quad (5)$$

Вектор невязки в уравнении (5) находится из соотношения

$$R_i(Q) = \frac{1}{V_i}\left\{ \oint_{\partial V_i} \left[F(\mathbf{n}, Q, \nabla Q) - (\mathbf{v}_b \cdot \mathbf{n}) Q \right] dS - \int_{V_i} H(Q, \nabla Q) d\Omega \right\}. \quad (6)$$

Под $\mathbf{v}_b = \{u_b, v_b, w_b\}$ понимается скорость перемещения границы ∂V_i контрольного объема V_i.

Среднемедианный контрольный объем V_i, связанный с узлом i=1,…, N гибридной сетки, где N – число узлов, строится таким образом, что геометрические центры ячеек сетки с вершиной в узле i соединяются друг с другом через середины разделяющих их граней. Пример контрольного объема показан на рис. 1.

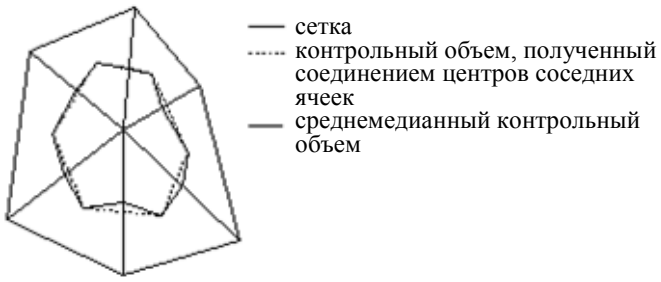


Рис. 1. Среднемедианный контрольный объем

Весовые множители (площади граней) внутренних граней контрольного объема являются антисимметричными, $\Delta s_{ij} = -\Delta s_{ji}$ для $\forall j \in E_i$, а весовые множители его граничных граней – симметричными, $\Delta s_{ik} = \Delta s_{ki}$ для $\forall k \in B_i$ [1]. При этом имеет место следующее соотношение:

$$\sum_{j\in E_i} \mathbf{n}_{ij}\Delta s_{ij} + \sum_{k\in B_i} \mathbf{n}_{ik}\Delta s_{ik} = 0\,. \tag{7}$$

Здесь E_i – множество внутренних граней, связанных с узлом i; B_i – множество граничных граней, связанных с узлом i; $\mathbf{n}_{ij}$ – внешняя единичная нормаль, задающая ориентацию грани (i, j); Δs_{ij} – площадь грани, соединяющей узлы i и j; $\mathbf{n}_{ik}$ – внешняя единичная нормаль к граничной грани (i, k); Δs_{ik} – площадь граничной грани, соединяющей узлы i и k.

Расчет потоков

Интеграл от потока в соотношении (6) разделяется на два слагаемых, связанных с внутренними и граничными гранями контрольного объема:

$$\oint_{\partial V_i} F(\mathbf{n}, Q, \nabla Q)dS = \sum_{j\in E_i} F(\mathbf{n}_{ij}, Q, \nabla Q)\big|_{\mathbf{x}=\frac{1}{2}(\mathbf{x}_i+\mathbf{x}_j)}\ \mathbf{n}_{ij}\Delta s_{ij} + \sum_{k\in B_i} F(\mathbf{n}_{ik}, Q, \nabla Q)\big|_{\mathbf{x}=\mathbf{x}_{ik}}\ \mathbf{n}_{ik}\Delta s_{ik}\,. \tag{8}$$

С учетом (8) соотношение (6) примет следующий дискретный вид:

$$R_i = \frac{1}{V_i}\left(\sum_{j\in E_i} F_{ij}\mathbf{n}_{ij}\Delta s_{ij} + \sum_{k\in B_i} F_{ik}\mathbf{n}_{ik}\Delta s_{ik} - H_i V_i\right). \tag{9}$$

Здесь F_{ij} – поток через внутреннюю грань (i, j), ориентация которой задается внешней единичной нормалью $\mathbf{n}_{ij}$; F_{ik} – поток через граничную грань (i, k), ориентация которой задается внешней единичной нормалью $\mathbf{n}_{ik}$; Δs_{ij} и Δs_{ik} – площади внутренней (i, j) и граничной (i, k) граней контрольного объема (рис. 2).

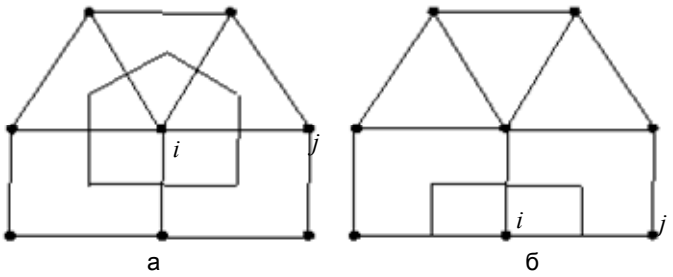


Рис. 2. Внутренний (а) и граничный (б) контрольные объемы

Внутренние грани. Поток через внутреннюю грань (i, j) контрольного объема вычисляется в серединной точке грани,

$$\mathbf{x}_{ij} = \frac{1}{2}\left(\mathbf{x}_i + \mathbf{x}_j\right),$$

как полусумма соответствующих узловых значений, умноженных на площадь грани:

$$F_{ij} = \frac{1}{2}\left(F_i + F_j\right)\ \mathbf{n}_{ij}\Delta s_{ij}\,.$$

Суммируя по всем внутренним граням (для $\forall j \in E_i$), получим

$$F = \frac{1}{2}\sum_{j\in E_i}\left(F_i + F_j\right)\,\mathbf{n}_{ij}\Delta s_{ij}\ .$$

С учетом соотношения (7) для замкнутого контрольного объема можно записать:

$$F = \frac{1}{2}\sum_{j\in E_i}\left(F_i + F_j\right)\,\mathbf{n}_{ij}\Delta s_{ij} - F_i\sum_{j\in E_i}\mathbf{n}_{ij}\Delta s_{ij}\ .$$

Тогда

$$F = \frac{1}{2}\sum_{j\in E_i}\left(F_j - F_i\right)\,\mathbf{n}_{ij}\Delta s_{ij}\ .$$

Учитывая, что $\Delta s_{ij} = -\Delta s_{ji}$ для $\forall j \in E_i$, вклад каждой грани (i, j) контрольного объема представляется в виде

$$F_{ij} = \frac{1}{2}\left(F_j - F_i\right)\,\mathbf{n}_{ij}\Delta s_{ij}\ .$$

Граничные грани. Поток через граничную грань (i, k) контрольного объема рассчитывается в точке

$$\mathbf{x}_{ik} = \frac{1}{2D+2}\sum_{j\in G_k}\left[1+(D+2)\delta_{jk}\right]\,\mathbf{x}_j\ ,$$

где D – размерность задачи; G_k – множество узлов в граничной ячейке k, δ_{jk} – символ Кронекера. Вклад грани, лежащей на границе области, записывается отдельно. Учитывая соотношение (7), имеем

$$F = \frac{1}{2}\sum_{j\in E_i}\left(F_i + F_j\right)\,\mathbf{n}_{ij}\Delta s_{ij} + F_{ik}\mathbf{n}_{ik}\Delta s_{ik} - F_{ik}\left(\mathbf{n}_{ik}\Delta s_{ik} + \sum_{j\in E_i}\mathbf{n}_{ij}\Delta s_{ij}\right).$$

В результате для расчета потоков через грани граничного контрольного объема получим такое же соотношение, что и для внутренних граней:

$$F = \frac{1}{2}\sum_{j\in E_i}\left(F_j - F_i\right)\,\mathbf{n}_{ij}\Delta s_{ij}\ .$$

Использование одинаковых выражений для расчета потоков через грани внутренних и граничных контрольных объемов обеспечивает более простую программную реализацию.

Невязкие потоки

Из соотношений (6) и (9) имеем

$$R_i^I(Q) = \frac{1}{V_i}\oint_{\partial V_i}F^I(\mathbf{n},Q)dS = \frac{1}{V_i}\left(\sum_{j\in E_i}F_{ij}^I\mathbf{n}_{ij}\Delta s_{ij} + \sum_{k\in B_i}F_{ik}^I\mathbf{n}_{ik}\Delta s_{ik}\right). \qquad (10)$$

Для дискретизации невязких потоков в (10) используется модифицированный вариант схемы MUSCL, которая представляет собой комбинацию центрированных конечных разностей 2-го и 4-го порядка, для переключения между которыми служит сглаживатель потока, построенный на основе характеристических переменных.

Разностная схема. Рассмотрим уравнение

$$\frac{\partial Q}{\partial t} + A\nabla Q = 0\ . \qquad (11)$$

Якобиан находится из соотношения

$$A = \frac{\partial F^I}{\partial Q} = \frac{\partial F_x^I}{\partial Q} + \frac{\partial F_y^I}{\partial Q} + \frac{\partial F_z^I}{\partial Q}\ .$$

Схема MUSCL для уравнения (11) записывается в виде

$$F_{ij}^I = \frac{1}{2}\left[F_{ij}^I(Q_j) + F_{ij}^I(Q_i) - \left|A_{ij}\right|\left(Q^+ - Q^-\right)\right]. \qquad (12)$$

Соотношение (12) представляет собой комбинацию центральной разностной производной 2-го порядка и диссипативного члена. Диссипативное слагаемое представляется в виде [5]

$$\left|A_{ij}\right|\left(Q^+ - Q^-\right) = \frac{1}{2}(1-\kappa)\ \left|A_{ij}\right|\ \left[\left(\frac{1}{2}Q_{j+} - Q_j + \frac{1}{2}Q_i\right) - \left(\frac{1}{2}Q_j - Q_i + \frac{1}{2}Q_{i-}\right)\right], \qquad (13)$$

где $\kappa \in [0,1]$, а Q_{j+}, Q_j, Q_i, Q_{i-} соответствуют точкам $\mathbf{x}_{j+}$, $\mathbf{x}_j$, $\mathbf{x}_i$, $\mathbf{x}_{i-}$, лежащим на равном расстоянии друг от друга. Соотношения (12) и (13) дают следующую разностную схему для расчета потока:

$$F_{ij}^I = \frac{1}{2}\left\{\left[F_{ij}^I(Q_j) + F_{ij}^I(Q_i)\right] - \frac{1}{2}(1-\kappa)\ \left|A_{ij}\right|\ \left(L_j(Q) - L_i(Q)\right)\right\}, \qquad (14)$$

где $L(Q)$ – псевдолапласиан. Для сохранения монотонности решения в схему (14) вводится ограничитель потока [6]. После линеаризации потоков получим

$$F_{ij}^I = \frac{1}{2}\left\{\left[A_{ij}Q_j + A_{ij}Q_i\right] - \left|A_{ij}\right|\left[\frac{1}{3}(1-\varphi)\left(\hat{L}_j^*(Q) - \hat{L}_i^*(Q)\right) + \varphi\left(Q_j - Q_i\right)\right]\right\}, \quad (15)$$

где $\hat{L}^*(Q)$ – модифицированный псевдолапласиан. Полагается, что к=1/3. Соотношение (15) представляет собой комбинацию конечных разностей 2-го и 4-го порядка точности. Для переключения между ними служит функция (сглаживатель потока) [6, 7].

На твердой стенке, вследствие условия непротекания, $F_{ik}^I = 0$ для $\forall k \in B_i$. На входной границе расчетной области полагается

$$F_{ik}^I = \frac{1}{2}\left[F_{ik}^I(Q_k) + F_{ik}^I(Q_\infty) - \left|A_{ik}\right|\left(Q_k - Q_\infty\right)\right],$$

где Q_∞ – консервативные переменные на бесконечности.

Расчет псевдолапласиана. Псевдолапласиан представляет собой обобщение центрированной разностной производной 2-го порядка на неструктурированную сетку:

$$L_i(Q) = \frac{1}{|E_i|}\sum_{j\in E_i}\left(Q_j - Q_i\right), \quad (16)$$

где $|E_i|$ – число элементов множества E_i. С использованием (16) оценки показывают, что

$$L_i(Q) \sim O(h^2)\nabla^2 Q\,|_{\mathbf{x}=\mathbf{x}_i},$$

поэтому диссипативный член в (14) имеет порядок $O(h^3)$. После подстановки в соотношение (15) и деления на объем в (9) получается погрешность порядка $O(h^2)$.

Разложение в ряд Тейлора в окрестности точки $\mathbf{x}_i$ дает

$$L_i(Q) = L_i(\mathbf{x})\nabla Q\,|_{\mathbf{x}=\mathbf{x}_i},$$

где $L_i(\mathbf{x}) = \{L_i x, L_i y, L_i z\}'$. Следовательно, на неравномерной сетке схема не имеет 2-го порядка точности. Если Q является линейной функцией пространственных координат, то $L_i(Q) \neq 0$. Указанные обстоятельства приводят к потере точности решения, в связи с чем псевдолапласиан переопределяется следующим образом [9]:

$$L_i^*(Q) = L_i(Q) - \nabla Q_i L_i(\mathbf{x}). \quad (17)$$

Если $Q = a\mathbf{x}+c$, то $L_i(Q) = aL_i(\mathbf{x})$, $\nabla Q = a$, поэтому $L_i^*(Q) = 0$. Вместе с тем, представление псевдолапласиана в виде (17) допускает потерю устойчивости численного решения и дает неточные результаты на сильно растянутых сетках, используемых для расчета течения в пограничном слое. Для обеспечения устойчивости решения вводится оператор масштабирования:

$$\hat{L}_i(Q) = \left(\sum_{j\in E_i}\frac{1}{\left|\mathbf{x}_j - \mathbf{x}_i\right|}\right)^{-1}\sum_{j\in E_i}\frac{Q_j - Q_i}{\left|\mathbf{x}_j - \mathbf{x}_i\right|}; \qquad \hat{L}_i(\mathbf{x}) = \left(\sum_{j\in E_i}\frac{1}{\left|\mathbf{x}_j - \mathbf{x}_i\right|}\right)^{-1}\sum_{j\in E_i}\frac{\mathbf{x}_j - \mathbf{x}_i}{\left|\mathbf{x}_j - \mathbf{x}_i\right|}.$$

После этого модифицированный псевдолапласиан находится из соотношения

$$\hat{L}_i^*(Q) = \hat{L}_i(Q) - \nabla Q_i \hat{L}_i(\mathbf{x}). \quad (18)$$

Псевдолапласиан в виде (18) дает ноль на линейной функции и позволяет обеспечить точные результаты на сильно растянутых сетках. Недостаток представления псевдолапласиана в виде (18) связан с анизотропныи масштабированием, что приводит к демпфированию высокочастотных гармоник только в направлении наивысшего сеточного разрешения (поперек пограничного слоя).

Расчет градиента. Рассмотрим контрольный объем *abcde*, связанный с узлом *i* неструктурированной сетки (рис. 3). С учетом того, что $Q_i = \text{const}$ и $\nabla Q_i = 0$, используя формулы Грина и теорему Стокса, получим

$$\nabla Q = \frac{1}{\Omega}\left(\int_\Omega \nabla Q d\omega - \int_\Omega \nabla Q_i d\omega\right) = \frac{1}{\Omega}\left(\oint_L Q dl - \oint_L Q_i dl\right). \quad (19)$$

Здесь Ω – расширенный контрольный объем 12345, участвующий в вычислении градиента; L – площадь граней расширенного контрольного объема; V – исходный контрольный объем *abcde*; S – площадь граней контрольного объема. Интегралы, входящие в (19), представляются в виде

$$\oint_L Q dl = \sum_{1,2,3,4,5}\frac{Q_i + Q_j}{2}\,\mathbf{n}_{ij}, \qquad \oint_L Q_i dl = Q_i L.$$

При этом имеют место следующие соотношения:

$$S = S_{ab} + S_{bc} + S_{cd} + S_{de} + S_{ea};$$

$L = L_{12} + L_{23} + L_{34} + L_{45} + L_{51} = 3 \times S$;
$\Omega = 3 \times V$.

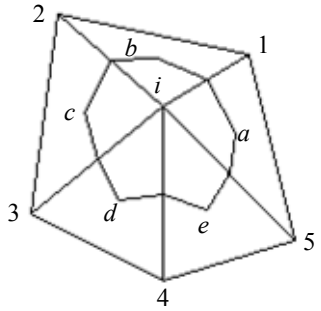


Рис. 3. Вычисление градиента

Нормали к граням контрольного объема находятся из соотношений

$\mathbf{n}_{ab} = \mathbf{n}_{51} + \mathbf{n}_{12}$; $\mathbf{n}_{bc} = \mathbf{n}_{12} + \mathbf{n}_{23}$; $\mathbf{n}_{cd} = \mathbf{n}_{23} + \mathbf{n}_{34}$;
$\mathbf{n}_{de} = \mathbf{n}_{34} + \mathbf{n}_{45}$; $\mathbf{n}_{ea} = \mathbf{n}_{45} + \mathbf{n}_{51}$.

Используя формулы Грина и учитывая (7), получим

$$(\nabla Q)_i = \frac{1}{V_i}\left[\sum_{j \in E_i} \frac{1}{2}\left(Q_j - Q_i\right)\ \mathbf{n}_{ij} \Delta s_{ij}\right]. \qquad (20)$$

Вязкие потоки

Из соотношений (6) и (9) имеем

$$R_i^V(Q) = \frac{1}{V_i}\oint_{\partial V_i} F^V(\mathbf{n}, Q, \nabla Q) dS = \frac{1}{V_i}\left(\sum_{j \in E_i} F_{ij}^V \mathbf{n}_{ij} \Delta s_{ij} + \sum_{k \in B_i} F_{ik}^V \mathbf{n}_{ik} \Delta s_{ik}\right). \qquad (21)$$

Учитывая, что вследствие условий прилипания и непротекания на стенке $F_{ik}^V = 0$ для $\forall k \in B_i$, и пренебрегая вязкими силами на входной границе расчетной области, соотношение (21) можно переписать в следующем виде:

$$R_i^V = \frac{1}{V_i}\sum_{j \in E_i} F_{ij}^V \mathbf{n}_{ij} \Delta s_{ij}. \qquad (22)$$

Соотношение (22) используется для всех контрольных объемов, включая граничные.
После линеаризации соотношения (22) получим

$$R_i^V = \frac{1}{V_i}\left[\sum_{j \in E_i} \frac{\partial F_{ij}^V}{\partial(\nabla Q_{ij})}\ \mathbf{n}_{ij} \Delta s_{ij}\right], \qquad \frac{\partial F_{ij}^V}{\partial(\nabla Q_{ij})} = BM^{-1}\frac{\partial Q}{\partial l}. \qquad (23)$$

где M – матрица перехода от консервативных переменных к примитивным, B – вязкий якобиан. Соотношение (23) приобретает вид

$$R_i^V = \frac{1}{V_i}\sum_{j \in E_i} \frac{Q_j - Q_i}{\left|\mathbf{x}_j - \mathbf{x}_i\right|}\ BM^{-1}\ \mathbf{n}_{ij} \Delta s_{ij}.$$

Для расчета градиента ∇Q в серединной точке каждой грани $\mathbf{x}_{ij} = (\mathbf{x}_i + \mathbf{x}_j)/2$ в соотношении (23) используется полусумма соответствующих узловых значений. Для расчета градиентов $(\nabla Q)_i$ и $(\nabla Q)_j$ в узлах сетки применяется соотношение (20). Однако среднее арифметическое центральных разностей не демпфирует высокочастотных гармоник решения [8, 9]. Хотя выражение для расчета невязких потоков и включает диссипативные слагаемые, демпфирующие высокочастотные осцилляции решения, этого недостаточно в пограничном слое, где вязкие члены становятся доминирующими. Исходя из этого, составляющая градиента в направлении наиболее короткой грани заменяется простыми разностями:

$$\nabla Q_{ij} = \left(\nabla Q_{ij}\right)^* - \left[\left(\nabla Q_{ij}\right)^* \cdot \delta\mathbf{s}_{ij} - \frac{Q_j - Q_i}{\left|\mathbf{x}_j - \mathbf{x}_i\right|}\right]\delta\mathbf{s}_{ij}, \qquad \delta\mathbf{s}_{ij} = \frac{\mathbf{x}_j - \mathbf{x}_i}{\left|\mathbf{x}_j - \mathbf{x}_i\right|}. \qquad (24)$$

Дискретизация по времени

Перепишем уравнение (3) в виде

$$\frac{dQ}{dt} = L(Q), \qquad (25)$$

где $L(Q)$ – дифференциальный оператор. После линеаризации уравнение (25) примет вид

$$\frac{dQ}{dt} = CQ,$$

где C – квадратная матрица.

