METHOD OF MULTI-MODAL VIDEO ANALYSIS OF HAND MOVEMENTS FOR AUTOMATIC RECOGNITION OF ISOLATED SIGNS OF RUSSIAN SIGN LANGUAGE

A. Axyonov, D. Ryumin, I. Kagirov
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg
Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russian Federation – (axyonov.a, ryumin.d, kagirov)@iias.spb.su

Commission II, WG II/5

KEY WORDS: Sign Language, Gestures, Speech Recognition, Computer Vision, Machine Learning, Neural Networks.

ABSTRACT:

This paper presents a new method for collecting multimodal sign language (SL) databases, which is distinguished by the use of multimodal video data. The paper also proposes a new method of multimodal sign recognition, which is distinguished by the analysis of spatio-temporal visual features of SL units (i.e. lexemes). Generally, gesture recognition is a processing of a video sequence, which helps to extract information on movements of any articulator (a part of the human body) in time and space. With this approach, the recognition accuracy of isolated signs was 88.92%. The proposed method, due to the extraction and analysis of spatio-temporal data, makes it possible to identify more informative features of signs, which leads to an increase in the accuracy of SL recognition.

1. INTRODUCTION

Despite the great practical potential of automatic sign language (SL) recognition systems, the problem of effective SL recognition has not yet been resolved due to serious differences in the vocabulary and grammatical structure of spoken and sign languages, so that a straightforward application of spoken language recognition methods to SL would be pointless (Stokoe W.C, 2005). This explains why nowadays there are no fully automated operating models and methods for sign language recognition systems. In order to create full-fledged models of this kind, a deep semantic and grammatical analysis of spoken languages is required (Battison R., 1978), implying a lot of preliminary work on creation of algorithms for text analysis, as well as databases. It is worth to mention that the abovementioned problems are caused by the general lack of universal methods for creating multimodal SL corpora. Moreover, there has been felt a lack of methods and algorithms that increase the efficiency of machine learning and the accuracy of automatic SL recognition, making use of various video capture devices that allow obtaining not only high-quality images in optical mode, but also additional data from the coordinates of graphical areas of interest (depth map mode, infrared mode, etc.) (Ryumin et. al., 2017).

A new method for collecting multimodal SL databases in this paper is distinguished by the use of multimodal video data. SL corpora are required for testing and comparing sign (hand gesture) recognition methods and algorithms. The SL corpora collection technique should include a layer of annotation of images obtained from the video stream (Kagirov et. al., 2020). Such annotation must be suitable for computer vision and machine learning tasks, that is, it should be oriented not only on linguistic tasks, but also on image recognition problems. In other words, the annotation should be based on features that could be used not only for linguistic notation, but also for computer analysis and gesture recognition. Using this technique, a collection and annotation of the multimodal corpus of Russian sign language was carried out. The used approach includes two main stages, involving a sequential execution of eight steps.

Besides methodology, the paper also proposes a new approach of multimodal sign recognition. This method is distinguished by the analysis of spatio-temporal visual features of SL units (i.e. lexemes). Generally, gesture recognition is a processing of a video sequence, which helps to extract information on movements of any articulator (a part of the human body) in time and space. The exception are static gestures, implying no articulator movements. Complex gestures cause quite serious recognition problems due to the relatively small size of the images of the articulators in comparison with the entire scenery. Moreover, the task of SL gestural information recognition includes other important matters, such as the size of the recognition dictionary, the variability of signs (including signer's unique language habits), and the parameters of the information transmission channel. The lexical components of SL (meaningful, significant hand gestures) are classified according to several parameters: the handshape, the localization of articulators, the manner of movement, mimics, articulation (Brentari D., 1998). The task of isolated signs recognition is important; however, an adequate processing of a sign series (including coarticulation problems) seems to be of more importance. Therefore, it is reasonable to build the SL recognition process taking into account the spatio-temporal component of SL utterances.

