IRSTI 20.53.19 DOI: https://doi.org/10.26577/JMMCS.2021.v112.i4.11

Y.N. Amirgaliyev[1], **I.N. Bukenova**[1,2*]
[1]International Information Technology University, Kazakhstan, Almaty
[2]Almaty Technological University, Kazakhstan, Almaty
E-mail: ibukenowa@mail.ru*

RECOGNITION OF A PSYCHOEMOTIONAL STATE BASED ON VIDEO SURVEILLANCE: REVIEW

Recognition of human activity based on video is currently one of the most active areas of research in the field of computer vision. Various studies show that the effectiveness of recognizing actions depends on the type of functions performed and how these actions are expressed. At the same time, determining the psychoemotional state of controlled persons, including students and schoolchildren, is an urgent socially significant problem. The pace of technology development and the growing interest of foreign and domestic specialists indicate that automation of the detection of psychoemotional reactions is an urgent and popular area of research. The main purpose of this study is to review various literary sources and study methods of image recognition, artificial intelligence technology as a means of determining the psychoemotional state observed based on video surveillance. The article discusses modern types of emotional artificial intelligence that allow a computer to recognize and interpret human emotions and respond to them. The camera reads a person's state, and the neural network processes data to detect emotions. The pattern recognition methods discussed in this article can solve many problems, and you can find a method that matches the programming language used.

Key words: Computer vision, image processing, cluster analysis, signal processing, neural network, filtering.

Е.Н. Амиргалиев[1], И.Н. Букенова[1,2*]
[1]Халықаралық ақпараттық технологиялар университеті, Қазақстан, Алматы қ.
[2]Алматы технологиялық университеті, Қазақстан, Алматы қ.
E-mail: ibukenowa@mail.ru*

Бейнебақылау негізінде психоэмоционалды жағдайды тану: шолу

Бейне негізінде адамның іс-әрекетін тану қазіргі уақытта компьютерлік көру саласындағы зерттеулердің ең белсенді бағыттарының бірі болып табылады. Әр түрлі зерттеулер іс-әрекеттерді танудың тиімділігі шығарылатын функциялардың түріне және осы әрекеттердің қалай көрсетілетініне байланысты екенін көрсетеді. Сонымен қатар бақылаудағы адамдардың, соның ішінде студенттер мен оқушылардың психоэмоционалды жағдайын анықтау әлеуметтік маңызы бар өзекті мәселе болып табылады. Технологияның даму қарқыны, шетелдік және отандық мамандардың қызығушылығының артуы психоэмоционалды реакцияларды анықтауды автоматтандыру зерттеудің өзекті және сұранысқа ие бағыты екенін көрсетеді. Бұл зерттеудің негізгі мақсаты бейне бақылау негізінде байқалатын психоэмоционалды жағдайды анықтау құралы ретінде әртүрлі әдеби дереккөздерге шолу және бейнелерді тану әдістерін, жасанды интеллект технологиясын зерттеу болып табылады. Мақалада компьютерге адамның эмоциясын тануға және түсіндіруге және оларға жауап беруге мүмкіндік беретін эмоционалды жасанды интеллекттің қазіргі түрлері қарастырылған. Камера адамның жағдайын оқиды, ал нейрондық желі эмоцияны анықтау үшін деректерді өңдейді. Мақалада қарастырылған үлгіні тану әдістері көптеген мәселелерді шешуге қабілетті және сіз қолданылған бағдарламалау тіліне сәйкес әдісті таба аласыз.

Түйін сөздер: Компьютерлік көру, суретті өңдеу, кластерлік талдау, сигналдарды өңдеу, нейрондық желі, сүзгілеу.

Е.Н. Амиргалиев[1], И.Н. Букенова[1,2*]
[1]Международный университет информационных технологий, Казахстан, г.Алматы
[2]Алматинский технологический университет, Казахстан, г.Алматы
E-mail: ibukenowa@mail.ru*

