Development of an algorithm for abnormal human behavior detection in intelligent video surveillance system

D Zh Satybaldina¹, N S Glazyrina², K A Kalyanova³, V S Stepanov⁴
¹,² L.N.Gumilyov Eurasian National University, Nur-Sultan, Pushkin str., 11, 010000, Republic of Kazakhstan
³ Kazakh National Women's Pedagogical University, Almaty, Ayteke bi str., 99, 05000, Republic of Kazakhstan
⁴Kazdream Technologies LLP, Nur-Sultan, Turkestan str., 10, 010000, Republic of Kazakhstan
E-mail satybaldina_dzh@enu.kz

Abstract. The aim of the work is to develop algorithms for analyzing video data in real time based on computer vision methods, as well as deep learning technologies for artificial neural networks for abnormal human behavior recognition near critical facilities using ATMs as an example. The article provides an overview of the initial research aimed at the choice of data capture devices, neural network architecture, software implementation and selection of experimental conditions (distance and illumination). Static and dynamic hand gestures were used as object’s movements. Experimental results show that using the Intel RealSense D435 Depth Camera provides more accurate dynamic gesture recognition under different experimental conditions.

Keywords: Video analytics; Security systems; Abnormal behavior; Pattern recognition; Neural networks

1. Introduction
Video surveillance is an integral component of the security system of modern cities and, in combination with computational analytics, can significantly expand its functionality, including face recognition [1], traffic monitoring [2], detecting suspicious objects and investigating crimes [3].

Key video surveillance applications for public safety, such as monitoring and tracking, need to transmit and process large amounts of video data in real time, which causes significant delays for communication networks [4]. The need for human resources to interpret the data is also growing. This results in the traditional video surveillance systems being designed for offline forensic analysis and used otherwise than as a proactive tool to prevent suspicious activity before damage occurs.

Recently proposed second generation video surveillance systems aim to minimize the role of human operators, instead of which intelligent machine learning algorithms process the collected video frames in the cloud to detect, track and report any unusual circumstances [5]. The introduction of video analytics technologies in recent decades has led to success in various fields, such as emergency detection [6],...
traffic control and management [7], analysis of the motor activity of people and any objects of interest [8, 9]. The availability, development and price of processors and sensors for depth vision, surveillance cameras with smart functions also led to the widespread use of video analytics tools [5].

In this article, we propose an approach that combines computer vision and pattern recognition methods, as well as modern video capture devices that can provide progress in the development of algorithms for recognizing physical movements of people in real time and alerting security personnel in case of classification of abnormal (suspicious, atypical) behavior near critical facilities using ATMs as an example. In works [10, 11], we have already presented an approach for recognizing static hand gestures based on a deep convolutional neural network using the concept of learning transfer.

Experimental studies have shown that the proposed model provides high recognition accuracy (99.4%). The purpose of this work is to develop a software tool for preprocessing data from two types of capture cameras under different experimental conditions in terms of distance to cameras and illumination, implementation of hand gestures classification based on a convolutional neural network.

The rest of this article is structured as follows. Section 2 provides an overview of the related work. Section 3 describes application classes for recognizing static and dynamic hand gestures. Section 4 presents the experimental results. The conclusions made and results obtained are provided in Section 5.

2. Related works

Nowadays, ATMs have become one of the necessities of modern life and are widely used for various banking operations. However, the number of ATM-related crimes has also increased. In general, ATM-related crimes can be divided into three main groups: logical attacks, fraud related to credit cards and currency and physical attacks. In addition, it is important to highlight fraud and illegal actions aimed at people using an ATM. Each type of crime is accompanied by certain movements on the part of the offender and certain suspicious (abnormal) behavior.

In modern video surveillance systems for public safety purposes, smart motion detectors are implemented, capable of classifying only simple types of movements, processes such as "loitering", crowding, falling of a person, holding a weapon in hands and leaving a suspicious object.

Human activity recognition (HAR) is the subject of a significant number of works, which show that each observation environment must satisfy a certain set of requirements [12]. These requirements contributed to the creation of specialized HAR tools, both for equipment (RGB and RGBD sensors, CCTV cameras) and for software (OpenCV libraries, etc.). The most common approaches include face recognition [1, 13], hand gestures [10-11,14], crowd behavior [15, 16], and atypical behavior [17, 18].