2. RELATED WORK

Recently, computer vision- and machine learning-based approaches to detecting and recognizing hand gestures have gained most popularity, since they support contactless human-machine interaction tools (Kaur et. al., 2016). However, there are also many problems due to the use of various hardware tools for gesture video capturing (optical, infrared, thermal and other cameras) (Sonkusare et. al.,

2015). Among the former are: 1) constant light changes; 2) occlusion effects; 3) dynamic background; 4) processing time, heavily depending on resolution and frame rate; 5) additional objects of the foreground and background resembling to human hands and/or having the same color (Garg et. al., 2009; Murthy et. al., 2009).

In (Uddin et. al., 2016), an approach to hand gesture recognition is used, that makes use of conversion of a RGB color image to HSV (Hue Saturation Value) color space. Then Gabor filters are applied to extract features; the filters scale and rotate the image in 5 and 8 different variations. The output is a convolution of the original image with the filtered images.

It is worth mentioning of convolutional neural networks that take raw images as input, independently extract distinctive visual features, and classify hand gestures (Alnaim et. al., 2019; Chung et. al., 2019).

The work (Xi et. al., 2018) presents a 3D approach for hand segmentation using the depth map of the Kinect v2 sensor, which determines the locations of the fingers using three-dimensional connections, Euclidean and geodesic distances (in English geodesic distance) from the pixels of the hand skeleton. Another 3D approach to hand gesture recognition based on a machine learning model using bidirectional convolutional neural networks is presented in (Devineau et. al., 2018).

In (Premaratne et. al., 2017), the authors propose a method of tracking hand gestures, which is based on the trajectories of the center of mass. In this case, 16 graphemes of the English alphabet were recognized, which were drawn by the hand in the air by the signer. The gesture classification algorithm applied was HMM.

In order to analyze dynamic hand gestures, networks with long short-term memory (abbreviated as LSTM) are also used, receiving consecutive frames in input. For instance, in (John et. al., 2016), a hybrid approach to hand gesture recognition was proposed, implying feeding a raw image to the input of a CNN, then the LSTM network is used for the task of hand gesture classification.

Approaches to hand gestures recognition, based on 3D models of the hand, use information about the distance of visual elements, making it possible to form a volumetric model of the hand. In (Tekin et.al., 2019), a model for recognizing a hand action using a single RGB image was proposed. In (Malik et. al., 2018), a new algorithm based on a 3D CNN was proposed, which learns to detect a hand from a 3D image. In (Ryumin et.al., 2019) also uses an approach to detect and recognize 3D one-handed gestures using CNN to recognize hand configurations. The drawbacks of 3D-approaches include the need for large datasets and high computational costs.

Current research results make it clear that DNN-based machine learning methods, if compared to "classical" approaches (Ivanko et. al., 2018; Ryumin et. al., 2020), which are based on linear classifiers (such as SVV, support vector machine), show quite good results in segmentation, classification, as well as recognition of both static and dynamic gestures.

3D CNNs (Ji et. al., 2010) can be used for simultaneous extraction of short-term spatio-temporal features. However, LSTM networks (Hochreiter et. al., 1997; Ryumina et. al., 2020) are best suited for storing temporal features. Therefore, it is argued that it is reasonable to use a 3D convolutional neural network (Ji et. al., 2010; Ryumina et. al., 2020) to extract short-term spatio-temporal characteristics and then use LSTM to extract spatio-temporal relationships from video sequences. Therefore, a 3D convolutional LSTM neural network, due to the storage of 3D spatial information, can form more efficient spatial and temporal characteristics of a gesture.

3. METHODOLOGY

The analysis showed (Ryumin et. al., 2019) the complete absence of RSL corpora with multimodal (multiple data types) representation of signs. In addition, it is revealed that most of the existing RSL corpora are aimed at researching the process of nonverbal communication exclusively through hand gestures and excluding such no less important communicative techniques of natural interaction as facial expressions and human posture in general. These shortcomings of existing SL corpora (including RSL) have revealed the need to develop their own universal methodology for collecting and annotating the multimodal SL corpus, which can be used for such scientific tasks as: 1) researching the features of articulations of sign languages; 2) determining the linguistic content of sign statements; 3) training various neural network models aimed at automatic interpretation of statements in sign language into text representation.