Распознавание психоэмоционального состояния на основе видеонаблюдения: обзор

Распознавание действий человека на основе видео в настоящее время является одним из самых активных направлений исследований в области компьютерного зрения. Различные исследования показывают, что эффективность распознавания действий в значительной степени зависит от типов извлекаемых функций и способа выражения этих действий. В то же время определение психоэмоционального состояния наблюдаемых по видео, в том числе студентов и учеников является актуальной задачей, имеющая социальную значимость. Темпы развития технологий и повышенный интерес зарубежных и отечественных специалистов показывают, что автоматизация определения психоэмоциональных реакций – актуальное и востребованное направление исследований. Основная цель данного исследования заключается в обзоре литературы и методов распознавания образов, технологии искусственного интеллекта, как средства определения психоэмоционального состояния наблюдаемых на основе видео наблюдений. В статье рассматриваются существующие виды эмоционального искусственного интеллекта, позволяющие компьютеру распознавать и интерпретировать человеческие эмоции и реагировать на них. Камера, считывают состояние человека, а нейросеть обрабатывает данные, чтобы определить эмоцию. Рассмотренные в статье методики распознавания образов способны решать широчайший спектр задач, и можно найти подходящий метод под используемый язык программирования.

Ключевые слова: Компьютерное зрение, обработка изображения, кластерный анализ, обработка сигналов, нейросеть, фильтрация.

1 Introduction

The task of automatic recognition of a psychoemotional state is interdisciplinary and constantly attracts researchers of different specialties-mathematicians, programmers, psychologists, and physiologists. The progress of modern automated control systems, security systems, emergency notification systems, etc. depends on its solution. The solution of this problem is of great scientific importance for all areas of basic human research and information technology. In recent years, interest in the analysis of video surveillance with the use of artificial intelligence technology, considered as the most convenient objective way to identify emotions, the emotional state of a person, has clearly increased. first, you need to control the child's emotions every second of learning. Recognition of human behavior and generalization of the image are complex tasks of computer vision.

Tragedies happen at school, so you need to use video surveillance technology to detect emotions at school. To identify young people who may pose a potential danger, it is not enough to catch them when they use weapons at school in a critical situation. Thus, it is impossible to avoid possible harm. Therefore, to influence young people and prevent this critical situation, it is necessary to identify them in advance.

Let us define the concept of "Emotion" . Through emotions, we experience a person's attitude to something at a certain moment. An emotion is a higher level of emotional response than an emotional tone in evolutionary development. This is a reaction of adaptation to a specific situation, and not a reaction to a specific stimulus [1]. An emotional tone can evoke emotions, but emotions can arise when evaluating a situation. Differentiated assessment of emotions in different situations. The emotional tone gives a broader assessment, and the

emotion more subtly reflects the meaning of a particular situation. This is not only a method of assessing the upcoming situation, but also a mechanism for early and adequate preparation through the mobilization of mental and physical energy. Like an emotional tone, it serves as a mechanism for predicting the significance of a particular situation for a person and a mechanism for consolidating positive and negative experiences (attempts at positive or negative reinforcement).

There are two main ways to analyze emotions:

1. Contact method. When a person is put on a device that reads his pulse, the electrical impulses of the body. These technologies allow you to determine emotions, stress levels.

2. Contactless. Emotion analysis is based on video and audio recordings. The computer learns facial expressions, gestures, eye movement, voice, and speech.

To train the neural network, they collect a sample of data, manually mark the change in the emotional state of a person. The program learns patterns and understands which signs relate to which emotions.

There is a database of images. Such images can be people's faces, images, different emotional states, different objects of the three-dimensional world. The task is to search for the desired image in the database. Moreover, the task can be formulated both for finding an accurate image, and as close as possible to the specified one. An important task is the selection of methods and algorithms for image processing that can provide a high-quality solution to the problem. Processing technologies in this case depend on many parameters: the size of the database, the size of the image, the image quality, the brightness and contrast parameters of the images, the presence of the background, the angles of the objects' location [1-2].

The purpose of this work is to review the literature and methods of image recognition, artificial intelligence technology, as a means of determining the psychoemotional state of the observed based on video observations.

2 Review of literature and methods

Many scientists from near and far abroad are engaged in this task. Russian scientists Zaboleeva-Zotova A.V., Orlova Yu.A., Fedorov O.S. are engaged in determining the emotional state of a person by his movements using neural networks. In their work, they wrote that a review of the developments of Ugobe, Machine Perception, NeuroSky, VibraImage, Sound Intelligence, TruMedia, FaceReader, Federal Express, ERIC laboratories, Affective Computing Research, the Massachusetts Institute of Technology (MIT), the Fraunhofer Institute, the Universities of Geneva and Tokyo, Microsoft, Apple, Sony shows that now there is no system that fully implements the analysis of all means of transmitting human emotional reactions [2]. The authors analyzed the theories of emotions, considered the fundamental and modern works of scientists. Rosaliev V.L. and Zaboleeva-Zolotova A.V. [7] consider information about human body movements presented in the bvh format as time series. In their opinion, to analyze information about body movements, it is necessary to formalize the activity of human body movements [7]. Activity is expressed in the number of body

movements of a person: the fewer body movements, and as a result, the fewer changes in the file channels, the lower the activity value.