Для дискретизации уравнения (25) используется k-шаговый метод Рунге–Кутта

$$Q^{(n+1)} = \Lambda(kC)Q^{(n)}, \qquad \Lambda(z) = \sum_{m=0}^{p} a_m z^m,$$

где $a_0 = a_1 = 1$ и $a_p \neq 0$. Область устойчивости $S = \{z = x + iy: |\Lambda(z)| < 1\}$ имеет вид окружности с радиусом r_c. На комплексной плоскости область устойчивости представляется в виде круга [7]:

$$z = r\exp(i\theta).$$

Радиус области устойчивости находится из соотношения

$$r_c = \min_{\theta} r(\theta), \qquad \frac{\pi}{2} \le \theta \le \frac{3\pi}{2}.$$

Приведем расчетные соотношения пятишагового метода Рунге–Кутты ($r_c = 2{,}7$):

$$Q_i^{(m)} = Q_i^{(0)} - \alpha_m \Delta t_i R_i^{(m-1)} \qquad (m = 1,\ldots,5),$$

где

$$R_i^{(m-1)} = C_i\left(Q_i^{(m-1)}\right) - B_i^{(m-1)};$$

$$B_i^{(m-1)} = \beta_m D_i\left(Q_i^{(m-1)}\right) + (1-\beta_m)B_i^{(m-2)}.$$

При этом $C_i\left(Q_i^{(m-1)}\right)$ представляет собой вклад конвективных слагаемых, $D_i\left(Q_i^{(m-1)}\right)$ учитывает вклад источниковых членов, а также физической и численной диссипации. Коэффициенты α_m и β_m имеют следующие значения:

$$\alpha_1 = \frac{1}{4}, \quad \alpha_2 = \frac{1}{6}, \quad \alpha_3 = \frac{3}{8}, \quad \alpha_4 = \frac{1}{2}, \quad \alpha_5 = 1;$$

$$\beta_1 = 1, \quad \beta_2 = 0, \quad \beta_3 = \frac{14}{25}, \quad \beta_4 = 0, \quad \beta_5 = \frac{11}{25}.$$

Поскольку $\beta_2 = 0$ и $\beta_4 = 0$, то $D_i\left(Q_i^{(2)}\right)$ и $D_i\left(Q_i^{(4)}\right)$ не рассчитываются.

Ускорение сходимости

Для ускорения сходимости используется многосеточный метод решения системы разностных уравнений. Для построения последовательности вложенных неструктурированных сеток используется метод схлопывающихся граней [9]. Два узла i и j неструктурированной сетки, связанных гранью, заменяются одним узлом, расположенным посередине между ними. Схлопывание ячейки производится в направлении наиболее короткой грани. Градиент находится из соотношения (24).

Применяется схема полной аппроксимации [10]. В отличие от методов линеаризации по Ньютону с адаптацией числа многосеточных итераций на каждой итерации или с фиксированным числом многосеточных итераций на каждом шаге, схема полной аппроксимации позволяет избежать глобальной линеаризации (линеаризация проводится внутри цикла на самой грубой сетке), не проводить расчета больших якобианов, а также использовать разнообразные алгоритмы сглаживания. Согласования внутренних и внешних итераций не требуется.

Результаты расчетов

Рассмотрим течение идеального сжимаемого газа в межлопаточном канале газотурбинной установки. Предполагается, что лопатки турбины вибрируют с заданной амплитудой и фазой. Разница фаз колебаний всех лопаток турбины считается постоянной величиной, что позволяет выделить в качестве расчетной области участок турбинного вала, содержащий две лопатки [11].

Профиль лопатки представляет собой модифицированный профиль NACA0006 [12], который строится при помощи суперпозиции распределений кривизны и толщины профиля вдоль координаты $0 \le x/L \le 1$, где L – хорда профиля. Распределения кривизны и толщины профиля даются соотношениями

$$C(x) = H_c - R + \left[R^2 - (x - 0{,}5)^2\right]^{1/2};$$

$$T(x) = H_t\left(2{,}969x^{1/2} - 1{,}26x - 3{,}51x^2 - 2{,}843x^3 - 1{,}036x^4\right).$$

Здесь $R = \left(H_c^2 + 0,25\right)/2H_c$. Величины H_t и H_c представляют собой максимальную толщину профиля и радиус кривизны в серединной точке хорды профиля соответственно. В расчетах принимается, что $H_t = 0,06$, $H_c = 0,05$ (рис. 4).

Рис. 4. Профиль лопатки турбины

Геометрия расчетной области показана на рис. 5. Угол расположения лопатки относительно оси x равняется $\beta = 45^\circ$. Поток поступает через входное сечение под углом α к горизонтальной оси (угол атаки профиля равняется $\alpha - \beta$). В качестве рабочей среды принимается воздух.

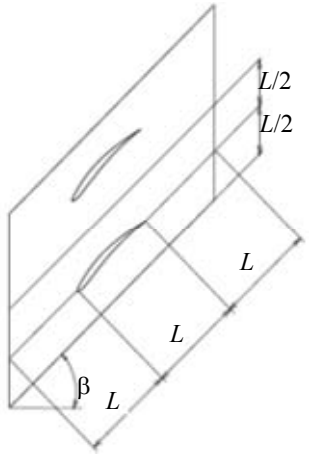


Рис. 5. Геометрия расчетной области

Рассматриваются два варианта течения в межлопаточном канале, соответствующие различным граничным условиям во входном сечении. Во входном сечении задается направление течения и число Маха (M = 0,7, $\alpha = 55^\circ$ в случае 1 и M = 0,8, $\alpha = 58^\circ$ в случае 2). Приведенные значения чисел Маха соответствуют перепадам давлений $p_1/p_0 = 0,8716$ в случае 1 и $p_1/p_0 = 0,8740$ в случае 2. Перепад давления в выходном сечении расчетной области определяется при помощи решения соответствующей стационарной задачи (течение в канале с неподвижными лопатками) при нескольких заданных значениях числа Маха на входе. Такая процедура требуется потому, что в [11] давление в выходном сечении не указывается (приводится только число Маха на входе). Полученный перепад давлений задается в качестве граничного условия в выходном сечении расчетной области. На направлении оси z выставляются периодические граничные условия. На поверхности профиля используется граничное условие непротекания для нормальной составляющей скорости.

Неструктурированная сетка в серединном сечении канала содержит 5443 треугольных элемента (5 вложенных сеток, сетка наилучшей разрешающей способности содержит 115 узлов на поверхности профиля, V-цикл). Сгущение узлов сетки производится у поверхности профиля. Шаг интегрирования по времени полагается равным $\Delta t = 3,25 \cdot 10^{-5}$ с.

Вибрация лопаток вопроизводится в виде двух последовательных циклов – восходящего и нисходящего движения профиля по гармоническому закону. Восходящее движение профиля происходит с амплитудой 0,01 м по нормали к его хорде. Амплитуда нисходящего движения составляет 2º относительно серединной линии профиля. Рассматривается 3 частоты ($\omega = 0,25$, 0,5, 0,75 рад/с) и 2 фазовых угла ($\varphi = 0^\circ, 90^\circ$). Производится расчет одного периода колебаний $T = 2\pi/\omega$. В качестве начального приближения используется стационарное решение задачи.

Смещение узлов, лежащих на поверхности лопатки, в пространстве и времени представляется в виде линейного гармонического возмущения их первоначального положения:

$$\mathbf{x}(t) = \mathbf{x}_0 + \sin\left(\omega t\right)\left(a^+ \delta\mathbf{x}^+ + a^- \delta\mathbf{x}^-\right).$$

Амплитуды возмущений вычисляются следующим образом:

$$a^+ = \frac{1}{2}\left\{1 + \operatorname{sgn}\left[\sin\left(\omega t\right)\right]\right\}, \qquad a^- = \frac{1}{2}\left\{1 - \operatorname{sgn}\left[\sin\left(\omega t\right)\right]\right\}.$$

Здесь $\mathbf{x}_0$ – начальное невозмущенное положение узлов; $\delta\mathbf{x}^+$ и $\delta\mathbf{x}^-$ – максимальные отклонения узлов от невозмущенного положения.

Новые координаты граничных узлов находятся из соотношения

$$\mathbf{x}^{n+1}(\mathbf{x},t) = \mathbf{x}_0(\mathbf{x}^n) + \mathrm{Real}\left\{ \left[d\mathbf{x}_r(\mathbf{x}^n) + i d\mathbf{x}_i(\mathbf{x}^n) \right] \exp(i\omega t) \right\},$$

причем

$$\mathbf{x} = \mathbf{x}_0 + d\mathbf{x}_r \cos(\omega t) - d\mathbf{x}_i \sin(\omega t), \qquad \frac{\partial \mathbf{x}}{\partial t} = -\omega \left[d\mathbf{x}_r \sin(\omega t) + d\mathbf{x}_i \cos(\omega t) \right].$$

Результаты расчетов показывают, что в случае 1 течение в канале является полностью дозвуковым. В случае 2 в лопаточном канале возникают локальные области сверхзвукового течения. В отличие от частотных, фазовые характеристики оказывают достаточно слабое влияние на параметры потока в межлопаточном канале.

Распределения коэффициента давления по поверхности профиля показаны на рис. 6 и рис. 7 для случаев 1 и 2 при двух положениях профиля. Коэффициент давления определяется следующим образом:

$$C_p = \frac{p_D - p_U}{\rho U^2 |a\omega|},$$

где p_U и p_D представляют собой давления на верхней и нижней поверхности профиля.

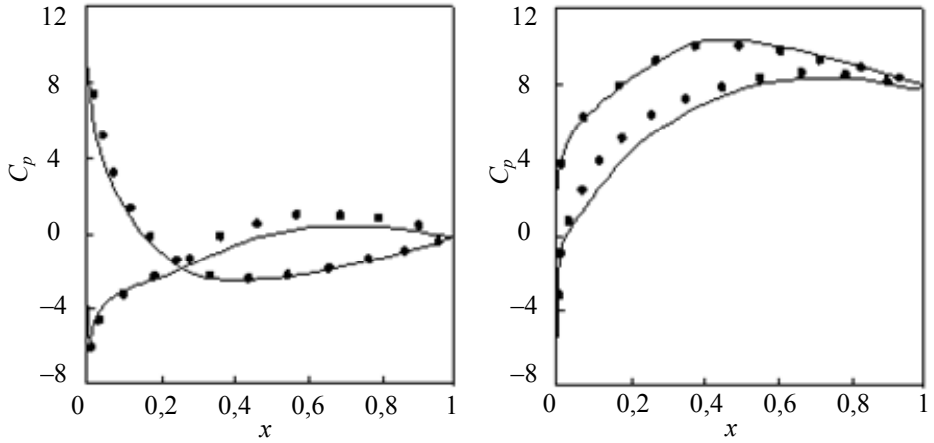


Рис. 6. Распределения коэффициента давления по поверхности профиля при восходящем (а) и нисходящем (б) движении в случае 1. Значки • соответствуют данным [13]

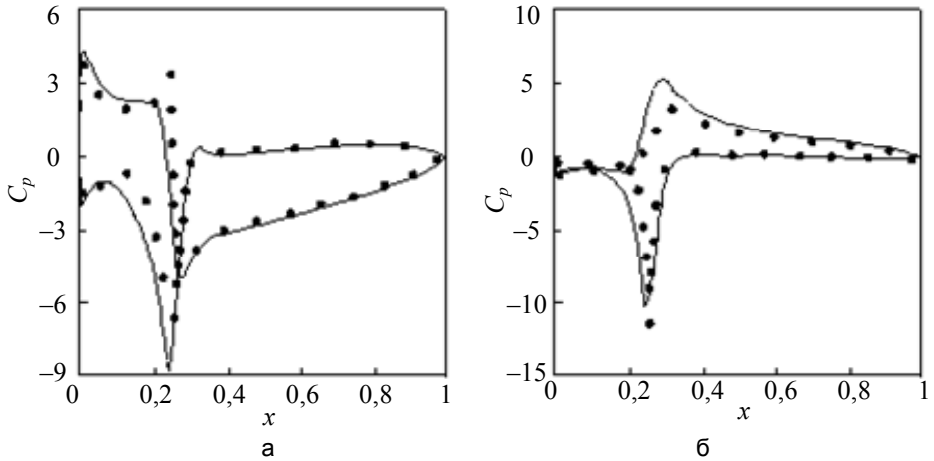


Рис. 7. Распределения коэффициента давления по поверхности профиля при восходящем (а) и нисходящем (б) движении в случае 2. Значки • соответствуют данным [13]

Заключение

Разработан подход к дискретизации нестационарных уравнений Навье–Стокса на неструктурированных сетках в рамках метода конечных объемов применительно к двух- и трехмерным задачам механики жидкости и газа. Предлагаемый подход использует разностные схемы высокого порядка по времени и по пространственным переменным, среднемедианный контрольный объем и модифицированные соотношения для расчета градиента и псевдолапласиана, позволяющие получить более точные результаты на сильно растянутых сетках в пограничном слое, одинаковые выражения для расчета потоков через грани внутренних и граничных контрольных объемов, что обеспечивает его более простую программную реа-

лизацию, позволяет легко реализовать стратегию адаптации сеток в соответствии с особенностями конкретных течений и дает обширные возможности для параллелизации процессов вычислений.

Литература

1. Волков К.Н., Емельянов В.Н. Вычислительные технологии в задачах механики жидкости и газа. М.: ФИЗМАТЛИТ, 2013. 468 с.
2. Cagnone J.S., Sermeus K., Nadarajah S.K., Laurendeau E. Implicit multigrid schemes for challenging aerodynamic simulations on block-structured grids // Computers and Fluids. 2011. V. 44. N 1. P. 314–327.
3. Deng X., Mao M., Tu G., Zhang H., Zhang Y. High-order and high accurate CFD methods and their applications for complex grid problems // Communications in Computational Physics. 2012. V. 11. N 4. P. 1081–1102.
4. Wang Z.J., Fidkowski K., Abgrall R., Bassi F., Caraeni D., Cary A., Deconinck H., Hartmann R., Hillewaert K., Huynh H.T., Kroll N., May G., Persson P.-O., van Leer B., Visbal M. High-order CFD methods: current status and perspective // International Journal for Numerical Methods in Fluids. 2012. V. 72. N 8. P. 811–845.
5. Luo H., Baum J.D., Lohner R. Edge-based finite element scheme for the Euler equations // AIAA Journal. 1994. V. 32. N 6. P. 1183–1190.
6. Crumpton P.I., Moinier P., Giles M.B. An unstructured algorithm for high Reynolds number flows on highly stretched grids // Proc. 10th International Conference on Numerical Methods in Laminar and Turbulent Flow. Swansea, UK, 1997. P. 561–572.
7. Brognies Z., Rajasekharan A., Farhat C. Provably stable and time-accurate extensions of Runge-Kutta schemes for CFD computations on moving grids // International Journal for Numerical Methods in Fluids. 2012. V. 69. N 7. P. 1249–1270.
8. Moinier P., Giles M.B. Stability analysis of preconditioned approximations of the Euler equations on unstructured meshes // Journal of Computational Physics. 2002. V. 178. N 2. P. 498–519.
9. Crumpton P.I., Giles M.B. Implicit time accurate solutions on unstructured dynamic grids // International Journal for Numerical Methods in Fluids. 1997. V. 25. N 11. P. 1285–1300.
10. Brandt A. Multi-level adaptive solutions to boundary-value problems // Mathematics of Computation. 1977. V. 31. N 138. P. 333–390.
11. Fransson T.H., Verdon J.M. Standard configurations for unsteady flow through vibrating axial-flow turbomachine-cascades // Unsteady Aerodynamics, Aeroacoustics and Aeroelasticity of Turbomachines and Propellers. Ed. H.M. Atassi. NY: Springer-Verlay, 1993. P. 859–889.
12. Verdon J.M. Linearized unsteady aerodynamic theory // United Technologies Research Center Report R85-151774-1. 1987.
13. Lawrence C., Spyropoulos E., Reddy T.S.R. Unsteady cascade aerodynamic response using a multiphysics simulation code // NASA Report TM-2000-209635. 2000.

Волков Константин Николаевич – доктор физико-математических наук, научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; старший лектор, Университет Кингстона, Лондон, SW15 3DW, Великобритания, k.volkov@kingston.ac.uk

Konstantin N. Volkov – D.Sc., Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation; Senior Lecturer, Kingston University, London, SW15 3DW, United Kingdom, k.volkov@kingston.ac.uk

Принято к печати 14.05.14
Accepted 14.05.14

УДК 004.942

МНОГОУРОВНЕВАЯ РЕКУРРЕНТНАЯ МОДЕЛЬ ИЕРАРХИЧЕСКОГО УПРАВЛЕНИЯ КОМПЛЕКСНОЙ БЕЗОПАСНОСТЬЮ РЕГИОНА

А.В. Маслобоев[a,b], В.А. Путилов[a,b], А.В. Сютин[a,c]

[a] Институт информатики и математического моделирования технологических процессов Кольского научного центра РАН, Апатиты Мурманской обл., 184209, Российская Федерация

[b] Кольский филиал Петрозаводского государственного университета, Апатиты Мурманской обл., 184209, Российская Федерация, masloboev@iimm.ru

[c] Университет Бергена, Берген, N-5020, Норвегия

Аннотация.

Предмет исследования. Целью исследования является разработка методов и средств математического и компьютерного моделирования систем информационного обеспечения региональной безопасности как многоуровневых иерархических систем. Эти системы характеризуются слабой формализованностью, многоаспектностью происходящих в них процессов и их взаимосвязанностью, динамичностью и высокой степенью неопределенности. Методологическая база исследования включает функционально-целевой подход и аппарат теории иерархических многоуровневых систем. В работе решаются задачи анализа и структурно-алгоритмического синтеза автоматизированных систем, ориентированных на информационную поддержку процессов управления и принятия решений в сфере региональной безопасности.