The proposed methodology for creating multimodal gesture corpora is illustrated in Figure 1 and consists of two main stages, which involve the sequential execution of eight steps.

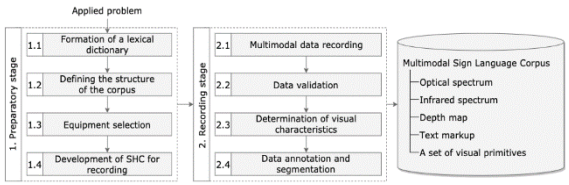

Figure 1. Diagram of the methodology for creating multimodal SL corpora.

Based on the application, for the solution of which the multimodal corpus of SL is written, the dialect of sign language is determined, and the first preparatory stage is performed, consisting of four steps.

1. The preparatory stages.
   1.1. The formation of the lexical dictionary is carried out dep ending on the scenarios for using the multimodal corpus of SL (for example, automatic gesture recognition in intelligent information systems, virtual reality, etc.).
   1.2. The structure of the multimodal corpus of SL is determined dep ending on the type of system (speaker-dependent / speaker-independent) and includes the number of informants with their total number of repetitions of letters / words / phrases / sentences (hereinafter lexical units) from the lexical dictionary. The logical component of the structure is presented in the form of a hierarchical model for the physical storage of gesture information and connections between its elements. As a result, all the

necessary data forms a file system, which consists of a root directory and a hierarchy of subdirectories with a set of files grouped by format.

1.3. The choice of equipment is made in accordance with a certain format of multimodal input video data, their quantity, and technical characteristics.

1.4. The final step of the preparatory stage is aimed at creating software for recording the multimodal corpus of SL.

2. The recording stages.

2.1. The recording of signers should be carried out in conditions that are close to the real conditions of using an automatic system, applying the developed software.

2.2. The recorded multimodal data must be checked for correctness and correspondence to lexical units from the previously formed lexical dictionary.

2.3. The determining the visual characteristics of a gesture depends on many factors, the main ones being the lexical vocabulary and the selected informants. It is important to understand that the demonstration of gestures by signers often differs in such details as hand configuration and localization, however, the nature of the gesture (movement) most often remains unchanged. Therefore, a notation that allows searching by corpus and sign recognition should consider not all hypothetically possible characteristics of a gesture, but only those that allow distinguishing one gesture from another.

2.4. In the last step, all collected multimodal data must be (semi) automatically annotated and segmented at the level of minimum gesture units (classes) in a semi-automatic or automatic way.

Using this methodology, a multimodal corpus of RSL elements was collected (hereinafter TheRuSLan). The TheRuSLan multimodal corpus contains video recordings of RSL gestures in color optical format, in depth map mode, and in the infrared range (Figure 2), making it a one-of-a-kind resource for RSL material.

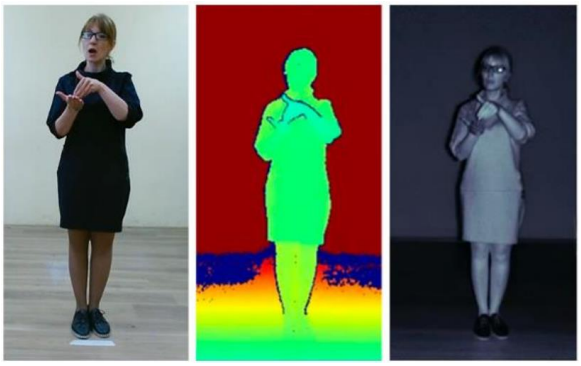

Figure 2. Examples of video frames showing RSL gestures in FullHD format, in depth map mode, and in the infrared range from the TheRuSLan multimodal corpus.

