AUTOMATIC DETECTION AND RECOGNITION OF 3D MANUAL GESTURES FOR HUMAN-MACHINE INTERACTION

D. Ryumin, I. Kagirov, D. Ivanko, A. Axyonov, A. A. Karpov

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, SPIIRAS, Saint-Petersburg, Russian Federation – dl_03.03.1991@mail.ru, kagirov@iias.spb.su, denis.ivanko11@gmail.com, a.aksenov95@mail.ru, karpov@iias.spb.su

Commission II WG II/5

KEY WORDS: Sign Language, Gestures, Face Detection, Computer Vision, Machine Learning, Neural Networks

ABSTRACT:

In this paper, we propose an approach to detect and recognize 3D one-handed gestures for human-machine interaction. The logical structure of the modules of the system for recording a gestural database is described. The logical structure of the database of 3D gestures is presented. Examples of frames showing gestures in the format of Full High Definition, in the map depth mode and in the infrared illustrated. Models of a deep convolutional network for detecting faces and hand shapes are described. The results of automatic detection of the area with the face and the shape of the hand are given. Identified the distinctive features of the gesture at a certain point in time. The process of recognizing 3D one-handed gestures is described. Due to its versatility, this method can be used in tasks of biometrics, computer vision, machine learning, automatic systems of face recognition, sign languages.

1. INTRODUCTION

In the modern information society, the task of increasing the level of automation and robotization of all human activities is one of the most important (Ryumin and Karpov, 2017). In this regard, scientists and leaders of developed, as well as developing countries, in collaboration with world-class research centers and companies, are paying attention to technologies for effective, natural and universal human interaction with computers and robots (Ivanko et al., 2018a).

Currently, interactive information systems are used in the areas of social services, medicine, education, robotics, military industry, public service centers, as well as for interacting with people in various emergency situations (Toyota Global Site, 2018). In addition, robots assistant, which are aimed at interacting with people to perform certain tasks, are becoming increasingly popular. In this case, many classic interfaces are not suitable enough. Instead, more intuitive and natural interfaces are necessary (gestural, speech, multimodal) (Ivanko et al., 2018b) etc.). For example, gestures can transmit simple commands to the robot. Gestures pass an unambiguous meaning and are effective at some distance from the robot, even in noisy environments when acoustic speech is ineffective. It is well-known, that hearing-impaired people are limited in their ability to communicate with hearing people. Sign language interpreters are usually needed when contacting various instances, however, the number of free translators is often not enough to satisfy the demand for them. Therefore, for hearing impaired people, sign language recognition technologies, using which they can interact with assistive mobile information robots, are necessary (Guo and Yang, 2016). For example, scientists from the American Institute of Robotics at Carnegie Mellon University (CMU) are working on a system that can analyze body language and gestures right up to the position of the fingers. This study is presented at the Computer Vision and Pattern Recognition Conference (Cao et al., 2018).

This paper presents an approach to automatic detection and recognition of both: static and dynamic 3D one-handed gestures in real time using an optical camera and a depth sensor (Kinect v2, 2019).

2. DATASET

For the purpose of training models and testing the approach, we used own database of 3D gestures of Russian sign language. The logical structure of the database of 3D gestures is presented in Figure 1.

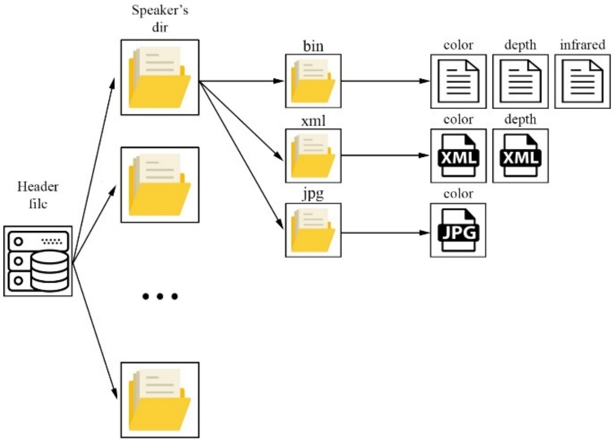

Figure 1. The logical structure of the database of 3D gestures of Russian sign language