In the human body, there are certain vibrations by which you can judge the psychoemotional state of a person and get information from them. To assess the psychoemotional state of a person as a form of the mental state of a person, the most effective methods are those that do not depend on the opinion of the subject. This method is a vibration imaging system developed by scientists.

The system, according to Nguyen D.K. and Yuzhakov M.M., is designed to register, analyze, and study the psychological and emotional state of a person, quantify the emotional level, polygraph, psychophysiological diagnostics, and remote identification of potential dangers. This system allows you to intuitively and automatically assess the psychophysiological state of a person based on the vestibular-emotional reflex, using software visualization of the vibration halo obtained by processing the components of the amplitude-frequency vibration image.

Belarusian scientists Brilyuk, D.I., Starovoitov, V.V. [8] in the article "Neural network methods of image recognition" shows the architecture of a multi-layer neural network (MNS) consisting of sequentially connected layers, where the neuron of each layer relates to all the neurons of the previous layer by its inputs, and the outputs are connected to the neurons of the next layer. From this article, we can learn that a neural network with two decision layers can approximate any multidimensional function with any accuracy. Neural networks that have a single layer of solutions can form a linear distribution surface, which significantly reduces the range of problems they solve. A neural network with a nonlinear activation function and two layers of solutions allows you to create any convex region in the solution space and has three layers of solutions of any complexity-regions, including non-convex regions. In addition, the multi-layer neural network does not lose its ability to generalize. According to the author, the multilayer neural network is prepared using the inverse error distribution algorithm, which is a method of gradient descent in the weight space to reduce the overall error of the network. In this case, the error (more precisely, the adjusted weight value) propagates from the input to the output through the weight of the neuron associated with the opposite. [8]. Also, in their article [8] Brilyuk D.I, Starovoitov V.V. talk about the use of a multi-layer neural network for direct classification of images - the input is either the image itself in some form, or a set of previously extracted key characteristics of the image, the output neuron with maximum activity indicates belonging to the recognized class (Figure 1).

Rosaliev V. L., Bobkov A. S., Fedorov O. S. [9] in the article "The use of neural networks and granulation in the construction of an automated system for determining the emotional reaction of a person" , developed an approach to delineating the human body. To increase the effectiveness of identifying the emotional state of a person, information about the body is divided into two zones: upper and lower. The upper zone contains the nodes of the body related to the arms, head, neck, and back. The lower zone includes nodes related to the legs and pelvis [9]. According to the authors, the initially recognized motion is fed to the input of the data preprocessing subsystem. Here, the data is filtered, and data blocks are allocated that describe the static and dynamic zones of the body. After the data preprocessing subsystem, the separated blocks are processed separately by the corresponding subsystems: the analysis of poses and body movements. The information obtained after the analysis is combined and a common result is formed in the results comparison subsystem. Then the data

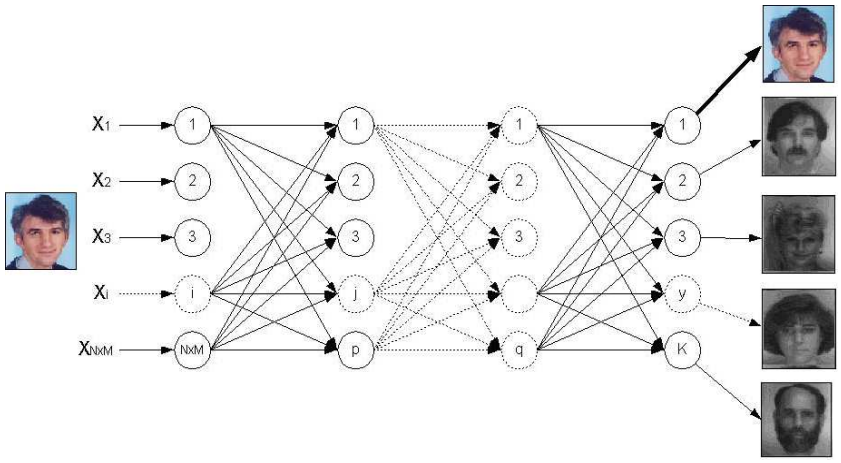

Figure 1: Multi-layer neural network for image classification. The neuron with the highest activity (here the first one) indicates that it belongs to the recognized class.

is supplemented with expert knowledge, which is stored in databases of characteristic poses and body movements and is sent to the user.