The presence of video data obtained from the depth map introduces a third dimension to the description, which allows more accurately determining the position of one object relative to another, in this case, the position of the hands relative to each other and the speaker's body. The distance between the active and passive hands and the distance of the hands from the body are a means of expressing a variety of lexical meanings in the SL.

4. DESCRIPTION OF THE METHOD

Generally, gesture recognition can be described as processing a video sequence, providing information about movements of articulators (hands, head etc.) in time and space (Cao et. al., 2018). Static gestures, however, are different, because the position of hands and fingers do not change (Oyedotun et. al., 2017). Besides, complex scenery on video frames causes serious recognition problems due to relatively small size of human hands in comparison to the entire scene. In addition, the task of recognizing gestural information of any sign language is characterized by other important features: the size of the recognition dictionary, signs varieties and signers' individual differences, characteristics of the transmission channel. The boundaries of words in the stream of continuous signing can be determined only in the process of recognition (decoding of signs) by selecting the optimal sequence of gestures that best matches the input stream using mathematical models. The lexical components of sign languages (meaningful signs) are built up of several components: hand(s) configuration, hand(s) localization, the manner of hand movement, facial expressions, articulation. The task of recognizing gestural information is important per se, however, a more urgent task is to understand the meaning of an utterance by a recognized series of gestures. Therefore, it is reasonable to build the gesture recognition process taking into account their spatio-temporal component. The functional diagram of the proposed method is shown in Figure 4.

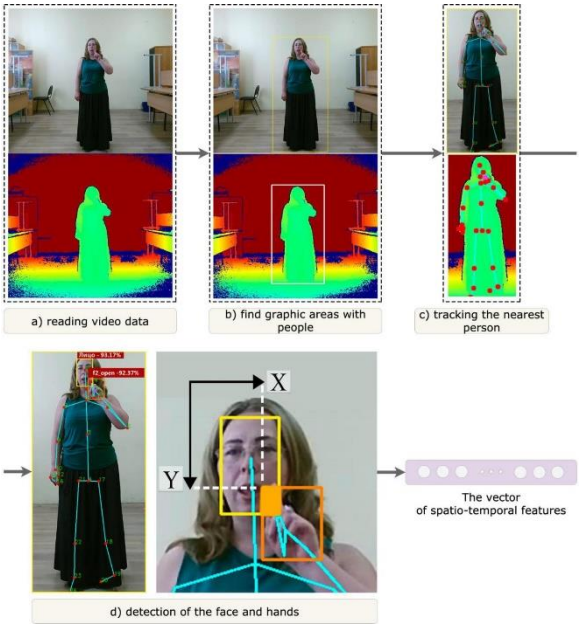

Figure 3. Illustration of the process of extracting spatio-temporal visual signs of a gesture.

The input multimodal video data of the method is a full-color (RGB) video stream and a depth map (Figure 3a), on which the signer demonstrates SL elements, standing at a distance of 1.2 to 3.5 m from the sensor. Color quality for RGB images is 8 bits per pixel with a video stream resolution of 1920 × 1080 (FullHD) pixels and a frequency of 30 fps, and for the depth map - 16 bits with a video stream resolution of 512 × 424 pixels and the same frame rate as color video stream. After that, a synchronous processing of modalities is performed.