For each gesture shown by the demonstrator, the following information is stored in the database as separate files on the disk:

- Video recordings of the gesture (color format, optical resolution 1920x1080 pixels (FullHD), for a depth map and an infrared mode — 512x424 pixels, frame rate — 30 frames per second);
- Data on the coordinates describing the position of the skeleton on the video;
- Images selected frame by frame from the video required for labeling.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W12-179-2019

This database was collected using the developed automatic 3D video stream recording system with the Microsoft Kinect 2.0 rangefinder sensor. The overall architecture of the developed system (MulGesRecDB) is presented in Figure 2 (Ryumin et al., 2019).

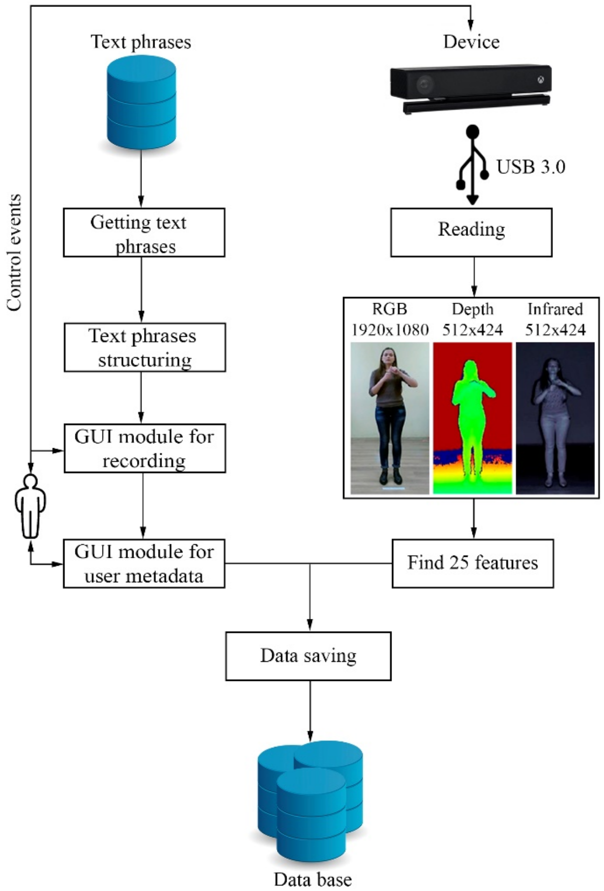

Figure 2. The logical structure of the modules of the MulGesRecDB system for recording a 3D gestural database

The words (phrases) were displayed on the screen and demonstrated by Russian sign language signers for 5 times at least. That is meant for further training of the automatic sign language recognition system based on probabilistic neural network models. The Kinect 2.0 camera was mounted at an optimal distance of 1.5-2 meters from the demonstrator. The demonstrators were recorded on a light background which are presented in Figure 3a. A distinctive feature of the presented database is the recording of gestures in a three-dimensional format (3D), which makes it one of a kind resource of the Russian sign language. The 3D format was obtained due to the fact that the Kinect 2.0 camera has the ability to record not only video data in color format with a resolution of 1920x1080 (FullHD), but also in the infrared (Figure 3b) and in the depth map mode (Figure 3c). The color designations on the depth map correspond to the ranges of the spectrum of visible light. In other words, the groups of dots most distant from the camera are colored red, and the closest ones are colored purple. The objects between these points are painted over with shades of green and yellow (Figure 3b).

A total of 13 speakers were recorded. All demonstrators speak Russian sign language. In addition, before recording the 3D database, the entire dictionary of gestures developed by the authors was adjusted and standardized by specialists from the rehabilitation center for people with hearing problems. All gestures were accompanied by oral articulation, but no audio data was recorded. The total number of gesture phrases was 2132. The main subject of the recorded database is devoted to food in the supermarket. In total, the database has 48 different one-handed gestures and 116 two-hand gestures. The method presented in this paper is aimed on automatically detecting and recognizing only 3D one-handed gestures. The static or dynamic orientation of a one-handed gesture in this case not considered.