Основные результаты. Разработана многоуровневая рекуррентная модель иерархического управления комплексной безопасностью региональных социально-экономических систем. Модель основана на функционально-целевом подходе и обеспечивает как формальную постановку и решение, так и практическую реализацию задач синтеза структуры автоматизированных систем и алгоритмов управления региональной безопасностью, оптимальных в смысле определенных критериев. Предложен подход к решению задач внутриуровневой и межуровневой координации в многоуровневых иерархических системах. Такая координация обеспечивается за счет удовлетворения требований взаимосвязи между показателями качества функционирования (целевыми функциями), оптимизируемыми различными элементами многоуровневых систем. Это позволяет достичь достаточной согласованности локальных решений, принимаемых на разных уровнях управления, в условиях децентрализованного принятия решений и высокой динамики внешней среды. Использование рекуррентной модели позволяет сформировать математические модели управления безопасностью региональных социально-экономических систем, функционирующих в условиях неопределенности.

Практическая значимость. Практическая реализация предложенной модели позволяет проводить автоматизированный синтез программной исполнительной среды информационно-аналитической поддержки принятия управленческих решений в сфере региональной безопасности. Модель сможет найти применение при создании методологии математического и компьютерного моделирования многоуровневых иерархических систем управления комплексной безопасностью сложных систем.

Ключевые слова: математическое моделирование, многоуровневая иерархическая система, управления, координация, рекуррентная модель, региональная безопасность, поддержка принятия решений.

Благодарности. Результаты работы получены в ходе исследований, проводимых по планам научно-исследовательских работ Института информатики и математического моделирования технологических процессов Кольского научного центра РАН (НИР № 01201452426 «Методы и когнитивные технологии создания, исследования и использования виртуальных систем поддержки управления комплексной безопасностью развития Арктической зоны Российской Федерации»). Авторы выражают благодарность своим коллегам по лаборатории за участие во всестороннем обсуждении результатов работы.

MULTILEVEL RECURRENT MODEL FOR HIERARCHICAL CONTROL OF COMPLEX REGIONAL SECURITY

A.V. Masloboev[a,b], V.A. Putilov[a,b], A.V. Sioutine[a,c]

[a] Institute for Informatics and Mathematical Modeling of Technological Processes Kola Science Center of the Russian Academy of Sciences, Apatity, 184209, Russian Federation

[b] Kola Branch of Petrozavodsk State University, Apatity, 184209, Russian Federation, masloboev@iimm.ru

[c] University of Bergen, Bergen, N-5020, Norway

Abstract.

Subject of research. The research goal and scope are development of methods and software for mathematical and computer modeling of the regional security information support systems as multilevel hierarchical systems. Such systems are characterized by loosely formalization, multiple-aspect of descendent system processes and their interconnectivity, high level dynamics and uncertainty. The research methodology is based on functional-target approach and principles of multilevel hierarchical system theory. The work considers analysis and structural-algorithmic synthesis problem-solving of the multilevel computer-aided systems intended for management and decision-making information support in the field of regional security.

Main results. A hierarchical control multilevel model of regional socio-economic system complex security has been developed. The model is based on functional-target approach and provides both formal statement and solving, and practical implementation of the automated information system structure and control algorithms synthesis problems of regional security management optimal in terms of specified criteria. An approach for intralevel and interlevel coordination problem-solving in the multilevel hierarchical systems has been proposed on the basis of model application. The coordination is provided at the expense of interconnection requirements satisfaction between the functioning quality indexes (objective functions), which are optimized by the different elements of multilevel systems. That gives the possibility for sufficient coherence reaching of the

local decisions, being made on the different control levels, under decentralized decision-making and external environment high dynamics. Recurrent model application provides security control mathematical models formation of regional socio-economic systems, functioning under uncertainty.

Practical relevance. The model implementation makes it possible to automate synthesis realization of the software executive environment for decision-making information and analytical support in the field of regional security. The model can find further application within mathematical and computer modeling methodology development of the multilevel hierarchical systems for security control of complex systems.

Keywords: mathematical modeling, multilevel hierarchical system, control, coordination, recurrent model, regional security, decision-making support.

Acknowledgements. Findings of this investigation are received within the bounds of research works carried out according to research plans of the Institute for Informatics and Mathematical Modeling of Technological Processes of the Kola Science Center of the Russian Academy of Sciences (project №01201452426 "Methods and cognitive technologies for engineering, analysis and application of the virtual systems for complex security management support of the of the Russian Federation Arctic zone development"). The authors express their thanks to their lab colleagues for assistance and participation within the comprehensive discussion of the research results.

Введение

Анализ тенденций развития социально-экономической и геополитической ситуации в глобальном, региональном и национальном масштабах показывает, что обстановка в Арктической зоне Российской Федерации является в целом сложной для нашей страны. В условиях активного геостратегического переустройства мира и борьбы мировых центров силы за контроль над ресурсами (природными, кадровыми, информационными и т.д.) проблемы становления новой системы обеспечения региональной безопасности в Российской Арктике не теряют своей остроты и актуальности. Эти проблемы дают определенный импульс развитию сферы компьютерных технологий для задач управления комплексной безопасностью региональных систем, так как их решение во многом требует интеграции, обработки и анализа больших объемов семантически и организационно разнородной информации для информационного обеспечения межведомственной деятельности в области региональной безопасности, а также поддержки принятия решений на разных уровнях управления.

Решение задач информационной поддержки управления региональной безопасности затрудняется отсутствием целостной многофункциональной информационной инфраструктуры региональной безопасности в арктических регионах [1], что препятствует эффективному использованию в практической деятельности субъектов безопасности интеллектуализированных многоуровневых автоматизированных систем управления комплексной безопасностью, интегрированных в единое информационное пространство региона. Такая информационная среда, согласно исследованиям [2–4], призвана обеспечить комплексную информационно-аналитическую поддержку процессов принятия стратегических и оперативных решений на разных уровнях управления на основе применения заложенного в нее методического инструментария и соответствующих информационных технологий, т.е. управления системой обеспечения комплексной безопасности региона.

Под региональной безопасностью понимается состояние защищенности региональной системы, при котором действие внешних и внутренних факторов не приводит к ухудшению или к невозможности ее функционирования или развития. Определение и формализация термина «региональная безопасность» подробно рассматриваются в работе [5]. Управление региональной безопасностью – сложная многокритериальная задача. Для успешного решения этой задачи на практике необходимо, чтобы процессы обеспечения региональной безопасности были управляемыми, т.е. существовала система управления этими процессами, способная к структурной реконфигурации, самоорганизации и адаптации состояния, параметров и режимов функционирования к динамике внешней среды в различных условиях и ситуациях. Процессы обеспечения безопасности компонентов региональных систем разнородны по динамике и составу участников. Субъекты безопасности, вовлеченные в эти процессы, как правило, территориально распределены. Это обусловливает динамичность и разнородность информационной среды региональной безопасности, необходимость в механизмах координации взаимодействия образующих ее подсистем в условиях децентрализованного управления и принятия решений. Такая среда характеризуется сетецентричностью [6] и синтезируется на базе объединения многоуровневых систем управления различными видами безопасности.

Управление региональной безопасностью по своей структуре многофункционально и в общем случае включает в себя такие функции управления, как целеполагание, стратегическое планирование, оперативное управление, а также функции контроля, учета, мониторинга и координации. Для эффективного децентрализованного принятия решений на разных уровнях управления безопасностью региона в условиях многокритериальности решаемых задач управления и различий в целеполагании разнородных субъектов безопасности необходимо обеспечить межуровневую и внутриуровневую координацию взаимодействия между ними за счет удовлетворения требований взаимосвязи между показателями качества

функционирования (целевыми функциями), оптимизируемыми различными элементами соответствующей системы управления комплексной безопасностью региона.

В работе рассматриваются различные теоретические аспекты организации и структурно-алгоритмического синтеза многоуровневых распределенных систем информационной поддержки управления региональной безопасностью. Предложена многоуровневая рекуррентная модель иерархического управления комплексной безопасностью региональных социально-экономических систем. Результаты работы получены на основе применения функционально-целевого подхода, развитого для класса задач с древовидными моделями предметной области [7] и предложенного профессором В.А. Путиловым в начале 80-х годов прошлого века для решения проблем управления сложными распределенными объектами [8], в том числе региональными социально-экономическими системами.

Моделирование многоуровневых распределенных систем управления региональной безопасностью на базе функционально-целевого подхода

В функционально-целевом подходе рекуррентная модель предметной области служит основой формализации главных задач структурно-алгоритмической организации автоматизированных систем и методов решения этих задач. Новизна рекуррентной модели определяется, во-первых, тем, что при построении модели целевого управления использована иерархия двухоперационных алгебр цепочек целей и совершенно аналогичных по структуре цепочек действий, обеспечивающих достижение этих целей. Во-вторых, иерархия целей в модели непосредственно порождает модель иерархии действий, что обеспечивает использование иерархии целей не только в качестве средства описания задачи, но и как средства проектирования системы управления. Модель основана на иерархической структуре задач управления региональной безопасностью и использует последовательно-параллельные композиции целей управления и действий по достижению этих целей.

В применении к задачам синтеза автоматизированных систем управления комплексной безопасностью на практике, как правило, используются модели в виде графа с произвольной структурой, нечеткие [9], когнитивные [10] и имитационные [5] модели опять же на базе таких графов. Для многоуровневых распределенных систем такие модели уже малопригодны, так как приводят к сложным моделям в виде системы вложенных графов произвольной структуры. Иерархические модели успешно применяются в других приложениях. Наиболее близок к настоящей работе программно-целевой подход [11], но здесь модель не древовидная, и это объясняется спецификой предметной области, где непременно должны присутствовать связи между элементами одного уровня. При решении оптимизационных задач в программно-целевом подходе используется достаточно сложный аппарат траекторной оптимизации.

В многоуровневых распределенных системах такие понятия, как «цель», «целенаправленная деятельность», «целенаправленные системы», тесно связаны с понятиями «принятие решений» и «системы принятия решений» [12]. Целенаправленное поведение, в сущности, представляет собой последовательность принимаемых и реализуемых решений. Вследствие возможности представления систем типа вход–выход в виде решающих систем [13], и наоборот, цели могут быть определены через решаемые задачи. В связи с этим цель считается достигнутой, когда найдено решение соответствующей задачи (задача может быть оптимизационной). В дальнейшем в соотношение между целями и решаемыми задачами в работе будет вкладываться именно такой смысл.

Для решения задач структурно-алгоритмического синтеза многоуровневой автоматизированной системы управления комплексной безопасностью региональных социально-экономических систем необходимо построить на основе функционально-целевого подхода модель данной предметной области и соответствующую ей в определенном смысле модель автоматизированной системы управления региональной безопасностью. Вместе с тем, необходимо задать критерий эффективности функционирования системы и определить механизм построения систем автоматизированного управления региональной безопасностью, эффективных в смысле заданного критерия, на базе созданных моделей.

Реализация функционально-целевого подхода, базирующегося на концепции управления через целеполагание и предполагающего соответствие функций системы управления целям предметной области, обеспечивает как формальную постановку и решение, так и практическую реализацию задач синтеза структуры автоматизированной системы и алгоритмов управления региональной безопасностью, оптимальных в смысле определенных критериев. Таким подходом обеспечивается и учет особенностей решаемых задач управления и принятия решений.

Иерархическое представление систем используется в разных приложениях, в том числе и для многоуровневых систем управления. Это объясняется простотой и наглядностью иерархических моделей, хорошо отражающих реальные взаимосвязи в окружающем нас мире, включая организации людей. Существуют и другие доводы в пользу иерархических многоуровневых систем [13, 14]:

- эти системы появляются при интеграции уже созданных систем;
- для решения общей задачи системы могут эффективно использоваться ограниченные возможности подсистем;

– системы лучше адаптируются к изменениям и усложнениям задач и обладают хорошими показателями надежности (неисправности в работе какой-либо подсистемы не всегда распространяются на всю систему).

Однако в многоуровневых иерархических системах возникают задачи координации, и использование этих систем оправданно, если удается упростить задачу координации до такой степени, чтобы она была значительно проще решаемой проблемы.

В работе [15] под общей многоуровневой иерархической системой понимается совокупность объектов и элементов, в которых рассматриваются процессы, управляемые соответствующими субъектами управления, имеющими собственные сферы интересов (цели) в условиях иерархической подчиненности основному субъекту управления. Система функционирует в конкретной среде при наличии неопределенности и имеет в своем составе следующие компоненты:

– средства ввода, вывода, приема, хранения, анализа, обработки и передачи данных, сопряженные с ее объектами или элементами;
– средства, позволяющие реализовать математические, логические и иные операции, сопряженные с ее объектами или элементами;
– средства для реализации информационных и управляющих связей между ее объектами или элементами;
– средства, позволяющие реализовать выбранные системы кодирования и декодирования данных, сопряженные с ее объектами или элементами.

Иерархическая структура заложена в самом понятии комплексной безопасности региона, заведомо образуемой различными по характеру согласованными составляющими региональной безопасности (экономической, экологической, социальной и др.). Каждая из составляющих региональной безопасности, в свою очередь, образуется набором объектов, субъектов, процессов и методов обеспечения безопасности, потенциальных угроз и опасностей. Такая детализация может продолжаться и далее.

Обращаясь к соотношению между целями и решениями задач (если найдено решение задачи, в том числе и оптимизационной, то достигнута соответствующая цель), видно, что процесс последовательной детализации задачи в области управления региональной безопасностью представляется деревом декомпозиции целей управления. Обратимся к задаче синтеза структуры многоуровневой системы управления региональной безопасностью, обеспечивающей достижение целей управления. Такая задача имеет много аспектов. Доказанная в работе [13] теорема о подсистемах многоуровневой системы показывает, что система в целом должна строиться из таких подсистем, которые обеспечивают покрытие соответствующих подзадач основной целевой задачи многоуровневой системы. Из теоремы также следует, что синтез структуры системы должен проводиться изоморфно построению основной цели из некоторой совокупности подцелей.

Рекуррентная модель иерархического управления региональной безопасностью

Перейдем к построению формальной рекуррентной модели иерархического управления региональной безопасностью и многоуровневых систем управления ей. Макроструктура многоуровневой системы управления комплексной безопасностью региона, построенная изоморфно декомпозиции основной целевой задачи обеспечения региональной безопасности, представляется в виде дерева. Корню дерева ставится в соответствие подсистема верхнего уровня (собственно система), вершинам дерева, отстоящим от корня на одно ребро, – подсистемы, реализующие классы безопасности, на три ребра – подсистемы, реализующие методы и средства обеспечения безопасности, и т.д.

Построение виртуальной макроструктуры системы и отображение ее на реальную структуру программно-аппаратных средств позволяет определить набор элементарных компонентов для структурно-алгоритмического синтеза системы. Синтез адекватной системы – трудоемкая задача, связанная с необходимостью удовлетворения условий изоморфизма на всех соответствующих уровнях декомпозиции задачи и организации системы, причем эти требования должны удовлетворяться для любой задачи рассматриваемой предметной области, что находится в противоречии с требованием гибкости системы по отношению к описанию предметной области. Поэтому для практических приложений актуальна задача синтеза покрывающих систем [13], обеспечивающих решение задач субъекта управления на всех уровнях организации системы с удовлетворяющими его значениями параметров качества цепочек действий. При этом по известным параметрам атомарных элементов нижнего уровня макроструктуры строятся отображения алгебры цепочек на алгебры соответствующих параметров «снизу вверх», до уровня иерархии системы, на котором субъект управления может принять решение либо о целесообразности использования синтезированной системы, либо о необходимости изменения постановки задачи или коррекции программно-аппаратного обеспечения системы с целью изменения параметров атомарных элементов.

Будем характеризовать любой элемент M макроструктуры системы состоянием S, управляющим воздействием U для задания режима работы элемента и его состояния, входной информацией W. Поскольку элементы макроструктуры – это программы, ориентированные на прием, переработку и переда-

чу информации, то результатом работы элемента *M* является некоторая выходная информация *V*. Будем рассматривать результирующую информацию *V* как некоторую функцию от состояния элемента макроструктуры *M*, входной информации и управляющего воздействия: $V = M\,(U, S, W)$.

Под элементарной неделимой единицей алгоритма управления безопасностью условимся понимать функциональную операцию *L* – некоторую совокупность действий исполнительной системы, зависящих от управляющего воздействия, состояния элемента макроструктуры и его внутренней структуры: $L = M\,(U, S)$. Функциональные операции выполняют преобразование входной информации *W* в выходную следующим образом: $L_{MUS} : V = L_{MUS}\,(W)$. Считывание состояния элемента макроструктуры достигается подачей специального управляющего сигнала U_0: $S = L_{MUS_0}\,(0)$.

Таким образом, определен алфавит функциональных операций (множество атомарных действий системы): $L = M \times U \times S$.

Представим содержательную информацию рассматриваемой предметной области (региональной безопасности) в виде формальных высказываний. Построим алгебраическую систему $A =< A, Q, R >$, состоящую из непустого множества *A,* семейства алгебраических операций *Q* и семейства отношений *R*. Для задания такой системы определим некоторые исходные объекты, которые будем рассматривать как неделимые; перечислим способы комбинирования исходных объектов между собой; укажем условие, которому удовлетворяют те и только те комбинации исходных объектов, которые считаются элементами системы; сформулируем условие, при котором два элемента системы считаются равными.

Отождествим семейство *Q* алгебраических операций с организацией достижений совокупной цели управления из известных атомарных целей, достижение которых реализуется атомарными действиями, заданными алфавитом *L*. Такие совокупные цели обеспечиваются комбинациями последовательного и параллельного (одновременного) достижения атомарных целей, т.е. композициями элементов функционального алфавита целей, построенными с использованием двух обобщенных операций:

1. операция $\odot$: достичь атомарной цели a_2 после достижения атомарной цели a_1;
2. операция $\oplus$: достичь атомарной цели a_2 одновременно с атомарной целью a_1.

Использование принципа управления через целеполагание обеспечивает организацию всего многообразия вариантов обеспечения региональной безопасности через композиции элементов функционального алфавита, построенные с использованием двух введенных обобщенных операций. Действительно, совокупная цель достигается последовательно-параллельной комбинацией подцелей нижнего уровня.

Введем точно такие же операции для атомарных действий – элементов функционального алфавита *L*. Зададим операцию как последовательное применение следующей функциональной операции к результату предыдущей:

$$\odot : L_j = L_{j-2} \odot L_{j-1} \to L_j(\mathrm{W}) = L_{j-1}\left(L_{j-2}(W)\right).$$

Операцию $\oplus$ зададим как одновременное выполнение двух атомарных воздействий:

$$\oplus : L_j = L_i \oplus L_k \to L_j(W) = \begin{cases} L_i(W_i), \\ L_k(W_k). \end{cases}$$

Операция $\odot$ производит последовательный запуск и исполнение выбранных атомарных элементов вычислительного процесса. Операция $\oplus$ производит параллельный запуск выбранных атомов.

