Deep learning based static hand gesture recognition

Dina Satybaldina[1], Gulzia Kalymova[2]
[1]National Research Nuclear University “MEPhI” (Moscow Engineering Physics Institute), Moscow, Russian Federation
[2]Information Technologies Faculty, L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan

Article Info

Article history:

Received Mar 15, 2020
Revised Jun 6, 2020
Accepted Jul 24, 2020

Keywords:

Computer vision
Convolutional neural network
Deep learning
Gesture recognition
Machine learning
VGG-16

ABSTRACT

Hand gesture recognition becomes a popular topic of deep learning and provides many application fields for bridging the human-computer barrier and has a positive impact on our daily life. The primary idea of our project is a static gesture acquisition from depth camera and to process the input images to train the deep convolutional neural network pre-trained on ImageNet dataset. Proposed system consists of gesture capture device (Intel® RealSense™ depth camera D435), pre-processing and image segmentation algorithms, feature extraction algorithm and object classification. For pre-processing and image segmentation algorithms computer vision methods from the OpenCV and Intel Real Sense libraries are used. The subsystem for features extracting and gestures classification is based on the modified VGG-16 by using the TensorFlow&Keras deep learning framework. Performance of the static gestures recognition system is evaluated using maching learning metrics. Experimental results show that the proposed model, trained on a database of 2000 images, provides high recognition accuracy both at the training and testing stages.

Corresponding Author:

Gulzia Kalymova
Computer Engineering Department
Faculty of Information Technologies
L.N. Gumilyov Eurasian National University
Nur-Sultan, Kazakhstan
Email: gulzia_kalymova@mail.ru

1. INTRODUCTION

One of the recent year’s trends is deep learning (DL) technology which is used in the digital image processing for solving complex problems (a classification, segmentation and image detection). DL techniques, such as convolutional neural networks (CNNs), have already influenced a wide range of signal processing activities within traditional and new advanced areas, including key aspects of machine learning and artificial intelligence [1]. In particular, CNNs showed superior performance in face detection applications [2, 3]. Furthermore, DL has made considerable progress in detection and classification of the hand gestures for implementation into the human computer interaction (HCI) technologies [4-6].

A gesture is a configuration and/or movement of a body part that expresses an emotion, intention or command. A set of gestures and their meanings form the gestures vocabulary. Gestures can be divided into two types: static and dynamic gestures. In static gestures the hand position does not change during the gesture demonstration. Static gestures mainly rely on the shape and flexure fingers angles. In the second case, hand position changes continuously so that dynamic gestures rely on the hand trajectories and orientations, in addition to the shape and fingers angles [5].

Intel Inc. created RealSense deep vision technology and developed (in collaboration with Microsoft) the tool for 3D face recognition, which provides access to Windows10 devices [7]. RealSense technology supported by an open source software development kit [8]. The second generation RealSense cameras and its stereo vision to calculate depth [9] were introduced in January 2018. Despite the fact that Intel RealSense D400 series devices appeared on the market only in recent years, they began to be used in many areas, for example, in security systems [10], robotics [11], medicine [12], and agriculture [13]. At the same time, there are no works that report on the RealSense D400 using for recognizing hand gestures that can be integrated into effective HCI systems.

In this regard, the aim of this work is to develop a gesture recognition approach based on the combined use of the RealSense D435 depth sensor for gesture capture and a pre-trained convolutional neural network with VGG-16 architecture for features extraction and objects classifying. RealSense libraries from Intel, OpenCV and open-source DL frameworks Keras and TensorFlow are used for software implementation of the gesture recognition system on Python. To determine the performance of our approach we collected a database of 1,000 images, which consists of 40 different types for 5 gestures, which were presented to the sensor by 5 people, and tested the recognition system on the database.

The rest of this paper is structured as follows. The details of the proposed method are described in Section 2. In Section 3, we briefly introduce our newly collected database and show the experimental results. Conclusion and future research are presented in Section 5.

2. RESEARCH METHOD

2.1. General framework of the hand static gesture recognition system

Hand gestures recognition is a difficult task due to objective and subjective differences associated with a large number of degrees of freedom of the hand and fingers bones, the differences in articulation, a relatively small area of the hands, and a different skin color. In addition, reliable algorithms for segmentation and detection of the hands and fingers positions should have invariance with respect to the size, speed and orientation of the gesture, the scene brightness, the background heterogeneity and other parameters. Therefore, a lot of difficulties can be introduced into the automating process of the hand gestures recognition.