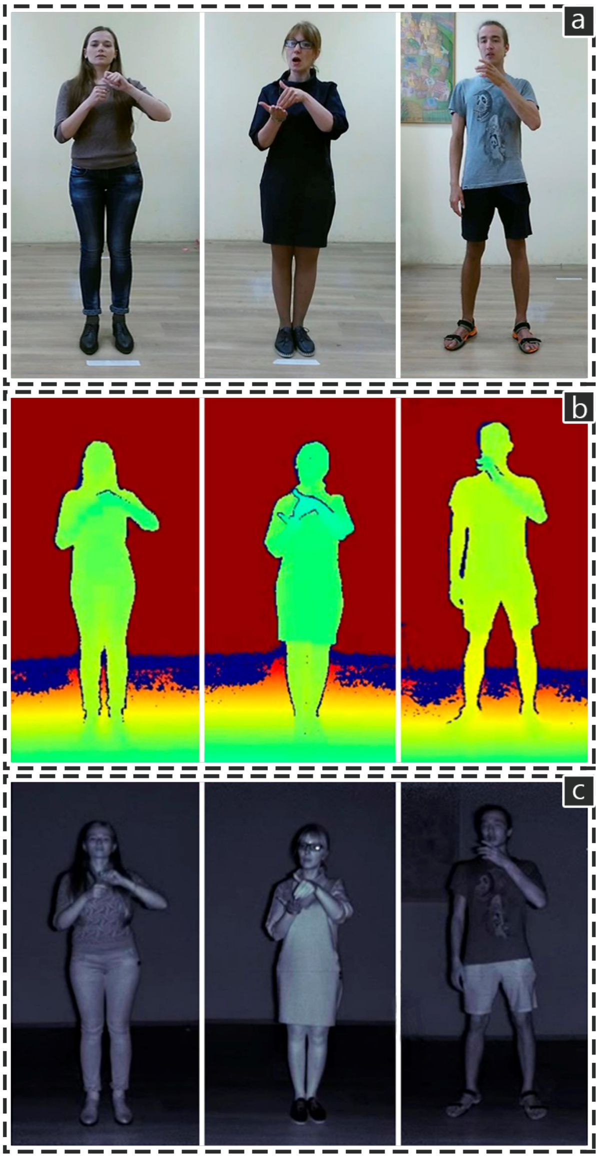

Figure 3. Examples of frames showing gestures in FullHD format (top row), in the depth map mode (middle row) and in the infrared range (bottom row)

3. DESCRIPTION OF THE METHOD

The input of the developed method is video data in a two-dimensional (RGB camera) and three-dimensional (depth map) format downloaded from the video file or directly from the sensor Microsoft Kinect v2. The resolution of color video frames is 1920x1080 pixels (FullHD), the resolution of depth map is 512x424 pixels with a frequency of 30 frames per second. The color quality for two-dimensional data is 8 bits, and for three-dimensional is 16 bits. Synchronous stream processing of two-dimensional (2D) and three-dimensional video data (3D) is carried out. On each frame, using a depth map and a software development kit (SDK) supplied with the Kinect v2 sensor, people are searched at a distance of 1.2 to 3.5 meters from the camera and 3D 25 reper-dots skeleton model is calculated for each person. Then, the 3D coordinates are converted to 2D and rectangular areas are formed with people on the 2D image (Figure 4a) and their 2D 25 reper-dots skeleton models (Figure 4b).