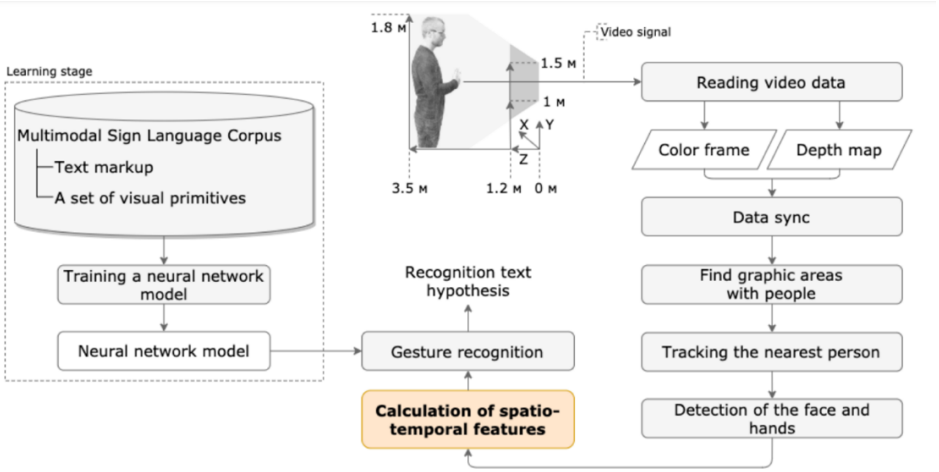

Figure 4. Functional diagram of the method of multimodal video analysis of hand movements and recognition SL elements.

For each frame of both modalities, a search for graphic areas containing people is performed (Figure 3b). Then, the z-axis of the three-dimensional space (depth map) determines the nearest human and sets tracking for him (Figure 3c).

4.1 Face and palm detection

At the next step, the graphic area of the face and the palms of the hands is detected within the formed rectangular area with the human (Figure 3d). Next, the spatio-temporal features (characteristics) of the reproduced SL element are calculated and normalized. The final stage is aimed at recognizing SL element, taking into account its spatio-temporal component.

A distinctive feature of the proposed new method of multimodal hand gesture recognition lies in the analysis of the SL elements as spatio-temporal (Figure 3).

In order to solve the problem of face detection, various methods of face detection were investigated using the multimodal Russian SL corpus TheRuSLan: 1) an improved Viola-Jones method (Viola et. al., 2004); 2) a method based on the Single Shot MultiBox Detector (SSD) architecture (Liu et.al., 2016) with a reduced model of the ResNet-10 network (He et. al.,2016); 3) based on HOG (Déniz et. al., 2011) and SVM methods (Chang et. al., 2011); 4) Max-Margin Object Detection (MMOD) method (King D.E., 2015). To score the 60 quality of detectors used quantitative indicators (metrics) in the object detection: Average Precision (averaged across all categories the values of average precision, hereinafter AP), AP_{50}, AP_{75}, $AP_{Small(S)}$, $AP_{Medium(M)}$, $AP_{Large(L)}$. Subscripts 50 and 75 define a minimum threshold crossing was found area with the annotated area, so in the case of 50% crossing positively detected area is the face area if the crossing with annotated area exceeds 50%. The subscripts S, M, and L show the value of the AP metric when the image is reduced to 32^2, from 32^2 to 96^2, and above 96^2, respectively. Comparative analysis is presented in the Table 1.

In the course of the experiments, it was revealed that for determining the graphic area of the face, the optimal face detector must be based on the Single Shot MultiBox Detector architecture with a reduced ResNet-10 network model, which is implemented in the open-source computer vision library OpenCV. When compared with other detectors, it was determined that it works at different face orientations, is resistant to occlusions, and also works in real time both on the Central Processing Unit (hereinafter CPU), and on the Graphics processor Unit (hereinafter GPU).

Face detector	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Viola-Jones	19	0,15	0,56	0,02	0,04	0,10	0,21
SSD OpenCV (Caffe)	62	0,41	0,86	0,25	0,06	0,37	0,46
SSD OpenCV (TensorFlow)	61	0,40	0,83	0,21	0,07	0,33	0,47
HOG & SVM (Dlib)	15	0,12	0,65	0,01	0,03	0,09	0,17
MMOD (Dlib)	9	0,08	0,44	0,01	0,02	0,07	0,13

Table 1. Comparative analysis of face detectors.