In order to handle the above-mentioned difficulties, we propose a recognizing static hand gestures system based on the digital processing of the color and depth images from a video stream in real time and extracting classification features from them inside convolutional neural network. The structure of the proposed system is shown in Figure 1, and operation principle of each stage is described in detail in the following subsections.

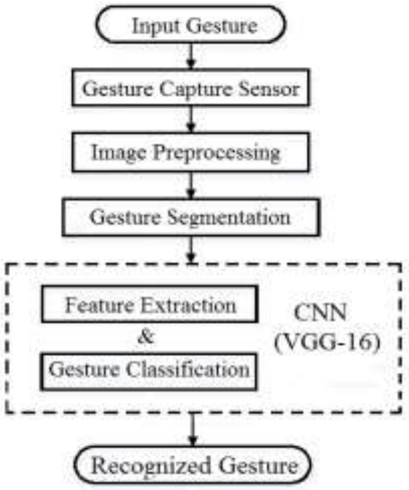

Figure 1. Block diagram of proposed static hand gesture recognition system

2.2. Capture images with hand gestures

The purpose of the gesture capture sensor is to digitize the gesture. Liao B. et al [14] proposed a static recognition system based on the deep learning model. A double-channel CNN to fuse the color information and depth information data were captured by Intel RealSense300 depth camera. Experimental results on hand gestures representing English alphabet demonstrate that the recognition accuracy is 99.4%. It was shown that depth cameras (known also as RGB-D sensors) greatly reduce various restrictions on experimental conditions as they obtain both RGB images and depth information in an actual time altogether.

Due to their versatility, accuracy and compactness such sensors and depth-sensing technology provided a significant boost for hand gesture and posture recognition algorithms [15-19]. The used methods for features extraction and gestures classification provide high recognition accuracy of both static and dynamic gestures. It should be noted that in the majority of papers, the effectiveness of recognition systems is obtained for the small gestures sets. It limits their mass application. This is confirmed by the fact that there are many open source developments for face, mouth and eye recognition (OpenCV), but there is reliable hand detectors absence. Therefore, the problem of improving static and dynamic posture recognition on a photo image or in a video stream remains relevant.

In our work the input sensor is an RGB-D camera (Intel® RealSense ™ D435) that gives us an RGB image as well as depths for each point. RealSense D435 is a compact (99mm × 25mm × 25mm; weight 72g) peripheral RGB-D device supporting USB 3.1 standard with a range of up to 10 m. The camera consists of Vision Processor D4, Depth Module and RGB camera. The characteristics of an active stereo depth deep resolution and RGB resolution are up to 1280×720 and 1920×1080, respectively.

A pair of identical cameras referred as imagers (left and right) and an infrared (IR) projector are used to stereo vision realization and to calculate depth as shown in Figure 2 [16]. The infrared projector projects non-visible static IR pattern to improve depth accuracy in scenes with low texture. The left and right imagers capture the scene and send imager data to the depth imaging (vision) processor, which calculates depth values for each pixel in the image by correlating points on the left image to the right image and via shift between a point on the Left image and the Right image. The depth pixel values are processed to generate a depth frame. Subsequent depth frames create a depth video stream.

Open library RealSense SDK 2.0 has standard functions for camera initialization, parameters setting, functions and methods for reading frames from the video stream, by calculating the distance from the hand to the depth camera, RGB images and depth maps saving methods [15]. It is possible to modify the algorithms available in the source code of the RealSence SDK 2.0. The methods and functions from Realsence SDK 2.0 were used to implement gestures image. Some additional functions for gesture capture were coded by the authors.

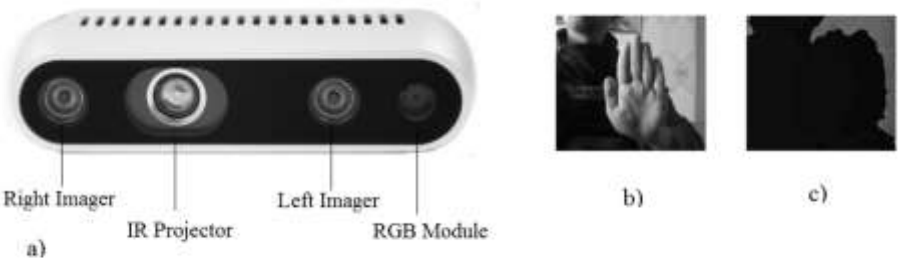

Figure 2. (a) RealSense D435, Output data from the sensor, (b) RGB image, (c) Depth image

2.3. Image preprocessing

Performing various operations in the subsequent steps of the gesture recognition system, such as segmentation and feature extraction, is much easier for pre-processed images. At this stage, usually the interferences and external noise are reduced by applying the operations of averaging and histograms leveling, color normalization is also carried out in accordance with the lighting conditions and light temperature [9].