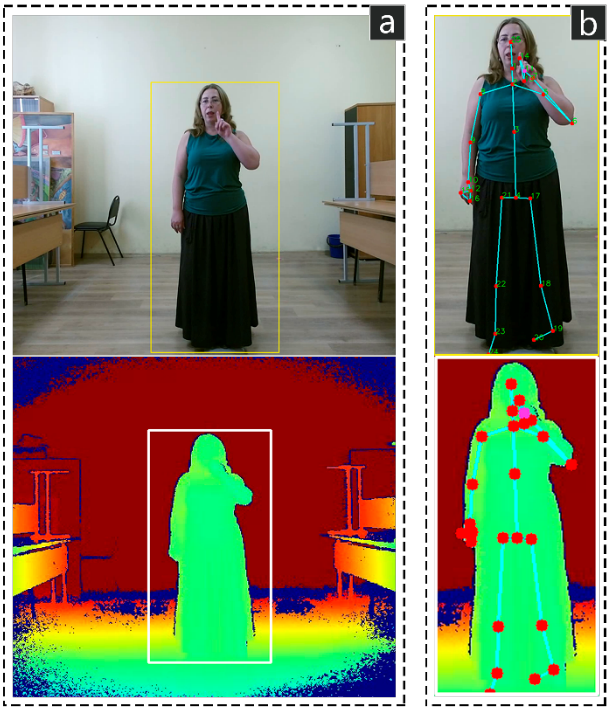

Figure 4. An example of calculating the 3D and 2D 25 reper-dots model of the skeleton for each person

Further, two models of the deep convolutional network are used to detect the face and hand shapes, presented in Figure 5.

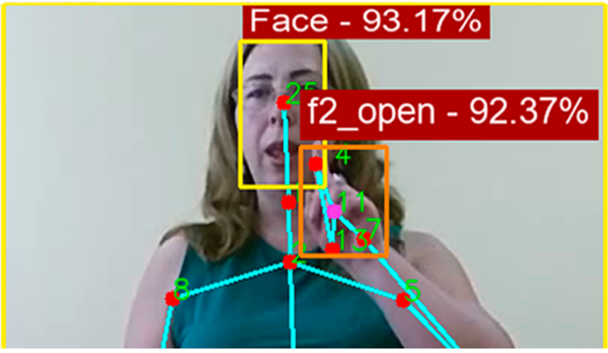

Figure 5. An example of face detecting area and hand-shaped area

The face detector is based on the structure of Single Shot MultiBox Detector (SSD (Liu et al., 2016)) with reduced network model ResNet-10 (He et al., 2016). This detector is included in the Deep Neural Networks (dnn) module of computer vision library Open Source Computer Vision Library (OpenCV, 2019). The model was trained using images available from the web, but the source is not disclosed. OpenCV provides 2 models for this face detector:

1. Floating point 16-bit version of the original caffe implementation;
2. 8-bit quantized version using Tensorflow.

We use the caffe model. This detector has the following features:

- Works in real time both on the central processing unit (CPU) and on the graphics processing unit (GPU);
- Works for different face orientations – up, down, left, right, side-face etc;
- Works even under substantial occlusion;
- Detects faces across various scales (detects big as well as tiny faces).

Hand shape detector based on SSD structure with MobileNetV2 network model (Sandler et al., 2018). This detector was trained on a multimedia database of 3D gestures of the Russian sign language collected and labeled by the authors. The MobileNetV2 network architecture is presented in Figure 6.