В работе [13] проведены исследования полученной алгебры строк (цепочек), определены свойства замкнутости, ассоциативности, коммутативности относительно введенных операций $\oplus$ и $\odot$, а также установлено наличие нулевого, единичного и обратных элементов. Не теряя общности, ограничим рассмотрение алгеброй действий, в которой нагляден физический смысл введенных обобщенных операций. Полученные результаты справедливы и для алгебры целей.

Зададим на алгебре цепочек *A* некоторое отношение эквивалентности *R*. Отношение эквивалентности может задаваться как совпадение параметров цепочек (например, длины или используемых операций), либо как совпадение параметров результата, т.е. при одинаковой входной информации в результате выполнения двух разных цепочек получаем результирующую информацию, принадлежащую в обоих случаях к одному некоторому множеству.

Известно, что заданное некоторым образом отношение эквивалентности *R* разбивает все множество цепочек на множество непересекающихся классов эквивалентности. Исходя из этого, все семантически одинаковые цепочки находятся в пределах одного класса эквивалентности. Классы эквивалентности $\{z_i\}$ характеризуются следующими соотношениями:

$$\{z_i\} : \begin{cases} 1)\ \forall\, a_1, a_2 \in z_i, a_1 R\, a_2\,, \\ \quad 2)\ \bigcup\limits_i z_i = A, \\ 3)\ z_i \cap z_j = \begin{cases} z_i, i = j, \\ \emptyset, i \neq j. \end{cases} \end{cases}$$

В каждом классе эквивалентности задается новое отношение эквивалентности, разбивающее каждый класс эквивалентности на подклассы, и т.д. В результате получается семейство алгебр классов эквивалентности

$$A^k = \langle \Sigma^k, \{\odot, \oplus\} \rangle,$$

где Σ^k – множество цепочек над алфавитом $\left\{z_{\mathbf{j}^k}^k\right\}$.

Таким образом, строятся модели декомпозиции целей управления на комплексной предметной области и декомпозиции действий соответствующей автоматизированной системы управления, обеспечивающих достижение этих целей. Они получены абстрагированием от конкретного содержания составляющих предметных областей и заменой их понятием классов эквивалентности функций (целей или действий в зависимости от приложения модели), т.е. множеств функций, эквивалентных в смысле их предметной направленности. В каждом классе эквивалентности задано новое отношение эквивалентности, относящее функции к разным поднаправлениям и разбивающее каждый класс эквивалентности на подклассы. Рекуррентный процесс детализации исходной функции продолжается вплоть до достижения уровня «примитивов» – элементарных функций, неделимых с точки зрения субъекта управления. Тем самым задание множества отношений эквивалентности функций определяет топологию на множестве функций. Базой этой топологии является множество примитивов.

Полученная декомпозиция предметной области представляется древовидным графом иерархии классов, в котором узлы – имена классов, ребра – отношения включения, корень – имя функции на комплексной предметной области, листья – примитивы:

$$\left\{z_{\mathbf{j}^k}^k\right\}_{k=1}^{K} \; ; \; z_{\mathbf{j}^k}^k = \bigcup_{j_{k+1}} z_{\mathbf{j}^{k+1}}^{k+1},$$

$$z_{\mathbf{j}^k}^k = \cap \; z_{\mathbf{j}^k}^k = \begin{cases} z_{\mathbf{j}^k}^k, & \mathbf{i}^k = \mathbf{j}^k \\ \emptyset, & \mathbf{i}^k \neq \mathbf{j}^k \end{cases}, \qquad (1)$$

$$\mathbf{j}^k = (\mathbf{j}^k, j_{k+1}),\ \mathbf{j}^1 = 1,\ k = \overline{1,K},$$

где k – индекс уровня декомпозиции; K – число уровней; $\mathbf{j}^k$ – вектор-индекс длиной k класса эквивалентности на k-м уровне декомпозиции; $j_i\left(i = \overline{1,K}\right)$ – i-й компонент вектор-индекса; $z_{\mathbf{j}^k}^k$ – имя класса на k-м уровне декомпозиции с вектор-индексом $\mathbf{j}^k$. Система (1) порождается системой отношений эквивалентности

$$\left\{R_{\mathbf{j}^k}^k\right\}_{k=1}^{k-1} : \forall j_{k+1} \; \forall x, y \;,\; x \; R_{\mathbf{j}^{k+1}}^{k+1} \; y \Rightarrow x \; R_{\mathbf{j}^k}^k \; y,$$

$$\forall \mathbf{j}^k \; \exists \; j_{k+1} : x \; R_{\mathbf{j}^k}^k \; y \Rightarrow x \; R_{\mathbf{j}^{k+1}}^{k+1} \; y \;,$$

$$\forall \mathbf{j}^k \; \forall \mathbf{i}^k \; : \; \mathbf{i}^k \neq \mathbf{j}^k : x \; R_{\mathbf{i}^k}^k \quad y \Rightarrow \neg \, x \; R_{\mathbf{j}^k}^k \; y \;,$$

где $R_{\mathbf{j}^k}^k$ – отношение эквивалентности, разбивающее $z_{\mathbf{j}^k}^k$ на $\left\{z_{\mathbf{j}^{k+1}}^{k+1}\right\}, x, y \in z_{\mathbf{j}^k}^k$. Построенная алгебраическая система A, состоящая из множества элементов, двух алгебраических операций и семейства отношений эквивалентности, является формальной моделью постановки и решения задач организации процесса управления региональной безопасностью, поскольку одинаковым образом описывает цели управления и действия по достижению этих целей на любом уровне декомпозиции исходной задачи.

В общем случае имеется множество классов эквивалентности

$$Z = \left\{z_{\mathbf{j}^k}^k\right\}_{k=1}^{K}, \qquad (2)$$

где K – число уровней декомпозиции; соответственно общая рекуррентная модель представляет собой иерархию алгебр

$$A^k = \langle \Sigma^k, \{\odot, \oplus\} \rangle, \qquad (3)$$

гомоморфно отображенных друг на друга «снизу вверх»:

$$\gamma_k : \; A^{k+1} \rightarrow A^k, \qquad (4)$$

где γ_k есть совокупность отношений $\left\{R_{\mathbf{j}^k}^k\right\}$.

Таким образом, построены формальная рекуррентная модель предметной области (региональной безопасности), основанная на рекуррентной декомпозиции целей управления, и модель соответствующей автоматизированной системы управления комплексной безопасностью в этой предметной области, основанная на адекватном декомпозиции целей управления процессе детализации действий по обеспечению региональной безопасности. Построение рекуррентной модели многоуровневой системы позволяет формализовать постановку задачи ее управления, сводящегося к выполнению набора различных примитивов. Таким образом, определяется множество действий системы для достижения поставленной цели управления комплексной безопасностью региона.

Координация управлений в многоуровневых распределенных системах

Процедуры синтеза и анализа многоуровневых иерархических систем предполагают, что составляющие систему элементы обладают ограниченными возможностями по решению задач, стоящих перед системой. В связи с этим глобальная задача, отражающая назначение системы в целом, разбивается на совокупность подзадач таким образом, что решение глобальной задачи эквивалентно решению этой со-

вокупности. Такой подход может применяться как при проектировании структур многоуровневых систем, так и при организации решения системой задач [13, 14]. В многоуровневых иерархических системах при этом возникают специфические проблемы управления.

Из рассмотрения структуры рекуррентной модели (2)–(4) следует, что имеются вполне определенные предпосылки применения к этой модели результатов, полученных для двухуровневых иерархических систем [16]. Действительно, формальная рекуррентная модель получена регулярным рекуррентным применением к процессу декомпозиции основной целевой задачи двухуровневой структуры, имеющей один элемент на верхнем уровне и заданное моделью предметной области число элементов нижнего уровня. Такой простой вид этой (элементарной) модели и регулярные правила построения модели на базе элементарной обеспечивают возможность получения для рекуррентной модели как общих результатов при исследовании вопросов координации, так и конкретных алгоритмов структурно-алгоритмического синтеза. При этом под координацией в настоящей работе понимается свойство системы находить оптимальные решения общей задачи управления при оптимизации подзадач управления, решаемых подсистемами. Другими словами, координирование означает такое воздействие элемента вышестоящего уровня на элементы нижестоящего уровня, которое заставляет нижестоящие элементы действовать согласованно. Для обеспечения координации требуется реализовать определенные ограничения на взаимосвязи между подсистемами.

Управление в многоуровневой системе может быть организовано разными путями в зависимости от степени распределенности общей задачи системы между уровнями. Наиболее методологически простое решение состоит в том, что элемент верхнего уровня (координатор) имеет точное описание поведения элементов нижнего уровня; такая постановка приводит к обычным задачам дискретной оптимизации. Более методологически содержательный подход состоит в формализации задачи координатора с учетом того, что она задается взаимодействием семейства взаимосвязанных подсистем (элементов) нижнего уровня; каждая из подсистем при этом решает свою задачу и преследует свои цели, поэтому координатор должен координировать взаимодействия между элементами нижнего уровня, а не управлять ими; соответственно формализация задачи координатора должна быть основана на информации о том, каким образом элементы нижнего уровня при выборе своих решений учитывают эти взаимодействия. Последний путь позволяет для решения задачи, стоящей перед всей системой в целом, использовать совокупность решающих элементов, расположенных на различных уровнях организации системы, даже если каждый элемент в отдельности (включая и координатора) не в состоянии решить общую задачу. Для решения общая задача разбивается на подзадачи, решение которых производится групповыми усилиями решающих элементов.

Заключение

В ходе проведенных исследований получены следующие основные результаты.

1. Предложено приложение методологических принципов функционально-целевого подхода и теории иерархических многоуровневых систем к задачам анализа и синтеза автоматизированных систем управления региональной безопасностью.
2. Разработана многоуровневая рекуррентная модель иерархического управления комплексной безопасностью региональных социально-экономических систем.

Модель обеспечивает основу для автоматизированного синтеза программной исполнительной среды информационно-аналитической поддержки принятия управленческих решений в сфере региональной безопасности, а также для решения задач координации в многоуровневых распределенных системах. Модель сможет найти применение при создании методологии математического и компьютерного моделирования многоуровневых иерархических систем управления комплексной безопасностью сложных систем.

Литература

1. Маслобоев А.В. Реализация трансграничных ИТ-проектов в сфере информационного обеспечения комплексной безопасности развития арктических регионов: состояние и перспективы // Информационные ресурсы России. 2014. № 3(139). С. 13–20.
2. Ямалов И.У. Моделирование процессов управления и принятия решений в условиях чрезвычайных ситуаций. М.: БИНОМ. Лаборатория знаний, 2007. 288 с.
3. Маслобоев А.В. Виртуальные когнитивные центры как интеллектуальные системы для информационной поддержки управления региональной безопасностью // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 2 (90). С. 167–170.
4. Zlatanova S., Peters R., Dilo A., Scholten H. Intelligent Systems for Crisis Management. Springer, 2013. 500 p.
5. Маслобоев А.В. Метод комплексной оценки и анализа глобальной безопасности региональных социально-экономических систем на основе когнитивного моделирования // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 5 (87). С. 154–164.

6. Sarbazi-Azad H., Zomaya A.Y. Large Scale Network-Centric Distributed Systems. John Wiley & Sons, 2013. 700 p.
7. Oleynik A., Putilov V. The conceptual modeling for information support of regional management // Applied Information Technology Research. Rovaniemi, Finland, 2007. P. 9–21.
8. Игнатьев М.Б., Путилов В.А., Смольков Г.Я. Модели и системы управления комплексными экспериментальными исследованиями. М.: Наука, 1986. 232 с.
9. Алексеев В.В., Богатиков В.Н., Палюх Б.В. Приложения метода разделения состояний к управлению технологической безопасностью на основе индекса безопасности. Тверь: ТГТУ, 2009. 398 с.
10. Информационная безопасность систем организационного управления. Теоретические основы. В 2 томах. / Под ред. Н.А. Кузнецов, В.В. Кульбы. М.: Наука, 2006. Т. 1. 496 с.
11. Бурков В.Н., Коргин Н.А., Новиков Д.А. Введение в теорию управления организационными системами. М.: Либроком, 2009. 264 с.
12. Larichev O.I., Petrovsky A.B. Decision support systems for illstructured problems: requirements and constraints // Organizational Decision Support Systems. Amsterdam, North-Holland, 1988. P. 247–257.
13. Кузьмин И.А., Путилов В.А., Фильчаков В.В. Распределенная обработка информации в научных исследованиях. Л.: Наука, 1991. 304 с.
14. Mesarovic M.D., Macko D., Takahara Y. Theory of Hierarchical Multilevel Systems. NY-London: Academic Press, 1970. 294 p.
15. Шориков А.Ф. Методология моделирования многоуровневых систем: иерархия и динамика // Прикладная информатика. 2006. № 1. С. 136–141.
16. Mesarovic M.D., Takahara Y. General Systems Theory: Mathematical Foundations. NY: Academic Press, 1975. 279 p.

Маслобоев Андрей Владимирович – кандидат технических наук, доцент, старший научный сотрудник, Институт информатики и математического моделирования технологических процессов Кольского научного центра РАН, Апатиты Мурманской обл., 184209, Российская Федерация; заведующий кафедрой, Кольский филиал Петрозаводского государственного университета, Апатиты Мурманской обл., 184209, Российская Федерация, masloboev@iimm.ru

Путилов Владимир Александрович – доктор технических наук, профессор, директор, Институт информатики и математического моделирования технологических процессов Кольского научного центра РАН, Апатиты Мурманской обл., 184209, Российская Федерация; директор, Кольский филиал Петрозаводского государственного университета, Апатиты Мурманской обл., 184209, Российская Федерация, putilov@iimm.ru

Сютин Алексей Викторович – кандидат технических наук, научный сотрудник, лаборатория системной динамики, факультет географии, Университет Бергена, Берген, N-5020, Норвегия; младший научный сотрудник, Институт информатики и математического моделирования технологических процессов Кольского научного центра РАН, Апатиты Мурманской обл., 184209, Российская Федерация, alexei.sioutine@geog.uib.no

Andrey V. Masloboev – PhD, Associate professor, senior research fellow, Institute for Informatics and Mathematical Modeling of Technological Processes Kola Science Center of the Russian Academy of Sciences, Apatity, 184209, Russian Federation; Department head, Kola Branch of Petrozavodsk State University, Apatity, 184209, Russian Federation, masloboev@iimm.ru

Vladimir A. Putilov – D.Sc., Professor, Director, Institute for Informatics and Mathematical Modeling of Technological Processes Kola Science Center of the Russian Academy of Sciences; Director, Kola Branch of Petrozavodsk State University, Apatity, 184209, Russian Federation

Alexei V. Sioutine – PhD, research fellow, University of Bergen, Bergen, N-5020, Norway; junior research fellow, Institute for Informatics and Mathematical Modeling of Technological Processes Kola Science Center of the Russian Academy of Sciences, Apatity, 184209, Russian Federation, alexei.sioutine@geog.uib.no

Принято к печати 26.09.14
Accepted 26.09.14

УДК 621.83.621.81.002.2

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПОГРЕШНОСТЕЙ ИЗГОТОВЛЕНИЯ ЭЛЕМЕНТОВ ЦЕВОЧНОЙ ПЕРЕДАЧИ ПЛАНЕТАРНОГО РЕДУКТОРА

И.М. Егоров[a], С.А. Алексанин[a], М.Е. Федосовский[a], Н.П. Кряжева[b]

[a] ЗАО «Диаконт», Санкт-Петербург, 195274, Российская Федерация, egrov@mail.ru

[b] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

Аннотация. Теоретические основы расчета планетарных цевочных редукторов типа *k-h-v* заложены сравнительно давно. Однако в последнее время к вопросам их проектирования вновь привлечено повышенное внимание. Это связано с тем, что подобные устройства входят во многие сложные технические системы, в частности, в мехатронные и робототехнические системы. Развитие современной технологической базы производства таких редукторов сегодня позволяет реализовать принципиальные возможности этих устройств – высокий коэффициент полезного действия, большое передаточное отношение, кинематическую точность и плавность хода. Наличие адекватной математической модели позволяет управлять кинематической точностью редуктора за счет рационального выбора допусков на изготовление его деталей. Появляется возможность автоматизировать процесс проектирования планетарных цевочных редукторов с учетом различных, в том числе технологических, факторов. В настоящей работе авторами разработана математическая модель и предложен математический аппарат, позволяющие моделировать кинематическую погрешность редуктора с учетом многочисленных факторов, учитывающих, в том числе, и погрешности изготовления. Погрешности рассматриваются в виде, удобном для прогнозирования кинематической точности редуктора уже на стадии изготовления по результатам измерения деталей на координатно-измерительных машинах. При моделировании погрешностей изготовления колеса они задаются отклонением радиуса окружности центров цевок, ее эксцентриситетом и отклонениями положений осей цевок относительно окружности центров. Погрешности изготовления сателлита задаются отклонением его эксцентриситета и эксцентриситетом зубчатого венца. Вследствие коллинеарности, погрешности диаметров цевок, отверстий под цевки и погрешности профилей зубьев сателлита для заданной точки контакта объединены в одном отклонении. Программная реализация модели позволяет оценить влияние указанных погрешностей на ошибку угла поворота сателлита и обоснованно выбирать точностные параметры технологических процессов изготовления деталей. Кроме того, она дает возможность по результатам измерения на координатно-измерительной машине оценить кинематическую погрешность редуктора или путем анализа кинематограммы редуктора диагностировать погрешности изготовления его деталей. Модель реализована в виде программы, разработанной в среде Microsoft Visual C++ 6.0. Полученные результаты нашли применение в системе автоматизированного проектирования планетарных цевочных редукторов.

Ключевые слова: математическое моделирование, планетарный редуктор, цевочная передача, циклоидальное зацепление, кинематическая точность, погрешность изготовления.

Благодарности. Работа подготовлена по результатам НИОКРТ «Создание высокотехнологичного производства прецизионных быстродействующих силовых электромеханических приводов нового поколения» в Университете ИТМО, при финансовой поддержке Министерства образования и науки Российской Федерации согласно постановлению Правительства Российской Федерации от 9 апреля 2010 г. № 218 «О мерах государственной поддержки развития кооперации российских высших учебных заведений, государственных научных учреждений и организаций, реализующих комплексные проекты по созданию высокотехнологичного производства».