In this paper, a bilateral filter from the OpenCV library is used to remove noise in the frame. Additional method for determining the average brightness of pixels in an image is applied. We implemented a function for calculating average brightness using non-linear conversion from RGB color model to the HSV (Hue, Saturation, Value), where Value is the brightness parameter [20].

2.4. Segmentation methods

Segmentation is an extraction of an interest object (hand) from the background and determining its location in the scene. At this step, the following operations are performed: searching for a Region of Interest (ROI) to detect hands, removing the background, converting the color image to grayscale frame, Gaussian filtering for cleaning of noised grayscale images, contouring the segmented object and final binary thresholding to isolate the hand as shown in Figure 3. To implement these operations, we used both methods from the OpenCV library and own code development.

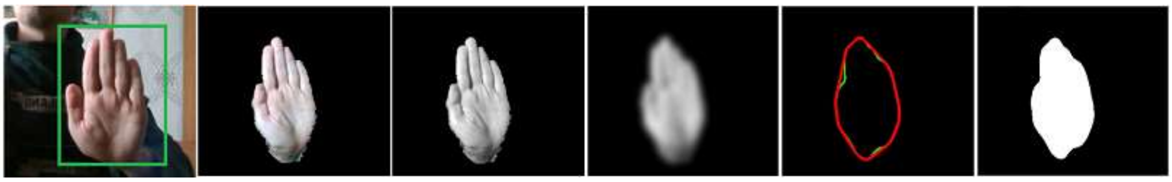

Figure 3. Hand gesture segmentation

2.5. Basic architecture VGG-16 and transfer training

A classification or recognition mechanism is a computational algorithm that takes an object representation and correlates (classifies) it as an instance of some known class type. In proposed method for hand static gestures recognition we have used deep convolutional neural network (DCNN) with VGG-16 model pre-trained on big image dataset. The VGG network architecture was developed by K. Simonyan and A. Zisserman [21]. Designing architecture with small convolution filters with 3×3 size shows a significant improvement on the neural networks that can be achieved by pushing the depth to 16-19 weight layers. The 5 convolutional blocks are followed by 5 max-pooling layers, and three fully-connected layers. The final 1000-way layer is a soft-max layer that outputs class probabilities. VGG-16 was trained and tested on ImageNet database to classify the 1.3 million images to the 1000 classes with accuracy of 92.7% [13].

DCNN with fixed weights of the convolutional layers, obtained by training on large-scale sets of natural images (ImageNet) shows the high accuracy of classifying images from other subject areas. Such approach based on VGG-16 and concept of a transfer learning and fine-tuning was used for classification of malware samples represented as byte plot grayscale images [22]. The transfer of training consists in transferring the parameters of a neural network trained with one data set to train an another neural network for a similar problem with different dataset images [23]. When the target dataset is significantly smaller than the base dataset, transfer learning can be a powerful tool to enable training a large target network without overfitting.

Fine-tuning is associated with the need to change and retrain fully connected layers in the classifier. The basic architecture of VGG-16 contains 1000 output neurons in the last layer according to the number of the object's classes. In a new classification task (for example, in the gesture recognition problem), the number of classes may differ from that in the original data set. In this case, the last layer in the VGG-16 architecture must be removed, and a new classifier with the required number of output neurons must be added. It is also necessary to replace the previous fully connected layers since their output vectors do not correspond to the new classification layer.