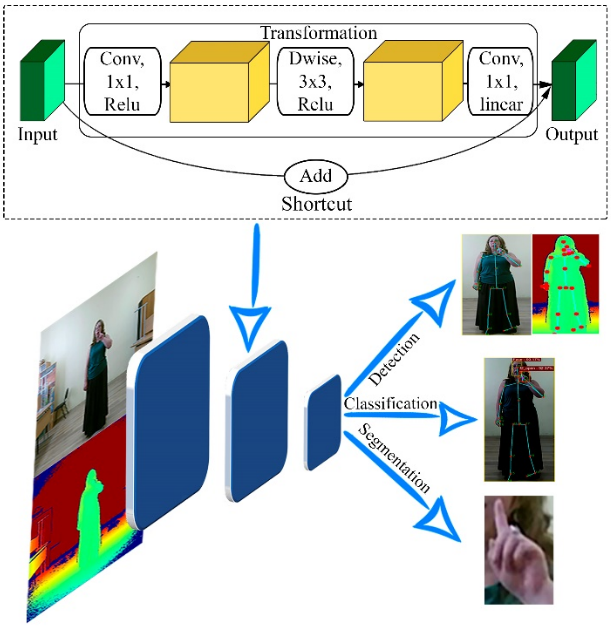

Figure 6. Overview of MobileNetV2 Architecture

To highlight the area with a gesture, a LabelImg (LabelImg, 2019) tool was used to graphically annotate images. Annotations saved as eXtensible Markup Language (XML) files in PASCAL Visual Object Classes (VOC (Everingham, et al., 2010)) format, which is used in ImageNet (Krizhevsky et al., 2012). In addition, it also supports YOLO (Redmon et al., 2016) format. For the detector training, the first 4 repetitions of the gesture with the coordinates of the necessary skeletal reper-dots were used. That is considered as train data, and the rest is used as test data.

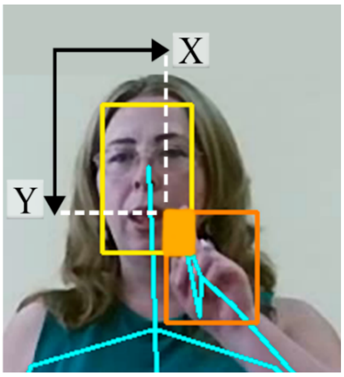

Figure 7. An example of calculating the distinctive characteristics of a gesture at a certain point in time

Then, coordinates of found faces and hand shapes are normalized. Based on the normalized coordinates, the 3D distance from the face to the hand is calculated. The area of the intersection of the hand and face coordinates is calculated as well (Figure 7).

As a result, the distinctive features of a gesture at a certain point in time are following:

- Normalized 3D distance from face to hand (gesture articulation zone);
- Normalized face and hand crossing area;
- Hand shape.

Testing has proven that it is possible to process video sequences of 23 frames on average in 1 second.

The final step of the method is to recognize one-handed gestures using the described features of the gesture and modern machine learning algorithms that are included in tools such as scikit-learn (Scikit-learn, 2019), Tensorflow Object Detection API (Huang et al. 2017), Keras (Keras, 2019), which simplify the creation, training, and deployment of both static and dynamic object detection models in real time. In the current study, recurrent neural network (RNN) with long short-term memory (LSTM (Hochreiter, 1997)) is used.

The features derived from 4 repetitions of the gesture of every demonstrator were used for training. The total number of training videos was 2496. LSTM was trained on the generated features of 15 frames sequences with 10 steps and the learning rate of 0.003. The neural network was trained using the Keras tool (Keras, 2019) with the RMSprop (Graves, 2014) optimization algorithm and asynchronous gradient descent.

Calculation of the average speed of work of the proposed method is to recognize one-handed gestures was made on computers with different performances, whose parameters are presented in Table 1.

Processor	Random access memory, GB	Type of hard disk	Processing speed of frame, ms
Intel Core i7- 8850H 2,6 GHz	16	SSD	≈80
Intel Core i5- 4210H 2,9 GHz	16	SSD	≈165

Table 1. The average processing speed of video frames using different computer systems

The results of the experiments showed that the average recognition accuracy, calculated by the formula (1), is 0.63 (63%):

$$x_c = \frac{x_1+x_2+\cdots+x_n}{n} \tag{1}$$

where n = number of gestures
x_n = gesture recognition accuracy

These results were obtained on 48 different 3D one-hand gestures from the multimedia database collected by the authors.

4. CONCLUSIONS AND FUTURE WORK

Thus, a method for detection and recognition of one-handed 3D gestures in real time, based on modern computer vision algorithms and deep neural networks, has been proposed. Due to its versatility, this method can be used in tasks of biometrics, computer vision, machine learning, automatic face recognition and sign language recognition.

In further research, we will use approaches based on 3D convolution and convolutional LSTM to recognize multimodal gestures. We also plan to expand the multimedia database with new demonstrators.

ACKNOWLEDGEMENTS

This research is financially supported by the Ministry of Science and Higher Education of the Russian Federation, agreement No. 14.616.21.0095 (reference RFMEFI61618X0095).