MODELING OF MANUFACTURING ERRORS FOR PIN-GEAR ELEMENTS OF PLANETARY GEARBOX

I.M. Egorov[a], S.A. Aleksanin[a], M.E. Fedosovskiy[a], N.P. Kryazheva[b]

[a] "Diakont", JSC, Saint Petersburg, 195274, Russian Federation, egrov@mail.ru

[b] ITMO University, Saint Petersburg, 197101, Russian Federation

Abstract. Theoretical background for calculation of *k-h-v* type cycloid reducers was developed relatively long ago. However, recently the matters of cycloid reducer design again attracted heightened attention. The reason for that is that such devices are used in many complex engineering systems, particularly, in mechatronic and robotics systems. The development of advanced technological capabilities for manufacturing of such reducers today gives the possibility for implementation of essential features of such devices: high efficiency, high gear ratio, kinematic accuracy and smooth motion. The presence of an adequate mathematical model gives the possibility for adjusting kinematic accuracy of the reducer by rational selection of manufacturing tolerances for its parts. This makes it possible to automate the design process for cycloid reducers with account of various factors including technological ones. A mathematical model and mathematical technique have been developed giving the possibility for modeling the kinematic error of the reducer with account of multiple factors, including manufacturing errors. The errors are considered in the way convenient for prediction of kinematic accuracy early at the manufacturing stage according to the results of reducer parts measurement on coordinate measuring machines. During the modeling, the wheel manufacturing errors are determined by the eccentricity and radius deviation of the pin tooth centers circle, and the deviation between the pin tooth axes positions and the centers circle. The satellite manufacturing errors are determined by the satellite eccentricity deviation and the satellite rim eccentricity. Due to the collinearity, the pin tooth and pin tooth hole diameter errors and the satellite tooth profile errors for a designated contact point are integrated into one deviation. Software implementation of the model makes it possible to estimate the pointed errors influence on satellite rotation angle error and reasonable selection of accuracy parameters for technological processes related to reducer parts manufacture. Additionally, it gives the possibility for estimation of the reducer kinematic error according to measurements by means of a coordinate measuring machine and diagnostics of reducer parts manufacturing errors by means of its

kinematogram analysis. The model is implemented as a program developed in Microsoft Visual C++ 6.0 environment. Obtained results have found their application in CAD of cycloid reducers.
Keywords: modeling, planetary gearbox, pin-gear drive, cycloid gear, kinematic accuracy, manufacturing error.
Acknowledgements. The paper has been prepared as a result of R&D work "Creating of high-tech production of precision high-performance forceful brand-new electromechanical actuators " in ITMO University, under financial support from the Russian Federation Ministry of Education and Science, according to the enactment of the Russian Federation Government dated April 9, 2010 № 218 "Measures of state support for development of cooperation between Russian universities, research organizations and companies which implement complex projects for high-tech production".

Введение

Планетарные цевочные редукторы (ПЦР) типа *k*-*h*-*v* (рис. 1) широко применяются в составе приводов мехатронных и робототехнических систем, к которым предъявляются повышенные требования по кинематической точности и жесткости.

В наше время появилась технологическая возможность реализации ПЦР, которые были теоретически разработаны еще в 1950–1980 годы. ПЦР обладают высоким коэффициентом полезного действия (до 0,95), высокой нагрузочной способностью, высокой адаптацией к условиям решения специальных задач; обеспечивают большие передаточные отношения в одной ступени (до 191), плавность работы, отсутствие мертвого хода, вибраций и высокую точность при значительном передаваемом моменте. Такие параметры обеспечили этим редукторам высокую применяемость в машиностроительных отраслях развитых стран мира. Циклоидальное зацепление обладает бóльшим КПД, чем традиционное эвольвентное, и при этом позволяет достигать в 7 раз большего передаточного отношения в одной ступени. Удельная масса редукторов планетарных редукторов с циклоидальным зацеплением меньше аналогичного показателя эвольвентных планетарных редукторов в 2–7 раз, что позволяет либо уменьшать габариты редуктора при одинаковой нагрузочной способности, либо увеличивать нагрузочную способность при тех же габаритах.

Использование в ПЦР циклоидального зацепления позволяет уменьшить разницу чисел цевок колеса и зубьев сателлита до единицы. При этом редуктор имеет передаточное число, равное числу зубьев сателлита, а коэффициент перекрытия теоретически превышает половину этой величины. Большое число одновременно зацепляющихся зубьев и различие длин профильных нормалей в зацеплениях существенным образом сказываются на характере зависимости кинематической погрешности редуктора от погрешностей изготовления элементов цевочной передачи. Именно такие редукторы получили наибольшее распространение.

Общая теория циклоидального зацепления изложена в работах [1, 2]. Ее применение при расчете ПЦР наиболее полно отражено в работах [3, 4]. Исследованию геометрии и нагрузочной способности ПЦР посвящены работы [3–6]. Общие методы расчета точности механизмов приведены в работе [7]. Точность механизмов с высшими кинематическими парами рассматривается в работах [8, 9]. Расчет ряда точностных параметров ПЦР приведен в работах [10–15].

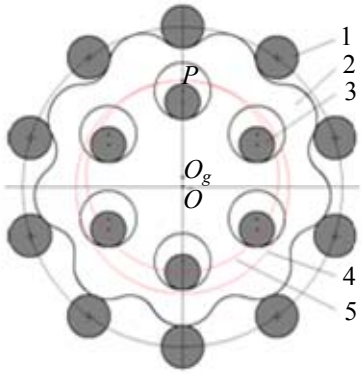


Рис. 1. Планетарный редуктор с цевочным зацеплением: 1 – цевка; 2 – сателлит; 3 – палец выходного вала; 4 – центроида колеса; 5 – центроида сателлита; O – центральная ось; O_g – ось сателлита; P – полюс зацепления

Первичные ошибки зацепления

Первичные ошибки зацепления *i*-ой цевки с зубом сателлита показаны на рис. 2.

Погрешности положения цевок колеса задаются отклонением радиуса окружности центров цевок, ее эксцентриситетом и отклонениями положений центров цевок относительно окружности центров. Окружность центров цевок определяется как окружность, для которой отклонения положений центров отверстий под цевки имеют минимальные значения. Такое представление погрешностей положения цевок колеса упрощает оценку влияния точностных параметров технологических процессов изготовления деталей на кинематическую погрешность редуктора и ее прогнозирование на основании результатов измерения на координатно-измерительной машине.

В неподвижной система координат *XOY* орт профильной нормали и рассматриваемые векторы первичных ошибок (рис. 2) имеют следующее представление:

$$\mathbf{e}_i = \begin{bmatrix} -\cos(\varphi_1 - \alpha_i) \\ \sin(\varphi_1 - \alpha_i) \end{bmatrix}; \quad d\mathbf{E} = dE \begin{bmatrix} \sin\varphi_1 \\ \cos\varphi_1 \end{bmatrix}; \quad \mathbf{E}_g = E \begin{bmatrix} \sin(\theta_g - \varphi_2) \\ \cos(\theta_g - \varphi_2) \end{bmatrix}; \quad \mathbf{E}_b = E_b \begin{bmatrix} \sin\theta_b \\ \cos\theta_b \end{bmatrix};$$

$$d\mathbf{R}_i = dR_i \begin{bmatrix} -\cos(\varphi_1 - \alpha_i) \\ \sin(\varphi_1 - \alpha_i) \end{bmatrix}; \quad d\mathbf{R}_{bi} = dR_b \begin{bmatrix} \sin(\varphi_1 + \psi_i) \\ \cos(\varphi_1 + \psi_i) \end{bmatrix}; \quad d\mathbf{x}_i = dx_i \begin{bmatrix} 1 \\ 0 \end{bmatrix}; \quad d\mathbf{y}_i = dy_i \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{1}$$

Угол давления $\alpha_i = \operatorname{arctg}\left(\dfrac{r_c \cdot \cos\psi_i - r_b}{r_c \cdot \sin\psi_i} \right)$.

Угол поворота выходного вала $\varphi_2 = \dfrac{\varphi_1}{u}$, где u – передаточное число редуктора.

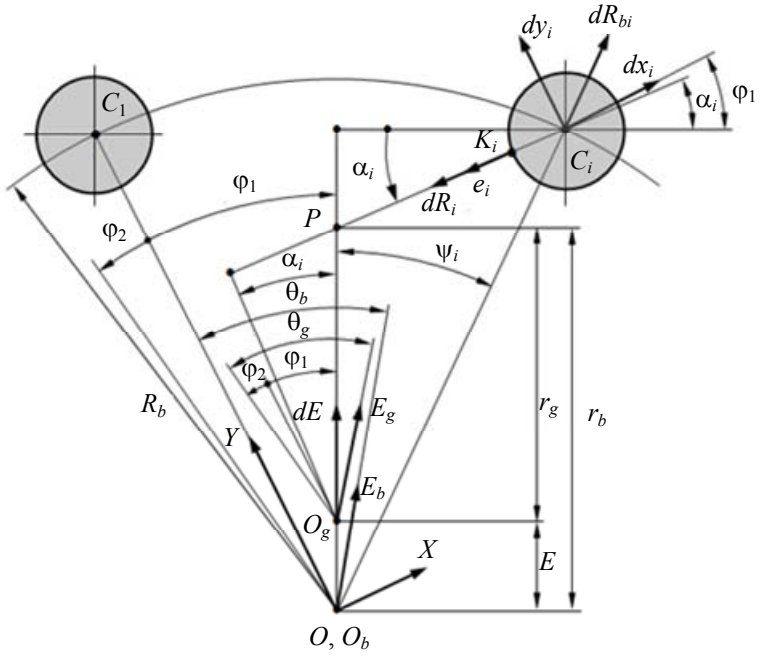


Рис. 2. Первичные ошибки зацепления *i*-ой цевки с зубом сателлита: *XOY* – неподвижная система координат, жестко связанная с корпусом редуктора; *O* – центральная ось; O_b – ось колеса; O_g – ось сателлита; *P* – полюс зацепления; φ_1 – угол поворота входного вала; φ_2 – угол поворота выходного вала; ψ_i – угловое положение *i*-ой цевки; r_b, r_g – радиусы центроид колеса и сателлита; K_i – точка контакта; $\mathbf{e}_i$ – орт профильной нормали; α_i – угол давления; h_i – плечо профильной нормали относительно оси сателлита; E, $d\mathbf{E}$ – эксцентриситет сателлита и его отклонение; $\mathbf{E}_g$, ϑ_g – эксцентриситет зубчатого венца сателлита и его фаза; $\mathbf{E}_b$, ϑ_b – эксцентриситет колеса и его фаза; R_b, $d\mathbf{R}_{bi}$ – радиус окружности центров цевок (радиус колеса) и его отклонение; $d\mathbf{x}_i$, $d\mathbf{y}_i$ – отклонения координат центра цевки; $d\mathbf{R}_i$ – отклонение радиуса цевки

Вектор отклонения, связанного с погрешностью профиля зуба сателлита, направлен вдоль профильной нормали и коллинеарен вектору отклонения радиуса цевки $d\mathbf{R}_i$. В связи с этим погрешность профиля f_i может быть включена в состав вектора $d\mathbf{R}_i$. Аналогично может быть учтена первичная ошибка, обусловленная зазором Δ_i в посадке цевки в отверстие корпуса редуктора:

$$d\mathbf{R}_i = \left(dR_i + f_i - \Delta_i\right) \cdot \begin{bmatrix} -\cos(\varphi_1 - \alpha_i) \\ \sin(\varphi_1 - \alpha_i) \end{bmatrix}.$$

Угловое положение *i*-ой цевки определяется по формуле $\psi_i = \tau_b \cdot (i-1) - \varphi_1$, где τ_b – угловой шаг цевок колеса. При заданном угле поворота входного вала в зацеплении находятся только цевки, расположенные по одну сторону от прямой *OP*. На рис. 2 это правая сторона. При разнице чисел цевок колеса и зубьев сателлита, равной единице, в зацеплении находятся все цевки, для которых $\psi_i \in [0, \pi]$.

Ошибка угла поворота сателлита

При определении ошибки угла поворота сателлита использована методика расчета погрешности механизма с высшей кинематической парой, изложенная в работе [8]. В соответствии с этой методикой,

ошибка положения звена, которое совершает вращательное движение, равна частному от деления суммы проекций векторов первичных ошибок на профильную нормаль на ее плечо относительно оси вращения звена:

$$\Delta\varphi = \frac{\mathbf{\Delta}\cdot\mathbf{e}}{h}, \qquad (2)$$

где $\mathbf{\Delta}$ – сумма векторов первичных ошибок; $\mathbf{e}$, h – орт и плечо профильной нормали. При выбранном на рис. 2 направлении орта нормали векторы первичных ошибок, относящихся к сателлиту, суммируются со знаком минус.

В соответствии с (1) и (2) для заданного угла поворота входного вала ошибка угла поворота сателлита, при зацеплении с i-ой цевкой определяется по формуле

$$\Delta\varphi_{gi}(\varphi_1) = \frac{\left(\mathbf{E}_b + d\mathbf{R}_i + d\mathbf{R}_{bi} + d\mathbf{x}_i + d\mathbf{y}_i - d\mathbf{E} - \mathbf{E}_g\right)\cdot\mathbf{e}_i}{h_i}; \quad h_i = r_g\cdot\cos\alpha_i .$$

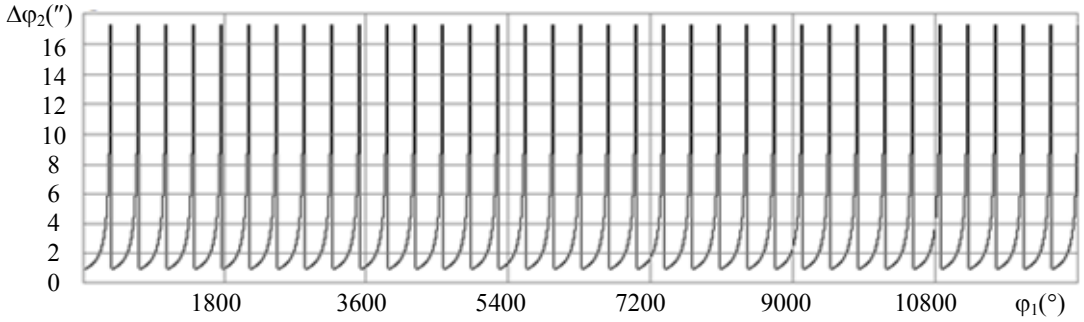
Максимальное значение $\Delta\varphi_{gi}(\varphi_1)$ для всех n цевок, находящихся в одновременном зацеплении, определяет ошибку угла поворота сателлита:

$$\Delta\varphi_g(\varphi_1) = \max\left\{\Delta\varphi_{gi}(\varphi_1)\right\}_{i=1}^{i=n}.$$

Следует отметить, что разность между максимальным и минимальным значениями ошибок угла поворота $\Delta\varphi_{gi}(\varphi_1)$ может привести к заклиниванию зубьев сателлита между цевками колеса даже при отсутствии погрешностей при передаче движения в противоположном направлении. Этот эффект является положительным: он создает предварительный натяг в зацеплении и может использоваться для выборки кинематического и снижения упругого мертвого хода в передаче.

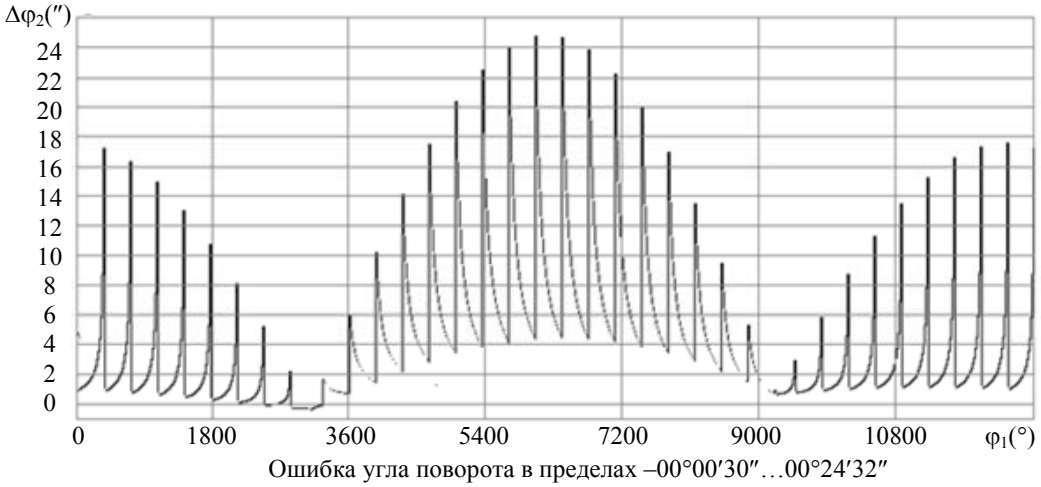
Программная реализация модели и результаты расчетов

Приведенная модель реализована в виде программы, разработанной в среде Microsoft Visual C++ 6.0, которая позволяет получить функции ошибки поворота сателлита при различных первичных ошибках в цевочной передаче.



Ошибка угла поворота в пределах 00°00′50″…00°17′22″

Рис. 3. Ошибки угла поворота сателлита при отклонении эксцентриситета сателлита +5 мкм



Ошибка угла поворота в пределах –00°00′30″…00°24′32″

Рис. 4. Ошибки угла поворота сателлита при эксцентриситете зубчатого венца сателлита 5 мкм

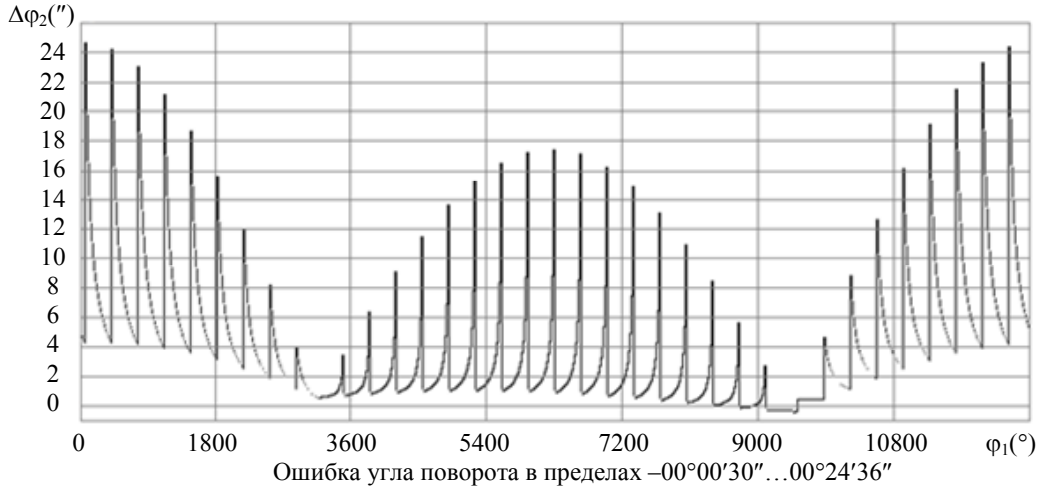


Рис. 5. Ошибки угла поворота сателлита при эксцентриситете колеса 5 мкм

На рис. 3–5 приведены результаты расчетов функции ошибки угла поворота сателлита на одном обороте выходного вала для передачи со следующими параметрами: передаточное число $u = 35$; $R_b = 100$ мм; $r_b = 70$ мм; $E = 0{,}972$ мм; $z_b = 36$; $z_g = 35$; диаметр цевок равен 5 мм.

Результаты расчетов показали, что наибольшее влияние на колебание ошибки угла поворота сателлита оказывают эксцентриситет колеса и эксцентриситет зубчатого венца сателлита. При этом функции ошибки имеют сходный характер. За счет большого числа одновременно зацепляющихся зубьев увеличение радиуса окружности центров цевок приводит к возникновению только постоянной составляющей ошибки угла поворота сателлита, равной –22" при $dR_b = +5$ мкм.