To recognize handshapes, several convolutional neural network (2D CNN) architectures with different hyperparameter configurations were investigated, which are included in the open-source software platform TensorFlow Object Detection API v1 and v2, as well as in the object recognition module of the open-source library Keras. Training of convolutional neural networks was performed using marked - up data with hand shapes and of 18 one-handed signs from the TheRuSLan multimodal corpus of RSL (RGB and depth map), which was divided into training and test sets in a ratio of 10:3 informants. The process of annotating the p alms of the hands to extract their shapes was carried out using the labeling tool LabelImg. Annotated areas are represented in the special PASCAL VOC format as XML text files. This format is widely used, for example, in the ImageNet visual database designed to study various approaches to recognizing visual objects.

It is revealed that the correct recognition of the hand shape with the elimination of false positives is performed under the following conditions:

– the best trained convolutional neural network model with EfficientDet-D7 architecture determines the shape of the hand;

– the center coordinate of the hand (element) received from the Kinect v2 sensor is located within the recognized hand-shaped area.

At the stage of formation of spatio-temporal features vectors, the coordinates of the areas of the face and the hands are calculated with a subsequent normalization. The 3D distance between the upper left coordinate of the face area and the same coordinate of the human hand area is calculated as

well. In addition, the intersection area of the face and hand regions is calculated. The informative spatio-temporal characteristics of a sign at a certain point in time are: 1) normalized 2D and 3D distances from face to hand (zone of gesture articulation); 2) normalized 2D area of intersection of the face and the hands (in the absence of intersection, the area is zero); 3) handshape (represented by a numerical value of the class).

4.2 LSTM neural network model

At the last step, it is proposed to recognize RSL one-handed signs using the LSTM neural network. In general terms, an LSTM network is a type of recurrent neural network. In turn, a recurrent neural network is a neural network that models events (phenomena) that change over time or sequence, for example, as sign recognition. This is done by feedback of the output of the neural network level at time t with the input of the same network level at time $t + 1$. However, the usual recurrent neural network has a disadvantage, which is a vanishing gradient. This problem occurs when the network tries to model a dependency within a long sequence of training set. This is because small gradients or weights (values less than 1) are repeatedly multiplied over several time steps, and therefore the gradients are compressed to zero. As a result, the weights of earlier steps will not be significantly changed, and therefore the network will not study long-term dependencies. The LSTM network solves this problem and that is why the choice was made in its favor. The architecture of the LSTM neural network model is shown in Figure 5.

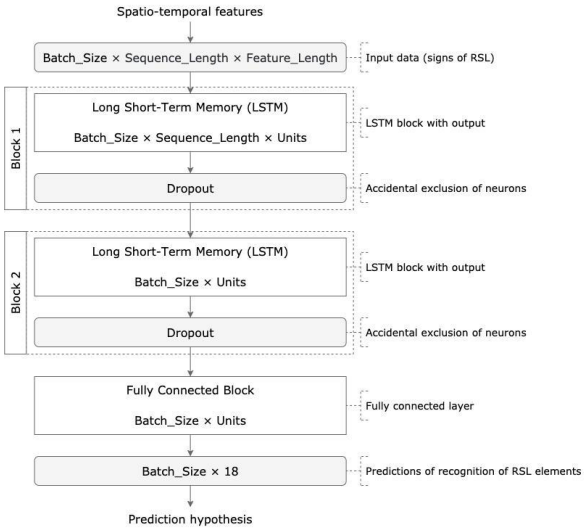

Figure 5. Architecture of the LSTM neural network model for RSL elements recognition.

As it can be seen, functional cores of signs, which consist of context-independent hand movements in relation to other signs, should be fed to the input of the LSTM network. In a more extended understanding of LSTM neural network takes a sequence of N frames × 4 values from the sign characteristics, in particular: normalized 2D and 3D distances from the face to the palm of the hand (represented by floating-point numbers); normalized 2D areas of the intersection of the face and the palm of the hand, which are also represented by floating - point numbers; shapes of the palm of the hand are integers. A comparative table of the proposed method with other methods is presented in Table 2.