Gorokhovatsky V. A. [5] in the article "Studying the properties of clustering methods in relation to sets of characteristic features of images" , he proposed recognition methods based on the transformation of the space of structural features by clustering and applying the cluster characteristics of the base of reference images. The advantages, according to the authors of structural methods in image analysis, are the representation of visual objects in the form of a set of independent structural elements, which allows the recognition process to make effective decisions on subsets of elements and provide the necessary resistance to interference in the analyzed image. Gorokhovatsky V. A. believes that the cluster transformation of the space of structural features reduces the amount of computational costs, and hundreds of times improves the speed of recognition while maintaining the desired efficiency.

The problem of aggression, including children's aggression, has long attracted the attention of psychologists. One of the oldest psychological theories of aggressiveness – psychoanalytic-explains aggression as one of the types of natural instincts in a person who, during socialization, finds acceptable channels for exit and ways of expression in society. In this case, the increased aggressiveness of the child is explained by the insufficient strength of the "I" , which controls the behavior, as well as the insufficient development of psychological mechanisms for using aggressive energy "for peaceful purposes" . Increases the aggressiveness of the child's acquired knowledge of himself as "unmanageable" , "angry" , "brawler" , etc. Until now, the psychoanalytic interpretation of aggressiveness remains one of the most popular.

3 Recognition of emotions by actions

The system for monitoring the psychoemotional state of a person (VibraImage) is designed to detect aggressive and potentially dangerous people, using contactless remote scanning to ensure security at airports, schools, and other protected facilities [1]. The system allows

you to automatically calculate and visually evaluate the psychoemotional state of a person using software processing of a television signal and its conversion into a vibration image. The psychoemotional state of a person is characterized based on patented algorithms for analyzing the vestibular-emotional reflex and macro-movements. The Natal project from Microsoft. The software provides full 3-dimensional recognition of body movements, facial expressions, and voice. Natal can recognize emotions by voice and face.

In the literature [10], many methods of classifying human actions by visual observation have been proposed. The recognition of human actions based on video surveillance data can be viewed from different angles. Two-dimensional analysis of the recognition of actions when a person interacts with a computer requires a good exposure of the human body to obtain video. The proposed methods, according to Omar Elharrouss et al. [10], can be divided into three categories: motion-based methods, appearance-based methods, and space-time-based methods. Motion-based methods consist of calculating parametric and general optical fluxes before comparing the results with motion patterns. For appearance-based methods, the motion history of the images is extracted for comparison with the active shape models. In spatiotemporal approaches, spatiotemporal characteristics with learning outcomes are used in the spatiotemporal domain.

Dasari R, Chen CW [11] classified human actions by tracking the trajectories of the human body CDVS. El-Masry et al. in their work [12] begin by selecting areas (patches) in a video that can be described as actions.They then generate blocks containing the detected movements, and each of these blocks is assigned a discrimination score. To recognize actions, the authors applied a clustering method to each block to identify different actions.

In [3], Tejero-de-Pablos et al. propose a new video generalization method that uses player actions as a hint to determine the main points of the original video. The deep neural network approach is used to isolate two types of action-related functions and to classify video segments into interesting and uninteresting parts. The proposed method is applicable to any sports in which games consist of a sequence of actions. Elharrouss O et al. [4] proposed a video summation method based on motion detection. Sensor noise (capture and digitization noise) and changes in scene illumination are the biggest limitations of background subtraction methods. To solve these problems, this paper presents an approach based on a combination of background subtraction and structure-texture-noise decomposition.

In the article Kamiński Ł, Maćkowiak S, Domański M (2017) [6], a new method for recognizing human activity is presented. The proposed solution uses a single stationary camera to detect common human actions, such as: waving hands, walking, running, etc. Unlike other methods that use different types of characteristic point descriptors to describe human postures, the proposed solution uses a CDVS descriptor that is part of the MPEG-7 standard. This allows you to efficiently compute a compact handle in the camera.