For the proposed system of the static hand gestures recognition we use modified VGG-16 architecture, in which fully connected layers with a large number of neurons are replaced by 4 dense layers with fewer neurons. The dropout layer has 5 output channels for 5 gestures from the training and test sets. The new layers weights are initialized with random values, after which the learning process on the training data set begins. We use an end-to-end approach proving surprisingly powerful. With minimum training data from humans the neuron's weights were corrected and recognition system was learnt to identify hand gestures. The input data to VGG-16 model is an RGB image with the fixed size of 224 × 224. Therefore, at the segmentation stage, processed frames with a segmented gesture saved in a different color resolution RGB model are converted to 224 × 224 format and transferred to the DCNN input. After going through stack of convolution and pooling layers, RGB images of hand gestures are converted into features maps and sent to dense layers. In the last step, the softmax model is used to gesture predict and output the result as a confusion matrix.

3. EXPERIMENTAL RESULTS

3.1. Datasets

We prepared a new database which contains images with segmented static hand gestures shown in Figure 4. We selected these gestures, which are also included in alternative dataset [24, 25] and can be used to cross-validate proposed static hand gesture recognition system model. Depth camera Intel RealSense D400 placed over a table, the subjects sat close to it, and moved their right hand over the sensor at a distance between 10 and 75cm in front of it as shown in Figure 5. The gestures were performed by 10 different subjects (5 women and 5 men).

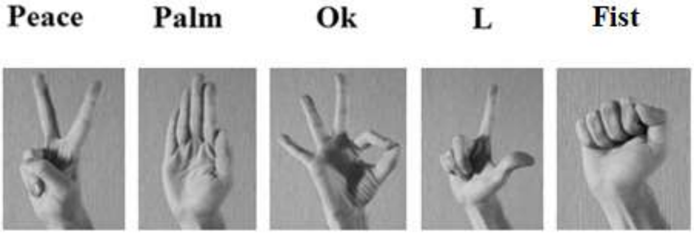

Figure 4. Dataset samples

Our dataset has a total of 2,000 pictures, including 1,000 RGB images and 1,000 depth maps, collected under different backgrounds in a few rooms with illumination changes. In order to increase the diversity of the database, we also tried image augmentation by using of flips, skews, rotations and etc. The static hand gestures database was used only for the training of the modified VGG-16. At the validation/testing stage RGB and depth images from RealSense D400 used directly to predict gesture.

DCNN training is the fine-tuned process based on the pre-trained models VGG-16 on ImageSet. At the start, the learning rate is 0.01 and then decreases by 10 times every 2000 iterations. Weight decay is set to 0.0004. At most 10,000 iterations were needed for the fine-tuning on training set. The experiments were done on a Intel(R) Core(TM) i7- 9750H CPU, NVIDIA GeForce GTX 1650, 16 GB RAM desktop. After performing transfer learning with a fine tuned last activation layer we were able to achieve an average accuracy of 95,5 % on the 5-class subset of the static hand gesture.

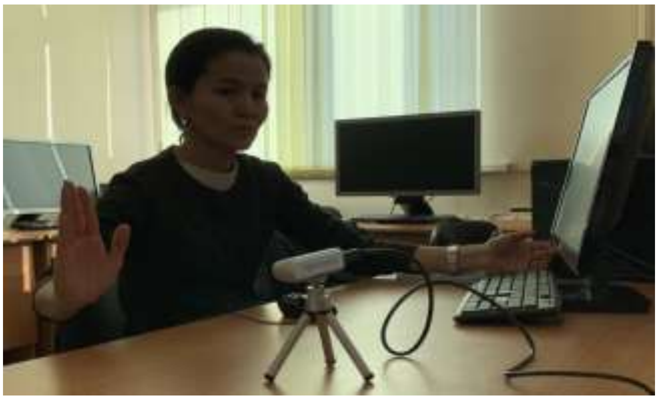

Figure 5. Sample human static hand gesture tracking from RealSense D400 sensor

3.2. Performance metrics

The confusion matrix (CM) has been used for obtaining quantitative results. The precision, recall, and F-Score measures are used as metrics for evaluation. CM of size *n*×*n* associated with a classifier shows the predicted and actual classification, where *n* is the number of different classes [25]. We use an alternative construction of CM: each column represents the percentage of hand gestures that belongs to each class, and along the first diagonal are the correct classifications, whereas all the other entries show misclassifications [24]. From CM matrix two measurements can be directly observed, the precision and the recall, which can be expressed as follows:

$$Precision = \frac{d}{d+b}, \tag{1}$$

$$Recall = 100 \times \frac{d}{d+c} \tag{2}$$

Recall demonstrates the ability of the algorithm to detect a given class in general, and precision demonstrates the ability to distinguish this class from other classes. F-Score is a harmonic mean of the precision and recall:

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

The F-Score reaches its maximum with precision and recall equal to unity and is close to zero if one of the arguments is close to zero.