Заключение

Погрешности элементов цевочной передачи оказывают различное влияние на ошибку угла поворота сателлита. Приведенная методика расчета позволяет обоснованно выбрать точностные параметры технологических процессов изготовления деталей редуктора. Она также дает возможность по результатам измерения на координатно-измерительной машине прогнозировать кинематическую погрешность редуктора и диагностировать ошибки изготовления деталей на основании анализа его кинематограммы.

Дальнейшее усовершенствование модели должно идти в направлении учета упругих деформаций звеньев цевочной передачи при отсутствии и наличии вращающего момента на выходном валу редуктора с учетом возможности двухпрофильного контакта при заклинивании зубьев сателлита между цевками колеса. Решение указанной задачи позволит уточнить значение ошибки угла поворота выходного вала редуктора и получить зависимость упругого мертвого хода от передаваемого момента и величины нормальных зазоров в зацеплениях, вызванных погрешностями изготовления его деталей. Полученные результаты позволят уточнить распределение нагрузки в зацеплениях при определении контактных напряжений.

Особый интерес представляет применение модели для проведения статистических экспериментов при различных возможных сочетаниях погрешностей изготовления деталей.

Полученные результаты применены при разработке системы автоматизированного проектирования планетарных цевочных редукторов с циклоидальным зацеплением.

Литература

1. Литвин Ф.Л. Теория зубчатых зацеплений. М.: Наука, 1968. 584 с.
2. Litvin F.L. Gear Geometry and Applied Theory. 2nd ed. Cambridge University Press, 2004. 800 p.
3. Кудрявцев В.Н. Планетарные передачи. М.-Л.: Машиностроение, 1966. 307 с.
4. Шанников В.М. Теория и конструирование редукторов с внецентренным циклоидальным зацеплением. В кн.: Зубчатые и червячные передачи. М.-Л.: Машгиз, 1959. С. 74–109.
5. Dascalescu A. Contribution to the Kinematics and Dynamics Studys of the Planetary Gears with Cycloid Toothing and Roller Teeth, PhD Theses. Cluj-Napoca, Romania, 2005. 9 p.
6. Fedosovskii M.E., Aleksanin S.A., Nikolaev V.V., Yegorov I.M., Dunaev V.I., Puctozerov R.V. The effect of a cycloid reducer geometry on its loading capacity // World Applied Sciences Journal. 2013. V. 24. N 7. P. 895–899.
7. Бруевич Н.Г. Точность механизмов. М.-Л.: ГИТТЛ, 1946. 332 с.
8. Литвин Ф.Л. Проектирование механизмов и деталей приборов. Л.: Машиностроение, 1973. 695 с.

9. Сергеев В.И. Методологические основы повышения точности механизмов с высшими кинематическими парами // Проблемы машиностроения и надежности машин. 2006. № 1. С. 3–9.
10. Guan T.M. Calculation and analysis on the return error resulting from cycloid-disk modification in the cycloid drive // Modular Machine Tool and Automatic Manufacturing Technique. 2001. V. 10. P. 15–18.
11. Hidaka T., Wang H., Ishida T., Matsumota K., Hashimota M. Rotational transmission error of K-H-V planetary gears with cycloid gear // Transactions of the Japan Society of Mechanical Engineers Series C. 1994. V. 60. N 570. P. 645–653.
12. Li C., Liu J., Sun T. Study on transmission precision of cycloidal pin gear in 2K-V planetary drives // Chinese Journal of Mechanical Engineering. 2001. V. 37. N 4. P. 61–65.
13. Shirokoshi N., Hidaka T., Kasei S. Studies of influences of geometrical errors to final performances in small backlash planetary gears. Relations among position deviations of planet gears, target of backlash and non-working flank load // Transactions of the Japan Society of Mechanical Engineers Series C. 2000. V. 66. N 646. P. 1950–1958.
14. Sun Y.G., Zhao X.F., Jiang F., Zhao L, Liu D, Lu G.B. Backlash analysis of RV reducer based on error factor fensitivity and Monte-Carlo simulation // International Journal of Hybrid Information Technology. 2014. V. 7. N 2. P. 283–292.
15. Yang D.C.H., Blanche J.G. Design and application guidelines for cycloid drives with machining tolerances // Mechanism and Machine Theory. 1990. V. 25. N 5. P. 487–501.

Егоров Иван Михайлович – кандидат технических наук, старший научный сотрудник, ведущий научный сотрудник, ЗАО «Диаконт», Санкт-Петербург, 195274, Российская Федерация, egrov@mail.ru

Алексанин Сергей Андреевич – директор дивизиона систем управления и технологического оборудования, ЗАО «Диаконт», Санкт-Петербург, 195274, Российская Федерация,aleksanin@diakont.com

Федосовский Михаил Евгеньевич – кандидат технических наук, генеральный директор, ЗАО «Диаконт», Санкт-Петербург, 195274, Российская Федерация,diakont@diakont.com

Кряжева Наталья Петровна – аспирант, начальник лаборатории техногенной безопасности, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, kryazheva@diakont.com

Ivan M. Egorov – PhD, senior researcher, “Diakont”, JSC, Saint Petersburg, 195274, Russian Federation, egrov@mail.ru

Sergei A. Aleksanin – Technological Equipment & Control Systems Division Director, “Diakont”, JSC, Saint Petersburg, 195274, Russian Federation, aleksanin@diakont.com

Mikhail E. Fedosovskiy – PhD, General Manager, “Diakont”, JSC, Saint Petersburg, 195274, Russian Federation, diakont@diakont.com

Natalya P. Kryazheva – Technological Safety Laboratory Head, ITMO University, Saint Petersburg, 197101, Russian Federation, kryazheva@diakont.com

Принято к печати 29.09.14
Accepted 29.09.14

8 ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ОБРАЗОВАНИИ
INFORMATION TECHNOLOGIES IN EDUCATION

УДК 37:004

ПРОБЛЕМА ПОДДЕРЖКИ КОГНИТИВНЫХ ФУНКЦИЙ В ПРОЦЕССЕ ЭЛЕКТРОННОГО ОБУЧЕНИЯ

Л.С. Лисицына[a], А.В. Лямин[a], А.С. Быстрицкий[b], И.А. Мартынихин[a,c]

[a]Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, lisizina@mail.ifmo.ru

[b] Университет UCLA, Лос Анжелес, 90095,США

[c] Первый Санкт-Петербургский государственный медицинский университет им. И.П. Павлова, Санкт-Петербург, 197022, Российская Федерация

Аннотация. Успешность развития таких важнейших когнитивных функций человека, как внимание, восприятие и скорость обработки информации, рабочая и долговременная память, мышление и т.п., является необходимой основой повышения результативности электронного обучения. Одним из путей развития когнитивных функций обучаемых в процессе электронного обучения являются компьютерные когнитивные тренинги, которые включаются в индивидуальную траекторию обучения для стимулирования его к успешному выполнению определенных учебных задач электронного курса. Проведен анализ проблем оценивания эффектов когнитивных тренингов (выраженности, устойчивости и трансфера) и предложены пути их решения. Показано, что биологической основой эффектов когнитивных тренингов является нейропластичность головного мозга, которая влияет на продолжительность и интенсивность проведения тренингов. Предложен подход к организации исследований эффектов когнитивных тренингов, основанный на применении случайных методов. Показана перспективность использования игровых механик для реализации когнитивных тренингов в электронном обучении. Проведен детальный анализ подходов к тренингу базовых когнитивных функций, в том числе рабочей памяти обучаемых. Практическая значимость данной работы состоит в выявлении первоочередных задач для разработки и исследования когнитивных тренингов в электронном обучении.

Ключевые слова: электронное обучение, когнитивный тренинг, базовые и комплексные когнитивные функции, тренинг рабочей памяти, нейропластичность, оценки эффектов когнитивных тренингов.

SUPPORT PROBLEM FOR COGNITIVE FUNCTIONS IN THE E-LEARNING

L.S. Lisitsyna[a], A.V. Lyamin[a], A.S. Bystritsky[b], I.A. Martynikhin[a, c]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, lisizina@mail.ifmo.ru

[b]University UCLA, Los Angeles, 90095, USA

[c] Pavlov First Saint Petersburg State Medical University, Saint Petersburg, 197022, Russian Federation, iam@psychiatr.ru

Abstract. Successful development of such important human cognitive functions as attention, perception and information processing speed, working and long-term memory, thinking, etc. is a necessary foundation for increasing the effectiveness of e-learning. One way for further developments of students' cognitive functions in the process of e-learning consists in computer cognitive training sessions, which are included in the individual learning paths to promote a learner to the successful implementation of specific learning tasks of e-course. Analysis of the estimating problems for cognitive training effects (severity, stability and transfer) is done and the ways for their solution are proposed. It is shown that the biological basis for cognitive training effects consists in the processes of neuroplasticity of the brain that influence the duration and intensity of training. An approach to the organization of research for the effects of cognitive training, based on the usage of random methods is suggested. The prospects of game mechanics application for cognitive training implementation in e-learning are shown. A detailed analysis of the approaches to the training of the basic cognitive functions, including working memory of learners, is carried out. The practical significance of this paper is to identify priorities for research and development of cognitive training in e-learning.

Keywords: e-learning, cognitive training, basic and complex cognitive functions, working memory training, neuroplasticity, assessment of cognitive training effects.

Введение

В настоящее время существенно повысился интерес к технологиям и методикам электронного обучения. Это связано, прежде всего, с появлением технологии MOOC (Massive Open Online Courses) и созданием на ее основе открытых онлайн-курсов для электронного обучения [1]. В процессе электронного обучения перегруженность учебной информацией напрямую не контролируется преподавателем и может привести к стрессу у обучаемого, к общему снижению показателей его психолого-физиологического состояния, что, в свою очередь, ведет к падению результативности обучения [2–4]. Низкая результативность является существенным недостатком современного электронного обучения; сегодня на практике наблюдается не более 5% успешно завершивших онлайн-курс. Для этого имеется несколько причин. Кроме того, что онлайн-курс должен быть интересным и актуальным по своему содержанию, необходимы новые подходы к поддержанию когнитивных функций обучаемых в процессе электронного обучения.

Формирование и устойчивое развитие когнитивных (познавательных) способностей у человека в течение всей жизни – непременный элемент любого образовательного процесса. Успешность развития таких важнейших когнитивных функций человека, как внимание, восприятие и скорость обработки информации, рабочая и долговременная память, мышление и т.п., является необходимой основой для повышения результативности электронного обучения. Одним из путей развития когнитивных функций обучаемых в процессе электронного обучения являются компьютерные когнитивные тренинги, которые включаются в индивидуальную траекторию обучения для стимулирования его к успешному выполнению определенных учебных задач электронного курса. Когнитивные тренинги в настоящее время разрабатываются на стыке нескольких наук – психологии, медицины, педагогики и информатики. В литературе часто можно встретить и другие термины, относящиеся к когнитивным тренингам, например, тренинг мозга, фитнес мозга, когнитивное улучшение (cognitive enhancement), когнитивная ремедиация. Исходя из этого, в настоящей работе проводится детальный анализ опыта разработки и применения когнитивных тренингов, который адресуется, в первую очередь, ученым и IT-специалистам, занимающимся проблемами электронного обучения.

Технологии разработки когнитивных тренингов

При разработке когнитивных тренингов применяются два вида технологий [5]: сверху вниз (top-down) – для тренировки комплексных когнитивных процессов, улучшающих, в том числе, и базовые когнитивные функции; снизу вверх (bottom-up) – для тренировки базовых когнитивных функций с целью создания основы для развития более сложных и комплексных когнитивных процессов у человека. В реальных проектах когнитивных тренингов часто используется комбинация этих технологий для разработки целой батареи тренингов, направленных одновременно на широкий спектр психических процессов.

От когнитивных тренингов ожидают эффект максимальной устойчивости и способности к их генерализации, т.е. воздействия тренировок на смежные когнитивные процессы и показатели реальной жизни, в том числе и на академическую успеваемость обучаемых. Эти ожидания не всегда оправдываются, но уже сейчас привели к разработке и широкому распространению разнообразных компьютерных когнитивных тренингов на основе коммерческих программных продуктов, например, CogMed (www.cogmed.com), Lumosity (www.lumosity.com), Jungle Memory (junglememory.com), CogniFit (www.cognifit.com) и т.д. Разработчики таких программных продуктов ведут агрессивную маркетинговую политику, дают громкие обещания значимых эффектов от тренировок для своих пользователей, но качественных и независимых исследований эффективности многих тренингов пока нет [6]. Тем не менее, среди преимуществ компьютерных когнитивных тренингов можно выделить, прежде всего, массовость и индивидуальность проведения тренингов; возможность создавать сложные задания для тренингов, в том числе с использованием различных игровых механик; сокращение трудозатрат на сбор, обработку и визуализацию результатов тренингов; оперативный учет психолого-физиологических особенностей тренируемых. В связи с этим в ближайшее время стоит ожидать новые научные результаты именно в области разработки и использования компьютерных когнитивных тренингов, предназначенных для поддержки когнитивных функций в процессе электронного обучения.

Биологическая основа эффекта когнитивных тренингов

Значимость той роли, которую в настоящее время отводят когнитивным тренингам, во многом связана с тем, что с точки зрения современных нейронаук когнитивные тренинги являются не просто муштрой для выполнения того или иного конкретного задания, в основе их эффектов лежат структурные и функциональные изменения в головном мозге, обусловленные явлениями нейропластичности.

Нейропластичность – способность нейронов и нейронных сетей в мозге изменять связи и функции в ответ на изменение поведения человека и окружающей среды, например, при учебной нагрузке в процессе обучения, при восстановлении после повреждений и т.п. Нейропластичность как способность к изменению и развитию является важной функцией нервной системы человека. Прежде считалось, что, единожды сформировавшись в детстве, структуры мозга у взрослых остаются неизменными. Однако согласно данным современных исследований мозг взрослого человека также претерпевает значительные изменения в меняющихся условиях [7]. Если генетические факторы обусловливают медленные изменения в строении мозга человека на эволюционном уровне, то нейропластичность обусловливает возможность его быстрого приспособления к изменяющимся условиям. Способность к изменению (нейропластический потенциал) значимо различается у различных отделов нервной системы, психических функций, возрастов, отдельных людей (генетические особенности и прежние влияния окружающей среды) [8]. Исходя из этого, эффекты от когнитивных тренингов и их устойчивость могут различаться для тренингов, направленных на разные психические функции при использовании в разных возрастных группах тренируемых [5, 8].

В последние годы описаны явления быстрой нейропластичности, в первую очередь, за счет активации незадействованных ранее связей и модуляции синаптической передачи. Так, например, в одном из

исследований нейроанатомические эффекты от тренингов были зафиксированы уже после двух часов тренировок [9]. Однако основная часть структурных изменений в головном мозге происходит, как показывает практика, после нескольких недель тренировок [5].

Известно, что эффекты тренировок имеют две фазы [10]. Первая фаза характеризуется быстрым повышением производительности, которое наблюдается в пределах одной или нескольких тренировок. Вторая фаза предполагает медленный рост производительности в результате многих тренировок в течение длительного времени [10]. В основе каждой фазы лежит свой мозговой механизм. Например, в тренингах двигательных навыков начальная быстрая фаза связана с активацией мозжечка, а затем и с фронто-стриальной системой, в то время как во вторую фазу тренингов вовлечена двигательная кора мозга [11]. При этом образование новых синапсов (синаптогенез) и реорганизация моторных отделов коры происходит только во время второй, медленной фазы [12].

Основные нейропластические изменения в головном мозге происходят только после достаточно интенсивного формирования конкретных навыков, а длительные тренировки задействуют более стойкие и глобальные механизмы нейропластичности [5, 8]. Существуют и данные о том, что нейротрофический фактор мозга может быть периферическим маркером эффективности когнитивных тренингов [13]. Потому при выборе продолжительности и интенсивности тренингов их разработчики должны учитывать процессы нейропластичности. При проведении компьютерных когнитивных тренингов эта задача может решаться индивидуально для каждого тренируемого согласно мониторингу прогресса и достижения ожидаемого эффекта во время второй фазы. С другой стороны, нейропластичность может иметь и негативный эффект, она обусловливает забывание, утрату тех или иных навыков в условиях отсутствия тренировки, в основе чего также лежат изменения связей и функций нейронов и нейронных сетей мозга, а гипер- или гипонейропластичность могут приводить к формированию заболеваний нервной системы у человека [7]. В связи с этим с учетом нейропластичности головного мозга когнитивные тренинги призваны поддерживать когнитивные функции обучаемых на любом этапе электронного обучения.

Проблемы оценки эффективности когнитивных тренингов

Для оценки эффективности когнитивных тренингов можно выделить три основных параметра – выраженность, устойчивость и генерализация (трансфер) эффекта.

Первая проблема заключается в том, какие именно эффекты от тренингов можно ожидать и как их оценивать? Возможно, что прогресс, который наблюдается в процессе тренировок, является не более чем тренировкой справляться с конкретным заданием тренинга, а не непосредственной тренировкой того или иного когнитивного процесса? В определенной степени здесь может иметь место ситуация, которая хорошо известна, например, при тестировании уровеня интеллекта. Натренированность тестируемого решать задачи конкретного теста может способствовать повышению показателя интеллекта IQ, измеренного данным тестом, но необязательно каким-либо образом значимо влияет на сам интеллект. Этот факт привел к тому, что самому интеллекту стали давать ироничные определения: «Интеллект – это то, что измеряется тестами интеллекта» [14]. Выходом из данного положения может стать подход, когда после завершения основного задания тренинга следует применять другие средства, оценивающие те же самые когнитивные функции [6].

Вторая проблема связана с устойчивостью эффекта тренингов. Известно, что результаты, достигаемые во время тренингов, угасают в течение нескольких месяцев после прекращения тренировок [8, 15]. Учитывая свойство нейропластичности, для закрепления эффекта тренингов их продолжительность и интенсивность должна быть достаточной для формирования структурных и функциональных изменений в головном мозге. Только адекватная дозировка тренингов может давать устойчивый эффект. Для определения таких дозировок необходимы исследования тренировок разной длительности для каждого вида тренинга, а сейчас дозировка тренингов часто подбирается эмпирически [5, 15].

С практической точки зрения наиболее важной является проблема оценки влияния тренингов на смежные и (или) более широкие когнитивные сферы, так называемые показатели из реальной жизни (real-life outcomes). И здесь речь идет о возможности переноса и (или) генерализации (трансфера) эффекта. Идеальный когнитивный тренинг должен не только быть отработкой выполнения того или иного задания, но и способствовать улучшению всей когнитивной сферы, на которую он направлен, и, по возможности, сказываться на смежных и более интегральных областях. В этой связи выделяют ближний и дальний трансфер. Примером ближнего трансфера может быть улучшение показателей визуально-пространственной рабочей памяти при тренинге вербальной памяти [15]. Дальний трансфер – это перенос эффектов тренинга одной когнитивной сферы на другие, например, влияние тренинга рабочей памяти на внимание, мышление, клинические симптомы и т.п. Наибольший интерес представляет установление влияния тренингов на подвижный интеллект как общую основу интеллектуальных способностей мыслить логически, анализировать и решать новые задачи независимо от предыдущего опыта. Разумеется, что демонстрация эффектов дальнего трансфера при отсутствии ближнего не имеет смысла, так как тогда не ясно, по какой причине происходят изменения за пределами тренируемой области [6].