REFERENCES

Ivanko et al., 2018a. Ivanko, D., Karpov, A., Fedotov, D., Kipyatkova, I., Ryumin, D., Ivanko, Dm., Minker, W., Zelezny, M. Multimodal Speech Recognition: Increasing Accuracy using High Speed Video Data. *Journal on Multimodal User Interfaces*, Springer, Vol. 12, Iss. 4. pp. 319-328, https://doi.org/10.1007/s12193-018-0267-1.

Ivanko et al., 2018b. Ivanko, D., Ryumin, D., Axyonov, A., Železný, M. Designing Advanced Geometric Features for Automatic Russian Visual Speech Recognition. In Proc. 20th International Conference on Speech and Computer SPECOM-2018, Leipzig, Germany, Springer, LNAI, Vol. 11096, pp. 245-254.

Ryumin and Karpov, 2017. Ryumin, D., Karpov, A. Towards Automatic Recognition of Sign Language Gestures using Kinect 2.0. In Proc. 19th International Conference on Human-Computer Interaction, HCII, Vancouver, Canada, Springer LNCS, Vol. 10278, pp. 89-104.

Toyota Global Site, 2018. Partner Robot, http://www.toyota-global.com/innovation/partner_robot (30 March 2019).

Guo and Yang, 2016. Guo, X., Yang, T. Gesture recognition based on HMM-FNN model using a Kinect. J. Multimodal User Interfaces, Springer, Vol. 11, Iss. 1. pp. 1-7, https://doi.org/10.1007/s12193-016-0215-x.

Kinect v2, 2019. Kinect for Windows SDK 2.0, https://www.microsoft.com/en-us/download/details.aspx?id=44561 (30 March 2019)

Ryumin et al., 2019. Ryumin, D., Ivanko, D., Axyonov, A., Kagirov, I., Karpov, A., Zelezny, M. Human-Robot Interaction with Smart Shopping Trolley using Sign Language: Data Collection. Proceedings of IEEE International Conference on Pervasive Computing and Communications, PerCom-2019, Kyoto, Japan, in press.

Cao et al., 2018. Cao, Z., Hidalgo, G., Simon, T., Wei, S-E, Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. IEEE Conference on Computer Vision and Pattern Recognition, CVPR-2018, arXiv preprint arXiv:1812.08008.

Liu et al., 2016. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C-Y., Berg A. SSD: Single Shot MultiBox Detector. European conference on computer vision, ECCV 2016, Springer, Cham, pp. 21-37.

He et al., 2016. He, K., Zhang, X., Ren, S., Sun, J. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition, CVPR-2016, Las Vegas, NV, pp. 770-778.

OpenCV, 2019. OpenCV library, https://opencv.org (30 March 2019).

Sandler et al., 2018. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. IEEE Conference on Computer Vision and Pattern Recognition, CVPR-2018, arXiv preprint arXiv:1801.04381v3.

LabelImg, 2019. LabelImg is a graphical image annotation tool, https://github.com/tzutalin/labelImg (30 March 2019).

Everingham, et al., 2010. Everingham, M., Van Gool, L., Williams, C. K., Winn, J., Zisserman, A. The pascal visual object classes (voc) challenge. International journal of computer vision, Vol. 88, Iss. 2, pp. 303-338.

Krizhevsky et al., 2012. Krizhevsky, A., Sutskever, I., Hinton, G. E. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, pp. 1097-1105.

Redmon et al., 2016. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788.

Scikit-learn, 2019. Scikit-learn Machine Learning, https://scikit-learn.org (30 March 2019).

Huang et al. 2017. Huang, J., Rathod. V., Sun. Ch., Zhu. M., Korattikara. A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR-2017, pp. 3296-3297.

Keras, 2019. Home – Keras Documentation, https://keras.io (30 March 2019).

Hochreiter, 1997. Hochreiter, S., Schmidhuber, J. Long short-term memory. Neural computation, Vol. 9, Iss. 8. pp. 1735–1780.

Graves, 2014. Graves, A. Generating Sequences with Recurrent Neural Networks. arXiv preprint arXiv:1308.0850v5.