In an article by authors Xu C, He J, and Zhang X (2017) [18], the authors believe that recognizing human movement-related actions using wearable sensors could potentially enable the use of various useful everyday applications. So far, most studies consider this as a separate problem of mathematical classification without considering the physical nature of human movements. Consequently, they suffer from data dependencies and face a dimensionality problem and a mismatch problem, and their models never become readable. Wang L, Huynh DQ, Koniusz P (2019) [19] in their paper analyze and compare 10 recent Kinect-based algorithms for both cross-action recognition and cross-action recognition using six control

data sets.

El-Henawy et al. [13] proposed a method for recognizing human actions using fast HOG3D and Smith-Waterman partial matching of the shape of each frame. First, the foreground of the video subsequence is extracted from the input stream. The keyframes of the current subsequence are then mixed before extracting the contour of the resulting frame. To classify the HOG3D functions, the author uses a nonlinear SVM decision tree. For video surveillance systems, the human body is in some cases incomplete, which is a problem for recognizing human actions [10]. This method recognizes multiple actions of multiple actors.

JIN CB et al [14] in their paper presented a sub-action descriptor for detailed action detection. The auxiliary action descriptor consists of three levels: pose, movement, and gesture level. The three levels give three categories of sub-actions for a single action aimed at solving the problem of representation. The proposed action detection model simultaneously localizes and recognizes the actions of multiple people in video surveillance using time-based appearance functions with multiple CNNs.

Another paper by AKULA A et al [15] demonstrates the use of IR cameras in the AAL region and discusses their effectiveness in recognizing human actions (HAR). Special attention is paid to one of the most responsible actions – falling. In the work, a set of IR image data was generated, consisting of 6 classes of actions - walking, standing, sitting on a chair, sitting on a chair with a table in front, falling on a table in front and falling / lying on the ground. To achieve reliable recognition of actions, the authors developed a controlled convolutional neural network (CNN) architecture.

4 Results and their comparison

If we consider using these methods to detect and recognize multiple human actions, we can compare the accuracy of the results of these methods. In this case, for comparison, I took the three methods described in [14,15,16].

The degree of accuracy for modern methods related to recognizing multiple human actions that use the same category of representation datasets:

Methods	Accuracy %
Jin CB and drl. 2017 [14]	83. 5% ICVL
Akula A et al. 2018 [15]	87.44% (infrared video)
CS-HOG-TD map [13]	97.5%

Bloisi DD et al. [16] presented the results of the multimodal background modeling method, which allows us to create a reliable initial background model, even without a clear framework. They used background subtraction as a method of detecting moving objects in image sequences.

In Elharrouss O et al [17], the authors presented a result on block background initialization using the sum of absolute differences (SAD), as well as simulations using block entropy estimation with low computational costs, which makes them suitable for an embedded platform. The author's method is effective for background generation and detection of moving objects.

To consolidate the visualized results, I used various metrics:

– total number of erroneous pixels (OKOP);

– signal-to-noise ratio (SSSH);

– multilevel structural similarity index (MISS);

– color image quality indicator (PKCI).

These indicators for the methods [13, 16, 17] are presented in Table 1.

Table 1: Performance results of the compared background modeling methods

Метод	OKOP	SSSH	MISS	PKCI
[16] IMBS-MT	9.8507	22.7278	0.9090	34.0028
[17] SAD	2.5313	27.7099	0.9892	39.6381
[13] HOG	1.1586	36.3866	0.9964	43.2006

5 Conclusion and discussion

The proposed methods suffer from some limitations, especially in the case of occlusion and very crowded scenes. Indeed, it is quite difficult to detect and recognize multiple human bodies in a crowded environment. One solution to overcome these limitations is to apply some pre-processing and learning process to deal with various occlusions and crowded scenarios. Also, it's worth mentioning that, to the best of our knowledge, there are no publicly available datasets for detecting people in very busy scenes that could be used in the training process.

Neural networks allow you to develop and modify the image recognition system due to the possibility of combining and interconnecting different networks. The effectiveness of cluster recognition significantly depends on the formed system of clusters in the application database of classes set by standards. The transition to the vector-cluster view significantly increases the speed of recognition by simplifying processing [5].

Life is rapidly being informatized, and the introduction of systems for monitoring psychoemotionality by video order transmitted over 3G networks or via the Internet (i.e., without significant delays and in good quality) will allow the introduction of systems that can monitor the life of society, revealing aggression, anger [1].