3.3. Results and discussion

Table 1 presents the CM obtained by using the proposed static hand gesture recognition system on test stage. As it can be seen, for all the gestures the precision, elements of the main diagonal, are over 97% and just a small percentage of samples are detected as belonging to other gestures, but its amount is less than 0.2%. This indicates that our algorithm is quite accurate in two measurements, precision and recall. Exception is recognition of palm, where the classification error is greater than 2.5%. This can be explained by a certain similarity between this gesture and the gesture of Peace. The average accuracy of gesture recognition (according to the first diagonal of the SM) reaches 99.4%.

Table 1. Confusion matrix for the proposed static hand gesture recognition system

	Predicted Gesture					
Actual gesture	Fist	L	Ok	Palm	Peace	*Precision*$_c$
Fist	99,9798	0,0018	0,0131	0,005	0,0003	0,9998
L	0	99,9978	0,0021	0	0,0001	1,0000
Ok	0,0008	0,0255	99,9726	0	0,0011	0,9997
Palm	0,0194	0,0159	0,0151	97,4405	2,5091	0,9744
Peace	0	0,0004	0,0032	0,6245	99,3719	0,9937
Recall	0,9998	0,9996	0,9997	0,9936	0,9754	

The proposed hand gesture recognition system has been compared with the one proposed in [24]. The near-infrared images of the same gestures acquired by a Leap Motion sensor. Along with the CM values, the F-Score measure is also used to compare the results of both solutions. This measure is present in Table 2. As it can be seen in Table 2, proposed system is the one that achieves comparable results.

Table 2. F-Score results for the proposed system and for the recognition system from [24]

	Hand Gesture Recognition System	
Gesture	Proposed	[24]
Fist	0,9998	0,99
L	0,9998	0,99
Ok	0,9997	0,99
Palm	0,9839	1
Peace	0,9845	0,99

The recognition system proposed in [24] use RGB sensor and a global image descriptor, called Depth Spatiograms of Quantized Patterns (DSQP), without any hand segmentation stage. Image descriptors reduced from DSQP are analyzed by a set of Support Vectors Machines to gesture classification. This comparison also allows us to notice that depth maps and convolution neural networks allow achieving same high performance.

4. CONCLUSION

We proposed static hand gesture recognition system that utilizes both RGB and depth information from Intel® Real Sense™ depth camera D435. We utilized depth sensor's unique property to segment hand poses and perform background denoising on images with gesture. Features extraction and gesture classification were performed on the deep convolution neural networks (DCNN) with VGG network architecture, pre- trained on the ImageNet database. The proposed DCNN with transfer learning was implemented by Keras which takes the TensorFlow as the backend. Our training database was collected manually using both RGB and depth images from depth sensor. This database consists of 2,000 samples performed by 5 women and 5 men. Modified VGG-16 model delivered the high accuracy on training set of images. Obtained recognition scores on the test stage prove the efficiency of the presented recognition framework, and claim a higher prominence of the depth camera D435 for future HMI applications. However, the static hand gestures recognition is insufficient for effective HCI systems. Therefore, future research is related to the method development for recognizing dynamic hand gestures and the expansion of the database and object classes.

ACKNOWLEDGEMENTS

This work has been supported by the Ministry of Digital Development, Innovations and Aerospace Industry of the Republic of Kazakhstan under project № AP06850817.

REFERENCES

[1] S. R. Sree, et al., "Real-World Application of Machine Learning and Deep Learning," in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1069-1073, 2019.

[2] T. V. Janahiraman and P. Subramaniam, "Gender Classification Based on Asian Faces using Deep Learning," in *2019 IEEE 9th International Conference on System Engineering and Technology (ICSET)*, pp. 84-89, 2019.

[3] R. I. Bendjillali, et al., "Illumination-robust face recognition based on deep convolutional neural networks architectures," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 1015-1027, 2020.

[4] A. K. H. AlSaedi and A. H. H. AlAsadi, "A new hand gestures recognition system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, pp. 49-55, 2020.

[5] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," *Computer Vision and Image Understanding*, vol. 141, pp. 152-165, 2015.

[6] B. K. Chakraborty, et al., "Review of constraints on vision-based gesture recognition for human–computer interaction," *IET Computer Vision*, vol. 12, no. 1, pp. 3-15, 2017.