Most security approaches focus on the occurrence of crime. The authors of [17] present an algorithm consisting of the stages of capturing motion scenes from a sensor, subtracting the background, tracking pedestrian movements in order to make a decision. This approach divides the frame into eight areas and looks for a speed shift for one of the tracked people, one of whom is attempting a theft. At the same time, the system can only warn when a person has already lost his possessions. The work [18] presents a real approach for detection of 13 patterns of suspicious behavior such as burglary, hand-to-hand combat, gunfire and vandalism. They label patterns into two categories (normal and abnormal) and use a 3D convolutional neural network (CNN) to extract features.

Recent approaches to wireless activity detection provide good results, although they require specialized hardware. The authors of [19] proposed a model for the propagation of WiFi signals in physical space in the presence of human activity. It has been shown that human motor activity interferes with the transmission of wireless signals from the access point to the receiver, while the information about the phase and amplitude in the channel (Channel State Information, CSI) changes considerably. Based on this model, the NotiFi system is proposed, which shows the detection of anomalous human motor activity with an accuracy of 89.2% in the line of sight, 85.6% outside the line of sight and 75.3% in through-the-wall scenarios.

We have previously used the WiFi signal propagation model to study the problem of navigation inside modern buildings with a complex structure [20]. In this work, we have developed our own method for positioning objects and an augmented reality application for indoor navigation.
There are also several previous works related to the definition of abnormal behavior of people near ATMs. The authors of the article [21] proposed a surveillance system for detecting abnormal activity at ATM locations. The system uses Kinect 3D cameras and is capable of classifying a person's aggressive and normal posture. Work [22] presents a system based on image processing and machine learning methods that classify normal and potentially criminal activities associated with theft near ATMs. In [23], a system is proposed that identifies abnormal actions aimed at users of ATMs that occur in rooms with ATMs. The system uses an approach that involves application of Motion history image (MHI) and Hu moments to extract relevant features from the video to identify abnormal human behavior. In [24], the classification of human actions into normal and abnormal is presented based on the histogram of directed gradients (HOG) and the Random forest machine learning algorithm. The authors of [25] developed the ArchCam expert system capable of detecting suspicious behavior in the vicinity of an ATM, such as squatting in front of an ATM, attempting to climb onto an ATM, and detecting a construction belt in a room with an ATM.

Summarizing a brief overview of the state of modern video surveillance systems and research in the field of HAR, it can be assumed that the development of algorithms for analyzing video data in real time, based on computer vision methods, deep learning technologies of artificial neural networks in combination with a WiFi signal propagation model, can be an effective approach for recognizing suspicious human movement near critical facilities.

3. Description of Python applications for recognizing static and dynamic hand gestures

This section presents a description of the main application classes, developed in Python 3.6, for recognizing static and dynamic hand gestures based on video preprocessing and classifying the extracted features using a neural network.

In our previous studies on static gesture recognition, we used an Intel® RealSense™ D435 depth camera as a data acquisition device, which provides an RGB image as well as depth for each point [10,11]. The application is designed with the use of both depth cameras and RGB cameras as capture sensors. For this, both methods from the RealSense and OpenCV libraries were used, as well as methods implemented by the authors themselves. The application class diagram is shown in Figure 1.

The methods and functions of the App class are used to build a graphical interface that is implemented using the Tkinter library. The Pillow library is used to draw frames in the graphical interface. This class also contains a keyboard handler.

The Settings class is a graphical application designed for the user to select a video source. Available options are as follows: webcam, Intel camera, and video read. For OS Windows 8 and below, the Intel camera cannot be selected. The GUI inherits from the App class. The Main class is the main program, the link between all classes. Having received a choice of a video source from the Settings class, it
initializes the Video_Capture, History_Motion and Neural_Network classes for further work. The Video_Capture, SimpleCamera, WebCamera, File and RSCamera classes are classes responsible for capturing video clips from a depth camera or RGB camera.

The Neural_Network and History_Motion classes are responsible for the preparatory stage and recognition of the gesture in the frame.

The Video_Capture class is the interface between the main class and the class required for the camera to work. The connection to the depth camera is carried out using the RealSense library, which has standard functions for initializing the camera, setting its operation parameters, functions and methods for reading frames from a video stream, calculating the distance from a hand to a depth camera, methods for storing RGB images and depth maps. Connection to the RGB camera is carried out through the OpenCV library. The functions of the class are also intended for capturing the frame, subtracting the background and detecting a hand in the frame. Depending on whether the hand was detected in the fragment or not, either a regular frame or a frame without a background is fed to the main class. This class was developed by us independently, with the main development idea based on the implementation of the possibility of connecting various sensors without changing the code of the main class.