The results of the action recognition may be affected by changes in the lighting. Detecting any lighting changes in the scene can be useful to improve the video captured before the action is recognized. The proposed methods for detecting changes in illumination allow you to apply an improvement in video quality only if there is any change.

Building a complex recognition system requires a preliminary analysis of all available information about the objects being studied. The variety and complexity of the tasks of recognizing the psychoemotional state do not make it possible to implement one universal approach to the solution. Thus, existing image recognition techniques are able to solve a wide range of tasks, and depending on the limiting factors (development budget, speed and accuracy of image recognition), you can find a suitable method for the programming language used.

References

[1] Orlova Yu.A., *Review of modern automated systems for recognizing human emotional reactions* (Volgograd, Russia: VolgSTU, 2011).

[2] Zaboleeva-Zotova A.V., Orlova Yu. A., Rozaliev V. L., Fedorov O. S., "Determination of the emotional state of a person by his movements using neural networks" , *Volgograd, Russia: Bulletin of the RSUPS* 2 (2011).

[3] Tejero-de-Pablos A., Nakashima Y., Sato T., Yokoya N., Linna M., Rahtu E., "Summarization of user-generated sports video by using deep action recognition features" , *IEEE Trans Multimed* 20 (8) (2018): 2000-2011.

[4] Elharrouss O., Al-Maadeed N., Al-Maadeed S., "Video Summarization based on Motion Detection for Surveillance Systems" , *In 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), IEEE* (2019): 366-371.

[5] Gorokhovatsky V.A., *Studying the properties of clustering methods in relation to sets of characteristic features of images* (Kiev, Ukraine, 2016).

[6] Kamiński Ł., Maćkowiak S., Domański M., "Human activity recognition using standard descriptors of MPEG CDVS" , *International Conference on Multimedia & Expo Workshops. IEEE* (2017).

[7] Rozaliev V.L., Zaboleeva-Zolotova A.V., "Modeling of the emotional state of a person based on hybrid methods" , *Volgograd, Russia: VOLGGU Izv.* (2008).

[8] Brilyuk D.I., Starovoitov V.V., "Neural network methods of image recognition" , *Minsk, Belarus: ITK NANB* (2013).

[9] Rozaliev V.L., Bobkov A.S., Fedorov O.S., "The use of neural networks and granulation in the construction of an automated system for determining the emotional reaction of a person" , *Volgograd, Russia: Izvestia* (2010).

[10] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, Ahmed Bouridane, Azeddine Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences" , *Appl Intell* 51 (2020): 690-712. https://doi.org/10.1007/s10489-020-01823-z

[11] Dasari R., Chen CW., "MPEG CDVS Feature Trajectories for Action Recognition in Videos" , *Conference on Multimedia Information Processing and Retrieval. IEEE* (2018).

[12] El-Masry M., Fakhr M.W., Salem M.A.-M., "Action recognition by discriminative EdgeBoxes" , *IET Comput. Vis.* 12 (4) (2017): 443-452.

[13] El-Henawy I., Ahmed K., Mahmoud H., "Action recognition using fast HOG3D of integral videos and Smith-Waterman partial matching" , *IET Imag. Process* 12 (6) (2018): 896-908.

[14] Jin C-B., Shengzhe L.I., Hakil et KIM, "Real-Time Action Detection in Video Surveillance using Sub-Action Descriptor with Multi-CNN" , *arXiv*: 1710.03383. (2017)

[15] Akula A., Shah AxK., et Ghosh R., "Deep learning approach for human action recognition in infrared images" , *Cogn. Syst. Res.* 50 (2018): 146-154.

[16] Bloisi D.D., Pennisi A., Iocchi L., "Parallel multi-modal background modeling" , *Pattern Recogn. Lett.* 96 (2017): 45-54.

[17] Elharrouss O., Abbad A., Moujahid D., et al., "Moving object detection zone using a block-based background model" , *IET Comput. Vis.* 12 (1) (2017): 86-94.

[18] Xu C., He J., Zhang X., "DFSA: A classification capability quantification method for human action recognition" , *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/ SCALCOM/ UIC/ ATC/ CBDCom/ IOP/ SCI* (2017)

[19] Wang L., Huynh D.Q., Koniusz P., "A comparative review of recent Kinect-based action recognition algorithms" , *IEEE Trans Image Process* 29 (2019): 15-28.