[7] L. Keselman, et al., "Intel R RealSense TM Stereoscopic Depth Cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-10, 2017.

[8] Intel® RealSense™ SDK 2.0. https://www.intelrealsense.com/developers/.

[9] Intel RealSense D400 Series Product Family. Datasheet. 2019 Intel Corporation. Document Number: 337029-007. https://www.intel.com/.

[10] R. D. Bock, "Low-cost 3D security camera. Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything," *International Society for Optics and Photonics*, vol. 10643, pp. 106430E, 2018.

[11] Q. Fang, et al., "RGB-D Camera based 3D Human Mouth Detection and Tracking Towards Robotic Feeding Assistance," in *Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference*, pp. 391-396, 2018.

[12] H. Aoki, et al., "Study on Non-Contact Heart Beat Measurement Method by Using Depth Sensor," in *World Congress on Medical Physics and Biomedical Engineering*, pp. 341-345, 2019.

[13] T. N. Syed, et al., "Seedling-lump integrated non-destructive monitoring for automatic transplanting with Intel RealSense depth camera," *Artificial Intelligence in Agriculture*, vol. 3, pp. 18-32, 2019.

[14] B. Liao, et al., "Hand gesture recognition with generalized hough transform and DC-CNN using realsense," in *2018 Eighth International Conference on Information Science and Technology (ICIST)*, pp. 84-90, 2018.

[15] M. B. Holte, et al., "Fusion of range and intensity information for view invariant gesture recognition," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-7, 2008.

[16] M. Van den Bergh, et al., "Real-time 3D hand gesture interaction with a robot for understanding directions from humans," in *2011 Ro-Man*, pp. 357-362, 2011.

[17] Z. Ren, et al., "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proceedings of the 19th ACM international conference on Multimedia*, pp. 1093-1096, 2011.

[18] D. Wu, et al., "One shot learning gesture recognition from rgbd images," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7-12, 2012.

[19] C. Keskin, et al., "Randomized decision forests for static and dynamic hand shape classification," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 31-36, 2012.

[20] V. Chernov, et al., "Integer-based accurate conversion between RGB and HSV color spaces," *Computers & Electrical Engineering*, vol. 46, pp. 328-337, 2015.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv: 1409.1556, 2014.

[22] E. Rezende, et al., "Malicious software classification using VGG16 deep neural network's bottleneck features," *Information Technology-New Generations*, pp. 51-59, 2018.

[23] Z. Liu, et al., "Improved kiwifruit detection using pre-trained VGG16 with RGB and NIR information fusion," *IEEE Access*, pp. 2327-2336, 2019.

[24] T. Mantecón, et al., "Hand Gesture Recognition using Infrared Imagery Provided by Leap Motion Controller," in J. Blanc-talon, et al. (eds.), Springer, Heidelberg, pp. 47-57, 2016.

[25] S. Visa, et al., "Confusion Matrix-based Feature Selection," *MAICS*, vol. 710, pp. 120-127, 2011.

BIOGRAPHIES OF AUTHORS

Dr. Dina Satybaldina is an Associate Professor at the National Research Nuclear University "MEPhI" (Moscow Engineering Physics Institute), Moscow, Russian Federation and a Professor of the Computer Engineering Department at L.N.Gumilyov Eurasian National University (ENU), Nur-Sultan, Kazakhstan. She received her PhD degree (Candidate Sciences in Physics and Mathematics) from E.A.Buketov Karaganda State University, Kazakhstan. In 2011 she obtained PhD degree in Computer Science from al-Farabi Kazakh National University, Almaty, Kazakhstan. Her research interests include Signal Processing, Computer Security, Machine Learning and Human-computer interaction. She has more than 75 publications related to these areas in various international journals and conference proceedings and has co-authored one book. Dr. Satybaldina is a Member of IEEE Communications Society and an active reviewer in many journals of the areas of computer science and computer security.

Gulzia Kalymova is currently a PhD student in the Computer Engineering Department at L.N.Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan. She received her Bachelor and Master degrees in Mathematics in Kazakh National University after Al-Farabi, Almaty, Kazakhstan. Her current research interests include Human-Computer Interactions, Machine Learning, Gesture Recognition, Computer Vision and Image Processing. She has published around 10 research papers in different international and national journals and conferences.