The SimpleCamera class is responsible for connecting to an RGB camera using the OpenCV library. The functions of the class are intended for reading a frame from the video stream, creating an object used to subtract the background from the frames and create a mask, based on which the frame without the background is calculated. The WebCamera and File classes are designed to work with cameras and video files, respectively. All functionality is inherited from the SimpleCamera class.

The RSCamera class is designed to connect to a depth camera using the RealSense library. An RGBD frame is obtained from a video stream and the average brightness of the frame is calculated. If the brightness is below the threshold value, then instead of the RGB frame, the deep one will be fed into the main class. The method for calculating the average brightness was developed by us independently, the idea of the method being to transfer the frame from the RGB system to the HSV (Hue, Saturation, Value) color model where V is the brightness value. HSV is a non-linear RGB conversion. Subtraction of the background is based on comparing the depth pixels with the threshold: if the value is greater than the threshold, then in the resulting frame, the pixel is colored black, otherwise - with the color of the current RGB frame. Hand detection functions are based on calculating the required number of pixels in the frame without the background in the fixed area. Since the entire background is painted black in the frame, the number of pixels that are not equal to zero is calculated. If the required number of pixels is collected, the system fixes the hand in the frame.

The Neural_Network class is responsible for connecting to the neural network using the TENSORFLOW library. The methods of the class are designed to transform the resulting frame with the image of the hand into the format required for feeding it to the input of the neural network. As a result of the work of the connected neural network, an estimate of the convergence with reference sets of movements is issued. This score is passed to the main class to display the result on the screen.

The History_Motion class is intended for displaying frames in dynamics in an informational form. The idea behind the method is to select key frames from a set obtained over a certain period of time. A frame is selected in two cases: if the number of pixels that differ from the frame taken as a basis is more than a certain threshold, or the position of the hand does not change for a certain period of time. Initially, the first frame in the set is taken as the basis for frame comparison. Each subsequently selected frame in this process will be used as a basis until the next frame is selected. After the frame has ceased to be the basis for comparison, it is moved to a new set with a fixed time interval indicating the duration of its use by the basis for comparison. This continues until all the frames in the set have been considered. After that, the arithmetic mean for the time indices is calculated from the selected frames, and those frames in which this indicator is less than the arithmetic one are eliminated from the fixed set. The filtered set is used to map the generated movement. The implementation of the class is based on the methods of the Numpy and OpenCV libraries.

4. Experimental research
The developed application was tested on the problem of static and dynamic hand gestures recognition. We prepared a database that contains images with segmented gestures presented in Figures 2 and 3.

A Logitech HD Pro Webcam C920 webcam and an Intel RealSense D435 depth camera were used as a gesture capture device.

Gesture recognition was carried out at a distance of 25 cm, 37.5 cm from the cameras; 43.75 cm, 50 cm, 62.5 cm; 75 cm. Tables 1, 2 show the results of static hand gesture recognition using RGB and RGBD cameras.

Table 1. Recognition of hand gestures based on the Logitech HD Pro Webcam C920

| Distance, cm | Fist   | L      | Okay  | Palm  | Peace |
|--------------|--------|--------|-------|-------|-------|
| 25           | 98,899 | 99,417 | 99,994| 77,766| 10,026|
| 37.5         | 98,284 | 99,992 | 99,999| 99,461| 22,391|
| 43.75        | 99,99  | 99,982 | 100   | 97,548| 63,666|
| 50           | 98,189 | 92,826 | 99,95  | 90,732| 77,887|
| 62.5         | 99,935 | 84,518 | 99,991| 41,755| 61,152|
| 75           | 99,972 | 39,8749| 99,99 | 0,8199| 91,743|

In Tables 1, 2, each column represents the percentage of gestures belonging to each class, and along the first diagonal the maximum values of the probability of correct gesture identification are specified, i.e. assignment of the observed object to a certain class of gestures.

After analyzing the resulting diagrams, we can see that when using the Intel RealSense D435 camera, more accurate gesture recognition is provided.