Важной проблемой является и сама организация проведения исследований по влиянию тренингов на когнитивные функции обучаемых. Только те исследования, в дизайне которых есть группа контроля и случайное распределение участников на активную и контрольную группы (рандомизация), можно считать адекватными для оценки эффективности любых вмешательств, в том числе и когнитивных тренингов [15, 16]. К сожалению, значительная часть исследований, продемонстрировавших значимые положительные эффекты тренингов, не имели адекватных групп контроля [15]. При этом даже в рандомизированных контролируемых исследованиях когнитивных тренингов чаще всего не использовался так называемый активный контроль, когда участникам из группы контроля не дают каких-либо заданий для тренинга [15]. Это может создавать преимущества для участников контрольной группы за счет так называемого эффекта Готорна [17], когда люди, вовлеченные в тренинг, склонны стараться лучше выполнять все выданные им задания из-за чувства собственной причастности к эксперименту, что может приводить к искажению результатов исследования.

Кроме того, значимые расхождения в оценках эффективности использования тренингов, которые можно отметить при сравнении рандомизированных контролируемых испытаний, свидетельствуют о чувствительности их результатов не только к общей методологии организации экспериментов, но и к более частным их особенностям. Например, S. Jaeggi [18] отмечает большое значение способов взаимодействия организаторов и участников эксперимента, психологических особенностей последних, мотивации (внутренней и внешней) к участию в тренингах, собственной оценки значимости и необходимости когнитивных тренингов, супервизии со стороны исследователей, качества инструкций, наличие обратной связи и т.д. Именно поэтому, чтобы нивелировать значение особенностей организации отдельных экспериментов для общей оценки эффективности когнитивных тренингов, следует применять метод мета-анализа с обобщением результатов всех доступных оригинальных исследований, посвященных одной проблеме [15].

Компьютерные игры в качестве когнитивных тренингов

Широкое распространение компьютерных игр вызывает большое число острых дискуссий, в которых обычно преобладают негативные оценки влияния компьютерных игр, их аддиктогенный потенциал, т.е. способность вызывать психическую зависимость [19, 20]. Лишь в последние годы появились работы, в которых анализируются позитивные последствия компьютерных игр, в первую очередь в отношении стимулирования к развитию когнитивных функций у игроков [21]. Обучение является эффективным, когда оно активно, основано на опыте, проблемно-ориентировано, включает немедленную обратную связь [22]. Компьютерные игры могут иметь все эти свойства, что вместе с их способностью увлекать игроков дает повод для оптимизма в отношении того, что они могут стать полезным и привлекательным методом обучения и развития когнитивных навыков [23]. Существуют убедительные данные о том, что по мере приобретения игрового опыта компьютерные игры обеспечивают некоторое положительное влияние на ряд когнитивных функций [24], но специально разработанные, персонифицированные компьютерные когнитивные тренинги оказываются более эффективными, чем обычные компьютерные игры [25]. В связи с этим в настоящее время значительная часть тренингов как для тренировки базовых когнитивных функций, так и для развития более сложных и комплексных когнитивных процессов, профессиональных навыков, создается в форме игр или с использованием различных игровых механик. При этом большое внимание уделяется разработке игр, специально предназначенных для образовательных целей, и так называемых «серьезных игр», т.е. игр, предназначенных для изменения поведения и отношения к чему-либо в более широком смысле. С результатами практического применения различных игровых механик в онлайн-курсе можно познакомиться, например, в работе [26].

Тренинг базовых когнитивных функций

В научных исследованиях популяциями особого интереса для тренингов базовых когнитивных функций стали подростки (тренинги для улучшения успеваемости в школе и вузе) и пожилые люди (тренинги для замедления возрастного угасания когнитивных способностей). Известно, что когнитивные тренинги, специально направленные на те или иные базовые когнитивные процессы, способны давать положительные эффекты у людей всех возрастов [27–29]. Важным аспектом, повышающим эффективность подобных тренингов, является и мультизадачность тренингов [30]. При этом, несмотря на то, что когнитивные тренинги в последнее время активно предлагаются на рынке и существует масса развивающих игр, созданных на основе тех или иных парадигм когнитивных тренингов, есть и серьезные сомнения в их эффективности за пределами тренируемой области.

В одном из крупнейших исследований эффективности когнитивных тренингов [27], в котором участвовало более 11 тысяч волонтеров в возрасте от 18 до 60 лет, случайным образом было сформировано 3 группы: участникам первой группы были предложены тренинги, направленные на решение логических задач, задач принятия решений и планирования; второй группы – тренинги краткосрочной памяти, внимания, математических вычислений; третьей группы – задания на поиск в Интернете ответов на

достаточно простые вопросы. Тренинги проводились не менее 10 мин 3 раза в неделю в течение 6 недель. Хотя в данном исследовании были получены положительные результаты для всех когнитивных процессов, на которые были направлены тренинги, трансфера эффектов установлено не было.

В мета-анализе исследований когнитивных тренингов у детей с синдромом дефицита внимания с гиперактивностью (СДВГ) [31] было выявлено значимое увеличение объема кратковременной памяти, но не было выявлено влияние тренингов на академическую успеваемость и оценку клинического состояния детей. В исследовании эффективности когнитивных тренингов на рабочем месте [32], когда офисным работникам в перерывах предлагались когнитивные тренинги по 20 мин в день 3 раза в неделю на протяжении 16 недель, было выявлено лишь незначительное улучшение когнитивных функций в активной группе. Зато неожиданно в группе контроля, которой в это же время давались для изучения различные материалы о природе, было отмечено большее субъективное улучшение качества жизни и общего самочувствия, снижение стресса.

В мета-анализе исследований эффективности когнитивных тренингов у здоровых пожилых лиц [28] также не был дан ответ об эффектах тренингов за пределами тренируемых функций. Здесь по результатам анализа 31 рандомизированного контролируемого испытания в группах испытуемых, прошедших когнитивные тренинги, отмечалось улучшение рабочей памяти, скорости обработки информации, воспроизведения лицо–имя и ряда других показателей. Участники тренингов субъективно положительно оценивали их эффект, но объективных данных о влиянии тренингов на повседневную деятельность, по мнению авторов этого мета-анализа, выявлено не было. В то же время в другом систематическом обзоре отмечается, что эффективность компьютерных когнитивных тренингов у пожилых не меньше или даже выше, чем у традиционных некомпьютеризированных вариантов [33].

Тренинги рабочей памяти

Областью особого интереса среди базовых когнитивных процессов является рабочая память. Рабочая память обеспечивает временное хранение и обработку информации, необходимой для решения сложных познавательных задач, что обусловливает связь показателей объема рабочей памяти с широким спектром реальных навыков [34], а также с процессами когнитивного развития и возникновения таких когнитивных расстройств, как расстройства чтения и счета, синдрома дефицита внимания и гиперактивности (СДВГ), дислексии и т.д. [35]. Известно, что рабочая память обусловливает успешность обучения в большей степени, чем показатели IQ [36]. Это связано с тем, что во время обучения учащимся часто приходится полагаться на рабочую память, чтобы выполнять различные задания – от запоминания инструкций преподавателя в отношении даже простых задач до хранения и обработки информации и отслеживания прогресса при решении трудных и многоэтапных заданий [35]. Исходя из этого, рабочую память рассматривают в качестве узкого места, «бутылочного горлышка» обучения [35, 36]. Так как обучение является длительным процессом, характеристики рабочей памяти могут оказывать существенное влияние на отдаленные результаты обучения. При этом есть убедительные данные, свидетельствующие о том, что емкость рабочей памяти может быть увеличена за счет тренировок [37, 38], что такие тренинги приводят к нейропластическим изменениям [39]. Но следует также отметить, что подтверждений воздействия тренингов рабочей памяти на результативность обучения пока недостаточно.

Популярность компьютерных тренингов рабочей памяти в качестве ключевой функции для обучения и общего развития человека росла с начала 2000-х годов, особенно после выхода работы Klingberg [40]. В это исследование были включены школьники с СДВГ. Им давались регулярные компьютерные тренинги рабочей памяти. Перед тренингами и через несколько недель после их начала была проведена оценка подвижного интеллекта, подтвердивших значимое улучшение его показателей. Затем такие же результаты были получены у молодых взрослых без СДВГ. Исследования были небольшими, но привлекли большое внимание специалистов из разных областей и масс-медиа. Их автор стал основателем коммерческой компании Cogmed, которая во многом является первопроходцем и законодателем мод на рынке компьютерных когнитивных тренингов, активно их рекламирует среди родителей школьников и учителей в качестве средства для улучшения когнитивных способностей [41], причем стоимость курса тренингов в США составляла в 2011 г. до 1500$ [6]. В 2008 г. S. Jaeggi [42], используя двойной *n*-обратный тест (dual *n*-back task), получила еще более впечатляющий результат, который сыграл значимую роль в популяризации когнитивных тренингов: тренировка рабочей памяти улучшала интеллект в весьма большой степени. Усредненно IQ повышался на половину балла после каждого часа тренинга, и это повышение имело дозозависимый эффект.

В дальнейшем гипотеза о том, что тренинги рабочей памяти оказывают значительное влияние на показатели интеллекта, стала подвергаться сомнению. Многие попытки других ученых воспроизвести полученные ранее результаты в исследованиях с более качественной методологией тренингов (с большими выборками, лучшим контролем и пр.) закончились неудачей. В мета-анализе эффективности тренингов рабочей памяти у детей и взрослых (как здоровых, так и с СДВГ), включающим результаты 23 исследований, было выявлено, что хотя данные тренинги и имели определенный (но не очень стой-

кий) эффект на объем рабочей памяти, достоверных данных о генерализации их эффектов не наблюдалось [15].

В связи с этим при разработке тренингов рабочей памяти следует учитывать, что не все известные парадигмы их тренингов одинаково эффективны. Например, компания Cogmed обещает, что тренинги рабочей памяти могут быть эффективны при условии использования правильных инструментов и протоколов, к которым относит, прежде всего, разработки своей компании [41]. В ответ на их экспансию на рынке школьного образования были проведены специальные исследования с использованием инструментов данной компании, в результатах которых имеются серьезные критические оценки, ставящие под сомнение эффективность таких тренингов [5]. А после критических замечаний [15, 43] в отношении исследований эффективности адаптивного двойного *n*-обратного теста [42] были проведены дополнительные исследования возможности дальнего трансфера, которые повторно подтвердили, что данный вариант тренинга рабочей памяти способствует улучшению показателей решения визуально-пространственных задач [18].

Таким образом, существующие в настоящее время исследования свидетельствуют о том, что когнитивные тренинги способны стимулировать развитие когнитивных функций у тренируемых, но остается открытым вопрос оценки их эффекта в отношении дальнего трансфера, в том числе для улучшения показателей подвижного интеллекта, академической успеваемости, адаптации к учебной нагрузке и профессиональной деятельности. И связано это не только со сложностью и многоаспектностью изучаемых явлений, но и, возможно, с недостаточной эффективностью существующих тренингов, что ставит задачу поиска новых решений. Одним из многообещающих направлений для ее решения, на наш взгляд, является разработка более эффективных заданий (парадигм) тренингов, исследование возможностей их сочетания, в том числе для потенцирования эффектов, обоснованного дозирования, генерализации и устойчивости их эффектов.

Заключение

Таким образом, когнитивные тренинги обязаны стать важными элементами процесса электронного обучения, направленными на поддержку когнитивных функций обучаемых на уровне, необходимом для успешного освоения электронного курса. Однако их разработка и применение в электронном обучении требует осторожности, о чем свидетельствует анализ опыта в нашей работе. Тем не менее, отметим ряд первоочередных задач, требующих своего решения в общей проблеме поддержки когнитивных функций в процессе электронного обучения. Разработка и проведение когнитивных тренингов должна основываться на явлениях нейропластичности головного мозга. Исходя из этого, их продолжительность и интенсивность должна соответствовать достижению второй фазы тренировок. Для оценивания когнитивных тренингов следует использовать контроль выраженности, устойчивости и трансфера их эффектов. При организации исследований эффектов когнитивных тренингов наиболее перспективными являются методы рандомизированного контролируемого исследования. Эффективность тренингов может быть существенно повышена за счет использования игровых механик в их реализации. Тренинги базовых когнитивных функций уже сейчас могут дать ощутимый эффект для повышения результативности электронного обучения.

Литература

1. Васильев В.Н., Стафеев С.К., Лисицына Л.С., Ольшевская А.В. От традиционного дистанционного обучения к массовым открытым онлайн-курсам // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 1 (89). С. 199–205.
2. Lisitsyna L., Lyamin A. Approach to development of effective e-learning courses // Frontiers in Artificial Intelligence and Application. 2014. V. 262. P. 732–738.
3. Lisitsyna L., Lyamin A., Skshidlevsky A. Estimation of student functional state in learning management system by heart rate variability method // Frontiers in Artificial Intelligence and Application. 2014. V. 262. P. 726–731.
4. Clark R.C., Nguyen F., Sweller J. Efficiency in Learning: Evidence-Based Guidelines to Manage Cognitive Load. San Francisco: Pfeiffer, 2006. 416 p.
5. Vinogradov S., Fisher M., de Villers-Sidani E. Cognitive training for impaired neural systems in neuropsychiatric illness // Neuropsychopharmacology. 2012. V. 37. N 1. P. 43–76.
6. Shipstead Z., Hicks K., Engle R. Cogmed working memory training: does the evidence support the claims? // Journal of Applied Research in Memory and Cognition. 2012. N 1. P. 185–193.
7. Pascual-Leone A., Freitas C., Oberman L., Horvath J.C., Halko M., Eldaief M., Bashir S., Vernet M., Shafi M., Westover B., Vahabzadeh-Hagh A.M., Rotenberg A. Characterizing brain cortical plasticity and network dynamics across the age-span in health and disease with TMS-EEG and TMS-fMRI // Brain Topography. 2011. V. 24. N 3–4. P. 302–315.

8. Fisher M., Holland C., Subramaniam K., Vinogradov S. Neuroplasticity-based cognitive training in schizophrenia: an interim report on the effects 6 months later // Schizophrtnia Bulletin. 2010. V. 36. N 4. P. 869–879.
9. Sagi Y., Tavor I., Hofstetter S., Tzur-Moryosef S., Blumenfeld-Katzir T., Assaf Y. Learning in the fast lane: new insights into neuroplasticity // Neuron. 2012. V. 73. N 6. P. 1195–1203.
10. Kleim J.A., Pipitone M.A., Czerlanis C., Greenough W.T. Structural stability within the lateral cerebellar nucleus of the rat following complex motor learning // Neurobiology of Learning and Memory. 1998. V. 69. N 3. P. 290–306.
11. Ungerleider L.G., Doyon J., Karni A. Imaging brain plasticity during motor skill learning // Neurobiology of Learning and Memory. 2002. V. 78. N 3. P. 553–564.
12. Kleim J.A., Hogg T.M., Vandenberg P.M., Cooper N.R., Bruneau R., Remple M. Cortical synaptogenesis and motor map reorganization occur during late, but not early, phase of motor skill learning // Journal of Neuroscience. 2004. V. 24. N 3. P. 628–633.
13. Vinogradov S., Fisher M., Holland C., Shelly W., Wolkowitz O., Mellon S.H. Is serum brain-derived neurotrophic factor a biomarker for cognitive enhancement in schizophrenia? // Biological Psychiatry. 2009. V. 66. N 6. P. 549–553.
14. Величковский Б.М. Когнитивная наука: основы психологии познания. Том 1. М.: Смысл, 2006. 448 с.
15. Melby-Lervag M., Hulme C. Is working memory training effective? A meta-analytic review // Developmental Psychology. 2013. V. 49. N 2. P. 270–291.
16. Власов В.В. Введение в доказательную медицину. М.: Медиа-Сфера, 2001. 392 с.
17. Хоторнский эффект – Википедия [Электронный ресурс]. Режим доступа: http://ru.wikipedia.org/wiki/Хоторнский_эффект, свободный. Яз. рус. (дата обращения 15.05.2014).
18. Jaeggi S.M., Buschkuehl M., Shah P., Jonides J. The role of individual differences in cognitive training and transfer // Memory and Cognition. 2013. V. 42. N 3. P. 464–480.
19. Sim T., Gentile D.A., Bricolo F., Serpollini G., Gulamoydeen F. A conceptual review of research on the pathological use of computers, video games, and the Internet // International Journal of Mental Health and Addiction. 2012. V. 10. N 5. P. 748–769.
20. Messias E., Castro J., Saini, A. Usman M., Peeples D. Sadness, suicide, and their association with video game and Internet overuse among teens: results from the Youth Risk Behavior Survey 2007 and 2009 // Suicide and Life-Threatening Behavior. 2011. V. 41. N 3. P. 307–315.
21. Granic I., Lobel A., Engels R.C.M.E. The benefits of playing video games // American Psychologist. 2014. V. 69. N 1. P. 66–78.
22. Boyle E.A., Connolly T.M., Hainey T. The role of psychology in understanding the impact of computer games // Entertainment Computing. 2011. V. 2. N 2. P. 69–74.
23. Connolly T.M., Boyle E.A., MacArthur E., Hainey T., Boyle J.M. A systematic literature review of empirical evidence on computer games and serious games // Computers and Education. 2012. V. 59. N 2. P. 661–686.
24. Boot W.R., Blakely D.P., Simons D.J. Do action video games improve perception and cognition? // Frontiers in Psychology. 2011. V. 2. Art. 226.
25. Peretz C., Korczyn A.D., Shatil E., Aharonson V., Birnboim S., Giladi N. Computer-based, personalized cognitive training versus classical computer games: a randomized double-blind prospective trial of cognitive stimulation // Neuroepidemiology. 2011. V. 36. N 2. P. 91–99.
26. Лисицына Л.С., Першин А.А., Усков В.Л. К вопросу повышения результативности массового онлайн-курса // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 5 (93). С. 164–171.
27. Owen A.M., Hampshire A., Grahn J.A., Stenton R., Dajani S., Burns A.S., Howard R.J., Ballard C.G. Putting brain training to the test // Nature. 2010. V. 465. N 7299. P. 775–778.
28. Kelly M.E., Loughrey D., Lawlor B.A., Robertson I.H., Walsh C., Brennan S. The impact of cognitive training and mental stimulation on cognitive and everyday functioning of healthy older adults: a systematic review and meta-analysis // Ageing Research Reviews. 2014. V. 15. P. 28–43.
29. Karch D., Albers L., Renner G., Lichtenauer N., von Kries R. The efficacy of cognitive training programs in children and adolescents – a meta-analysis // Deutsches Arzteblatt International. 2013. V. 110. N 39. P. 643−652.
30. Anguera J.A., Boccanfuso J., Rintoul J.L., Al-Hashimi O., Faraji F., Janowich J., Kong E., Larraburo Y., Rolle C., Johnston E., Gazzaley A. Video game training enhances cognitive control in older adults // Nature. 2013. V. 501. N 7465. P. 97–101.
31. Rapport M.D, Orban S.A., Kofler M.J., Friedman L.M. Do programs designed to train working memory, other executive functions, and attention benefit children with ADHD? A meta-analytic review of cognitive, academic, and behavioral outcomes // Clinical Psychology Review. 2013. V. 33. N 8. P. 1237–1252.
32. Borness C., Proudfoot J., Crawford J., Valenzuela M. Putting brain training to the test in the workplace: a randomized, blinded, multisite, active-controlled trial // PLoS ONE. 2013. V. 8. N 3. Art. e59982.