Recognition of static gestures using the Intel RealSense D435 camera was also carried out under different lighting conditions; the recognition results are shown in Table 3.
Table 3 - Performance of the static gesture recognition system under different lighting conditions

| Experimental conditions                                    | Fist    | L      | Okay   | Palm    | Peace   |
|------------------------------------------------------------|---------|--------|--------|---------|---------|
| late in the evening with no lights on                       | 95,522  | 98,080 | 99,402 | 58,548  | 75,120  |
| late in the evening with the lights on                       | 63,502  | 28,072 | 99,726 | 57,521  | 38,320  |
| late in the evening with the lights on and the lamp in front of the camera | 95,577  | 98,308 | 97,305 | 43,271  | 50,324  |
| daytime                                                    | 98,015  | 99,713 | 99,984 | 98,399  | 56,74   |
| late evening without the lights on, only the depth pixels used | 98,542  | 99,899 | 99,729 | 82,629  | 97,838  |

The results of recognition of dynamic hand gestures based on the processing of depth images are presented in Table 4. As can be seen, for all gestures, the value of the elements of the main diagonal is more than 88%.

Considering that the resulting accuracy of the classifier is calculated as the arithmetic mean of its accuracy for all classes, the performance of the proposed approach in terms of the recognition accuracy of dynamic gestures was obtained at the level of 0.9131, or ≈91.31%.

Based on the presented analysis of the results of gesture recognition at various distances from the cameras, under different lighting conditions, as well as studying the technical characteristics of the cameras, Intel RealSense D435 cameras were chosen to detect people in front of the ATM and their suspicious actions.

Table 4. Performance of the dynamic gesture recognition system

|                   | Swap Up | Swap Down | Swap Left | Swap Right | Circle Left | Circle Right |
|-------------------|---------|-----------|-----------|------------|-------------|--------------|
| Swap Up           | 89,112  | 0,060     | 0,056     | 10,353     | 0,237       | 0,178        |
| Swap Down         | 0,154   | 92,474    | 0,154     | 2,002      | 5,144       | 0,071        |
| Swap Left         | 2,177   | 2,260     | 92,088    | 0,939      | 1,262       | 1,274        |
| Swap Right        | 0,021   | 3,428     | 3,437     | 92,906     | 0,197       | 0,011        |
| Circle Left       | 1,0797  | 3,0598    | 1,0789    | 0,3019     | 88,69       | 5,772        |
| Circle Right      | 0,0471  | 0,7771    | 0,263     | 3,4161     | 2,93        | 92,561       |

5. Conclusion
The article provides an overview of the results of application development and its main classes for recognizing static and dynamic hand gestures, experimental studies of the use of the Logitech HD Pro Webcam C920 webcam and the Intel RealSense D435 depth camera as capture sensors under different conditions of distance and illumination are carried out.

It has been shown that the accuracy of gesture recognition depends on the conditions under which hand poses are demonstrated (illumination, distance to the camera).

The use of a convolutional neural network was found to contribute to a sufficiently high classification accuracy even in the case of video capture from an RGB camera. But the obtained higher accuracy of recognition of static and dynamic hand gestures from an RGBD camera under different experimental conditions confirms the potential performance expectations due to the pixels of depth.

Thus, in the course of the research, video data capture sensors for an intelligent video surveillance system were identified in terms of performance.

6. Acknowledgments
This work was supported by the Ministry of Education and Science of the Republic of Kazakhstan within the framework of project No. AP09260786.

References

[1] Ji P, Kim Y, Yang Y and Kim Y S 2016 Face occlusion detection using skin color ratio and LBP features for intelligent video surveillance systems FedCSIS (Gdansk) ACSIS (Polish Information Processing Society) V 8 pp.253-259 https://doi.org/10.15439/2016F508

[2] Zhou Y, Liu L, Shao L and Mellor M 2018 Fast automatic vehicle annotation for urban traffic surveillance IEEE Trans. Intell. Transport. Syst. (IEEE) vol. 19 no. 6 pp.1973– 1984

[3] Puvvadi U L N, Benedetto K D, Patil A, Kang K-D and Park Y 2015 Cost-effective security support in real-time video surveillance IEEE Trans Ind. Informat. (IEEE) vol. 11, no. 6 pp.1457–1465

[4] Prakash P, Suresh R and Pr D. K 2019 Smart city video surveillance using fog computing Int. J. of Enterprise Network Management (Springer, Cham) vol. 10, no. 3-4 pp. 389-399

[5] Shidik, G. F. et al. 2019 A Systematic Review of Intelligence Video Surveillance: Trends, Techniques, Frameworks, and Datasets IEEE Access (IEEE) vol. 7 pp.170457-170473

[6] Filonenko A, Hernández D C and Jo K-H 2018 Fast smoke detection for video surveillance using CUDA IEEE Trans. Ind. Informat. (IEEE) vol. 14, no. 2 pp. 725–733