33. Kueider A.M., Parisi J.M., Gross A.L., Rebok G.W. Computerized cognitive training with older adults: a systematic review // PLoS One. 2012. V. 7. N 7. Art. e40588.
34. Unsworth N., Engle R.W. On the division of short-term and working memory: an examination of simple and complex span and their relation to higher order abilities // Psychological Bulletin. 2007. V. 133. N 6. P. 1038–1066.
35. Gathercole S.E., Alloway T.P. Short-term and working memory impairments in neurodevelopmental disorders: diagnosis and remedial support // Journal of Child Psychology and Psychiatry. 2006. V. 47. N 1. P. 4–15.
36. Alloway T.P., Alloway R.G. Investigating the predictive roles of working memory and IQ in academic attainment // Journal of Experimental Child Psychology. 2010. V. 106. N 1. P. 20–29.
37. Klingberg T. Training and plasticity of working memory // Trends in Cognitive Sciences. 2010. V. 14. N 7. P. 317–324.
38. Diamond A., Lee K. Interventions shown to aid executive function development in children 4 to 12 years old // Science. 2011. V. 333. N 6045. P. 959–964.
39. Takeuchi H., Sekiguchi A., Taki Y., Yokoyama S., Yomogida Y., Komuro N., Yamanouchi T., Suzuki S., Kawashima R. Training of working memory impacts structural connectivity // Journal of Neuroscience. 2010. V. 30. N 9. P. 3297–3303.
40. Klingberg T., Forssberg H., Westerberg H. Training of working memory in children with ADHD // Journal of Clinical and Experimental Neuropsychology. 2002. V. 24. N 6. P. 781–791.
41. Ralph K. Cogmed Working Memory Training. Pearson. Clinical Assessment [Электронный ресурс]. Режим доступа: http://www.pearsonclinical.co.uk/Cogmed/Downloads/cogmed-claims-and-evidence.pdf, свободный. Яз. англ. (дата обращения 15.05.2014).
42. Jaeggi S.M., Buschkuehl M., Jonides J., Perrig W.J. Improving fluid intelligence with training on working memory // Proc. of the National Academy of Sciences of the United States of America. 2008. V. 105. N 19. P. 6829–6833.
43. Kaufman S.B. New Cognitive Training Study Takes on the Critics [Электронный ресурс]. Режим доступа: http://blogs.scientificamerican.com/beautiful-minds/2013/10/09/new-cognitive-training-study-takes-on-the-critics/, свободный. Яз. англ. (дата обращения 15.05.2014).

Лисицына Любовь Сергеевна – доктор технических наук, профессор, заведующая кафедрой, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, lisizina@mail.ifmo.ru

Лямин Андрей Владимирович – кандидат технических наук, доцент, директор центра дистанционного обучения, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, lyamin@mail.ifmo.ru

Быстрицкий Александр Станиславович – M.D., Ph.D., профессор, Университет UCLA, Лос Анжелес, 90095, США, ABystritsky@mednet.ucla.edu

Мартынихин Иван Андреевич – кандидат медицинских наук, доцент кафедры, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; ассистент кафедры, Первый Санкт-Петербургский государственный медицинский университет им. И.П. Павлова, Санкт-Петербург, 197022, Российская Федерация, iam@psychiatr.ru

Liubov S. Lisitsyna – D.Sc., Professor, Department head, ITMO University, Saint Petersburg, 197101, Russian Federation, lisizina@mail.ifmo.ru

Andrei V. Lyamin – PhD, Associate professor, Director of the Center for Distance Learning, ITMO University, Saint Petersburg, 197101, Russian Federation, lyamin@mail.ifmo.ru

Alexander S. Bystritsky – M.D., PhD, Associate Professor, Professor of Psychiatry, University UCLA, Los Angeles, 90095, USA

Ivan A. Martynikhin – PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation; assistant, Pavlov First Saint Petersburg State Medical University, Saint Petersburg, 197022, Russian Federation, iam@psychiatr.ru

Принято к печати 30.06.14
Accepted 30.06.14

КРАТКИЕ СООБЩЕНИЯ
BRIEF PAPERS

УДК 535.55

УПРАВЛЕНИЕ МОДОВЫМ СОСТАВОМ ИЗЛУЧЕНИЯ НА ВЫХОДЕ ОПТИЧЕСКОГО ЖГУТА

Е.П. Конькова[a]

[a] ВолГУ, Волгоград, 400062, Российская Федерация, kon_ele@mail.ru

Аннотация. Рассмотрена возможность управления модовой структурой излучения в жгуте из оптических волокон путем варьирования его пространственной геометрии, изгибного профилирования. В отличие от известного дифракционного профилирования предложен подход, основанный на управлении оптическими неоднородностями волокна, вызванными изгибом волокна. На примере винтовой укладки жгута показана возможность изменения распределения интенсивности излучения на его выходе путем изменения параметров укладки: диаметра и шага намотки. Отмечено возникновение регулярных и спекл-полей с круглым профилем пучка. Обнаружено, что с увеличением радиуса и шага намотки увеличивается число мод в пучке. Также наблюдается плавное изменение (как уменьшение, так и увеличение) интенсивности от центра пучка к краям.

Ключевые слова: световые поля, спеклы, моделирование геометрии волокна, оптические неоднородности.

MODE COMPOSITION CONTROL ON OPTICAL TWISTED STRIP OUTPUT

E.P. Kon`kova[a]

[a]Volgograd State University, Volgograd, 400062, Russian Federation, kon_ele@mail.ru

Abstract. The paper deals with possibility of mode structure control for light scattering in a twisted strip fiber by variation its space geometry. Unlike the known diffraction profiling, an approach is proposed based on control of fiber optic discontinuities caused by fiber twist. On the example of spiral pilling of a twisted strip we show the possibility of distribution changing for light scattering intensity on its output by changing of pilling parameters: radius and winding step. Fiber geometry alteration leads to the alteration for a number of optical modes. The increase of a bending radius and winding step of a fiber leads to the growth of the modes number. Regular optical fields and speckles are registered within this work. Gradual intensity change is observed (both, decrease and increase) from the beam centre to its edges.

Keywords: optical fields, speckles, fiber geometry modeling, optical discontinuities.

Известно, что в изогнутом оптическом волокне распределение интенсивности излучения по поперечному сечению на выходе, длина пути и количество отражений отдельных лучей иные, чем для прямого волокна [1, 2]. Это явление, в частности, нашло применение в разработках датчиков физических величин [3–5]. Знание характера распределения интенсивности в лазерном пучке в плоскости, перпендикулярной направлению его распространения (профиль пучка), особо важно для всех промышленных применений лазеров [6]. Аналогичные явления, как можно предположить, наблюдаются и в жгуте, изготовленном из оптических волокон.

Рассмотрена возможность управления модовой структурой оптического жгута путем варьирования его пространственной геометрии – изгибного профилирования. В отличие от известного дифракционного профилирования, основанного на использовании дифракционных элементов, предложенный подход основан на управлении оптическими неоднородностями волокна в результате моделирования геометрии волокна. В качестве исходной выбрана винтовая геометрия жгута, используемая, в частности, в волоконно-оптическом гироскопе [7]. В результате плотной (2 витка на 1 см) намотки жгута диаметром 3 мм и длиной 1,5 м вокруг сердечника длиной 30 см и диаметром 8 мм (радиус изгиба сопоставим с диаметром жгута) в материал волокна по всей длине с постоянным шагом принудительно вносились оптические неоднородности. Использовался жгут из регулярно уложенных оптических волокон, обычно являющийся частью оптической системы передачи изображения гибкого эндоскопа, с установленной на выходном торце короткофокусной (фокусное расстояние менее 20 мм) собирающей линзой (рис. 1).

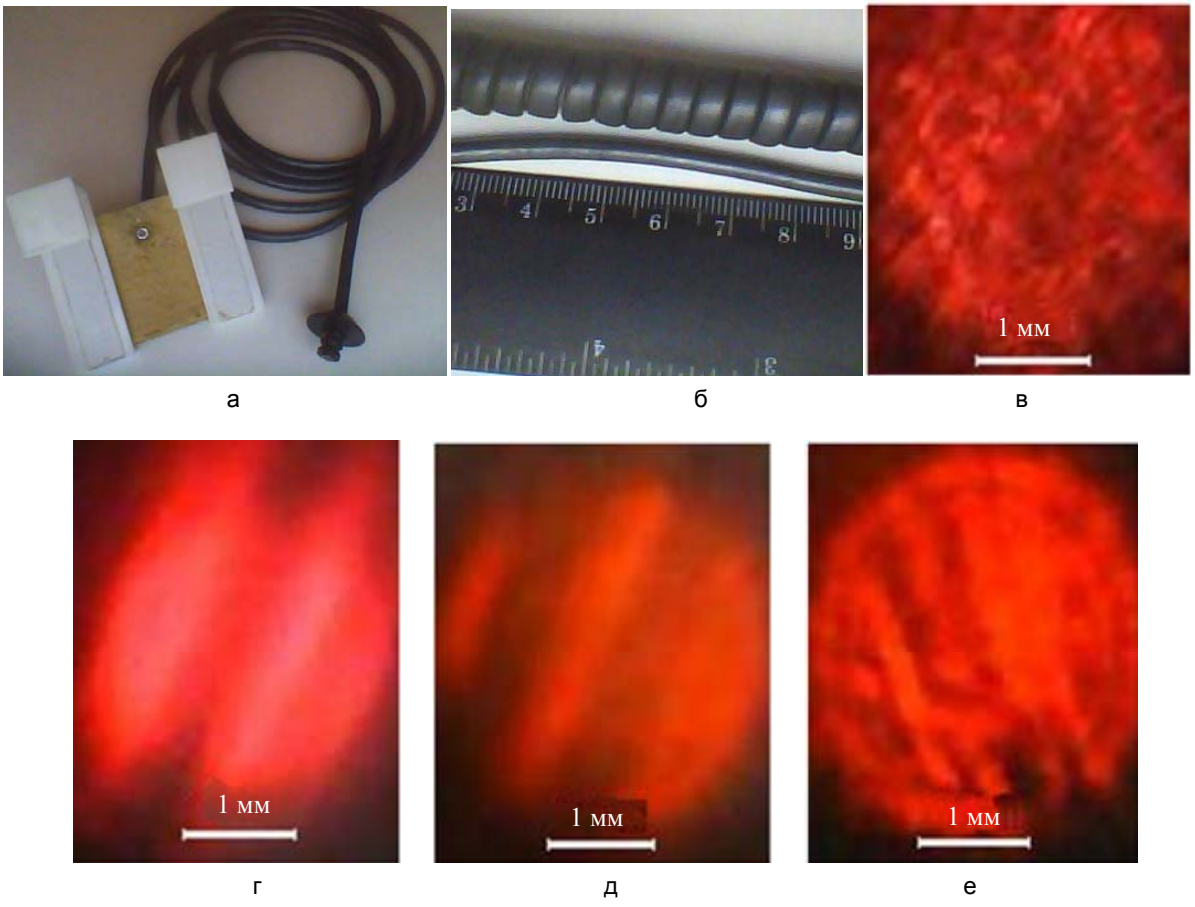
Известно, что световые потери в результате многократного изгиба жгута с радиусом до пяти его диаметров являются допустимыми. Исходя из этого, нами был выбран существенно меньший радиус изгиба, при котором световые потери, в отличие от потерь, связанных с натяжением намотки, оказывают значительное влияние на оптические свойства поверхности волокна.

В качестве источника излучения использовался полупроводниковый лазер (лазерный InGaAIP диод) с длиной волны 650 нм и мощностью излучения 5 мВт, работающий в режиме непрерывной генерации. Излучение фокусировалось (диаметр фокального пятна 0,5 мм) линзой на входной торец жгута под разными апертурными углами, что приводило к формированию регулярных и спекл-полей (рис. 1). Угол ввода излучения в жгут в данном случае не регистрировался, поскольку значение имел сам факт зависимости от него сформированного светового поля. Выходной торец жгута закреплялся вплотную к web-камере компьютера.

Отметим, что при отсутствии намотки зарегистрировать изображение на выходе жгута не представляется возможным из-за чрезмерной засветки web-камеры. Винтовая намотка жгута (рис. 1, б) приво-

дила к резкому возрастанию потерь, в результате чего камера переходила в рабочий режим. Картины распределения интенсивности излучения по поперечному сечению жгута диаметром 3 мм на его выходе при различных углах ввода излучения представлены в удобном для восприятия увеличенном виде. На рис. 1, в, показано образование спекл-поля. Рис. 1, г–е, иллюстрируют модовую структуру на выходе пучка при негауссовом распределении. Интенсивность полос равномерна по диаметру пучка.

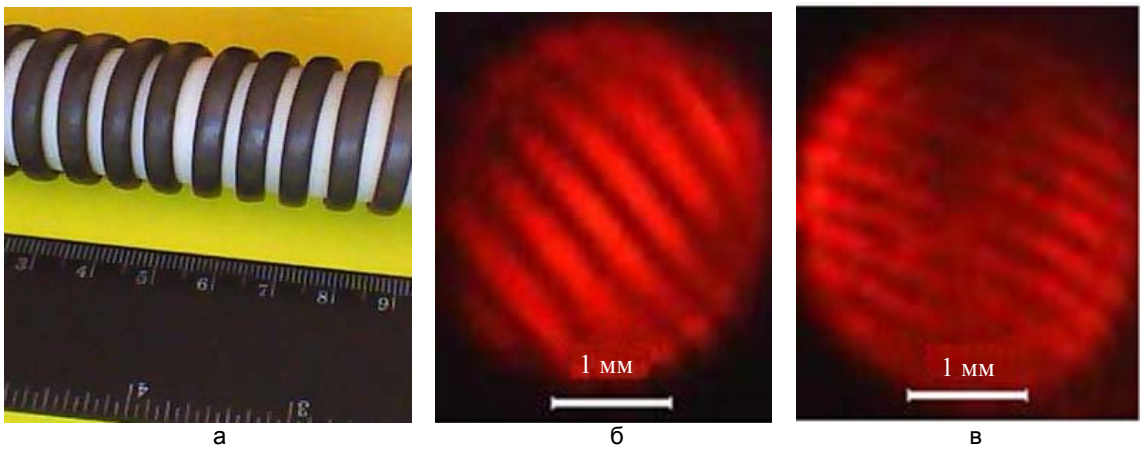
Автором показано, что модовый состав излучения на выходе оптического жгута меняется от диаметра намотки и ее шага. Пример такой намотки показан на рис. 2, а. С увеличением шага намотки одновременно увеличивается число мод в пучке (рис. 2, б, в), а интенсивность полос плавно меняется от центра пучка к его краям.



а б в

г д е

Рис. 1. Внешний вид исходного жгута с фиксатором (а) и жгут в виде плотной намотки (б). Зарегистрированные картины распределения интенсивности излучения по поперечному сечению пучка диаметром 3 мм на выходе жгута в виде плотной намотки: спекл-поле (в); пучок разделен на две составляющие (г); пучок разделен на три составляющие (д); пучок разделен на четыре составляющие (е)



а б в

Рис. 2. Жгут диаметром 3 мм в виде намотки с увеличенным шагом (а). Картины распределения интенсивности излучения по поперечному сечению пучка диаметром 3 мм на выходе жгута: пучок разделен на шесть составляющих (б); пучок разделен на семь составляющих (в)

Таким образом, на примере винтовой укладки оптического жгута продемонстрирована возможность формирования различной модовой структуры на его выходе. Представляет интерес выявление конкретной геометрии жгута, формирующей различный модовый состав излучения, определение апертурных углов, предельных для возбуждения регулярных и спекл-полей. В качестве отдельного направления возможных исследований отметим исследование обнаруженных автором спекл-полей.

Литература

1. Morshnev S.K., Gubin V.P., Isaev V.A., Starostin N.I., Sazonov A.I., Chamorovsky Yu.K., Korotkov N.M. Concerning the question about physical model of birefringent spur fiber // Optical Memory and Neural Networks. 2008. V. 17. P. 258–262.
2. Morshnev S.K., Ryabko M.V., Chamorovsky Y.K. Measuring of an embedded linear birefringence in spun optical fibers // Proceedings of SPIE – The International Society for Optical Engineering. 2007. V. 6594. Art. 65940R.
3. Кизеветтер Д.В. Поляризационные и интерференционные эффекты в многомодовых волоконных световодах: автореф. дис. … д-р. физ.-мат. наук. СПб.: СПбГПУ, 2008. 36 с.
4. Trufanov A.N., Smetannikov O.Y., Trufanov N.A. Numerical analysis of residual stresses in preform of stress applying part for PANDA-type polarization maintaining optical fibers // Optical Fiber Technology. 2010. V. 16. N 3. P. 156–161.
5. Моршнев С.К. Оптические свойства изогнутых волоконных световодов: автореф. дис. … д-р. физ.-мат. наук. М.: ИРЭ им. В.А. Котельникова РАН, 2009. 35 с.
6. Турунен Я. Дифракционное профилирование распределения интенсивности частично пространственно когерентного светового пучка. Патент РФ №2343516, опубл. 10.01.2009.
7. Шрамко О.А., Рупасов А.В., Новиков Р.Л., Аксарин С.М. Метод исследования зависимости *h*-параметра анизотропного световода от радиуса изгиба // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 1 (89). С. 26–31.

Конькова Елена Петровна – кандидат физ.-мат. наук, старший преподаватель, ВолГУ, Волгоград, 400062, Российская Федерация, kon_ele@mail.ru

Elena P. Kon`kova – PhD, senior lecturer, Volgograd State University, Volgograd, 400062, Russian Federation, kon_ele@mail.ru

Принято к печати 26.06.14
Accepted 26.06.14

Уважаемые читатели!

Уважаемые подписчики научно-технической литературы!

Журнал выходит 6 раз в год.

На журнал «Научно-технический вестник информационных технологий, механики и оптики» можно офорить подписку в любом отделении связи. Подписной индекс – 47197 (полугодовая подписка) и 70522 (годовая подписка) по каталогу агентства Роспечать (Газеты. Журналы).

Годовая подписка оформляется до 10 декабря, полугодовая подписка оформляется до 10 июня и 10 декабря.

Срок подписки и ее стоимость можно уточнить в редакции:

Санкт-Петербург, Кронверкский пр., 49, комн. 330.

Сайт журнала http://ntv.ifmo.ru