[7] Finogeev A, Finogeev A, Fionova L, Lyapin A and Lychagin K A 2019 Intelligent monitoring system for smart road environment J. Ind. Inf. Integr (Amsterdam: North-Holland/American Elsevier) vol. 15 pp. 15–20

[8] Murugan A S, Devi K S, Sivaranjani A and Srinivasan P 2018 A study on various methods used for video summarization and moving object detection for video surveillance applications Multimedia Tools Appl.(Springer, Singapore) vol. 77, no. 18 pp. 23273–23290

[9] Zhao L, He Z, Cao W and Zhao D 2018 Real-time moving object segmentation and classification from HEVC compressed surveillance video IEEE Trans. Circuits Syst. Video Technol. (IEEE) vol. 28, no. 6 pp. 1346–1357

[10] Satybaldina D., Kalymova G. and Glazyrina N. 2020 Application Development for Hand Gestures Recognition with Using the Depth Camera Communications in Computer and Information Science (Springer, Cham) V.1243 pp. 55-67

[11] Satybaldina D Zh, Ovechkin G V and Kalymova G A 2020 Static hand gestures recognition system with using depth camera Vestnik of RSRUE (Ryazan: Ryazan State Radio Engineering University Press) N. 72 pp.79-105. https://doi.org/10.21667/1995-4565-2020-72-93-105

[12] Martínez-Mascorro G A et al. 2005 Suspicious Behavior Detection on Shoplifting Cases for Crime Prevention by Using 3D Convolutional Neural Networks (arXiv preprint arXiv:02142v1[cs.CV])

[13] Wang Y, Bao T, Ding C and Zhu M l 2017 Face recognition in real-world surveillance videos with deep learning method IEEE 2nd International Conference on Image, Vision and Computing (Chengdu) (IEEE) pp. 239-243

[14] Satybaldina D Zh Kalymova G A Stepanov V S 2020 Application for recognition of static gestures by using deep images Certificate of state registration of the right to the computer program, № 10319 (National Institute of Intellectual Property)

[15] Direkoglu C, Sah M and O'Connor N E 2017 Abnormal crowd behavior detection using novel optical flow-based features 14th IEEE International Conference on Advanced Video and Signal Based Surveillance(Lecce) (IEEE) pp. 1-6

[16] Rabiee H et al. 2018 Detection and localization of crowd behavior using a novel tracklet-based model International Journal of Machine Learning and Cybernetics (Springer Nature)V. 9, №. 12 pp.1999-2010

[17] Tsushita H and Zin T T 2018 A Study on Detection of Abnormal Behavior by a Surveillance Camera Image Int. Conference on Big Data Analysis and Deep Learning Applications (Miyazaki) (Springer, Singapore) pp. 284-291

[18] Sultani W, Chen Ch and Shah M 2018 Real-world anomaly detection in surveillance videos IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE) pp.6479-6488
[19] Zhu D et al. 2017 A ubiquitous WiFi-based abnormal activity detection system *IEEE International Joint Conference on Neural Networks* pp.1766-1773

[20] Stepanov V S and Glazyrina N S 2018 Development of an application for navigation in the room *Herald of PSU* (Pavlodar: Toraighyrov University Press) №3 pp.171-180

[21] Nar R, Singal A and Kumar P 2016 Abnormal activity detection for bank ATM surveillance *IEEE Int. Conference on Advances in Computing, Communications and Informatics* (IEEE) pp. 2042-2047

[22] Mandal R and Choudhury N 2016 Automatic video surveillance for theft detection in ATM machines: An enhanced approach *3rd International Conference on Computing for Sustainable Global Development, INDIACom* (New Delhi) (IEEE) pp. 2821-2826

[23] Tripathi V, Gangodkar D, Latta V and Mittal A 2015 Robust Abnormal Event Recognition via Motion and Shape Analysis at ATM Installations *Journal of Electrical and Computer Engineering* (Hindawi Publishing Corporation)1(10)

[24] Tripathi V, Mittal A, Gangodkar D and Kanth V 2016 Real time security framework for detecting abnormal events at ATM installations *Journal of Real-Time Image Processing* (Springer Nature)pp. 1-11

[25] Lee W K, Leong C-F, Lai W-K, Leow L K and Yap Th-H 2018 ArchCam: Real Time Expert System for Suspicious Behaviour Detection in ATM Site *Expert Systems with Applications* 109 (Elsevier) pp. 12-24