For comparison, the following methods were selected: 1) the Restricted Graph-Based Genetic Programming (hereinafter RGGP) (Liu et. al., 2013); 2) the Elliptical Density Shape Model (hereinafter EDS) (Tung et. al., 2014); 3) the Multi-Stream Recurrent Neural Network (hereinafter MRNN) (Nishida et. al., 2015); 4) Recurrent 3D Convolutional Neural Network (hereinafter R3DCNN) (Molchanov et. al., 2016); 5) the Multi-Dimensional Convolutional Neural Networks (hereinafter MultiD-CNN) (Elboushaki et.al., 2020).

Method	Modality	Accuracy, %
RGGP	RGB	69.74
	Depth map	53.07
	RGB + depth map	74.28
EDS		77.43
MRNN	RGB	68.23
	Depth map	73.54
	RGB + depth map	79.98
R3DCNN		84.67
MultiD-CNN		87.38
Proposed method		88.92

Table 2. Comparison table of the proposed method with other methods.

All the methods from table 2 were implemented based on primary sources in the form of papers and were trained on 18 one-handed signs from the TheRuSLan multimodal corpus of RSL.

5. CONCLUSIONS AND FUTURE WORK

Thus, in paper a universal methodology for creating multimodal sign corpora is proposed, which is distinguished by the use of multimodal video data, with the use of which the collection and annotation of the multimodal corpus of the Russian sign language elements was carried out. Also, A new method of multimodal hand gesture recognition is proposed, which is distinguished by the analysis of spatio-temporal visual features of sign language elements.

Experiments have shown that the lowest recognition accuracy is shown by signs where the hand shapes are similar, and the articulation area is in the face region. With this approach, the recognition accuracy of isolated signs was 88.92%.

In further research, we plan to expand the multimedia database with new demonstrators. We also plan to research and develop new methods based on neural networks.

ACKNOWLEDGEMENTS

The research is financially supported by the Russian state research № 0073-2019-0005.

REFERENCES

Stokoe W.C., 2005: Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of deaf studies and deaf education*, 10(1), 3–37. doi.org/10.1093/deafed/eni001.

Kagirov I., Ryumin D., Axyonov A., Karpov A., 2020: Multimedia Database of Russian Sign Language Items in 3D. *Voprosy Jazykoznanija*., 20(1), 104–123. doi.org/10.31857/S0373658X0008302-1.

Battison R., 1978: Lexical borrowing in American Sign Language. *Linstok Press*.

Ryumin D., Karpov A., 2017: Towards automatic recognition of sign language gestures using kinect 2.0. *International Conference on Universal Access in Human-Computer Interaction*, 10278, 89-101. doi.org/10.1007/978-3-319-58703-5_7.

Brentari D., 1998: A prosodic model of sign language phonology. *MIT Press.*

Kaur H., Rani J., 2016: A review: Study of various techniques of Hand gesture recognition. *IEEE In International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, 1–5. doi.org/10.1109/ICPEICES.2016.7853514.

Sonkusare J.S., Chopade N.B., Sor R., Tade S.L., 2015: A review on hand gesture recognition system. *IEEE, In International Conference on Computing Communication Control and Automation*, 790–794. doi.org/10.1109/ICCUBEA.2015.158.

Garg P., Aggarwal N., Sofat S., 2009: Vision based hand gesture recognition. *World academy of science, engineering and technology*, 49(1), 972–977. doi.org/10.5281/zenodo.1074855.

Murthy G.R.S., Jadon R.S., 2009: A review of vision based hand gestures recognition. *International Journal of Information Technology and Knowledge Management*, 2(2), 405–410.

Uddin M.A., Chowdhury S.A., 2016: Hand sign language recognition for Bangla alphabet using Support Vector Machine. *IEEE, In Proceedings International Conference on Innovations in Science, Engineering and Technology*, 1–4. doi.org/10.1109/ICISET.2016.7856479.

Alnaim N., Abbod M., Albar A., 2019: Hand Gesture Recognition Using Convolutional Neural Network for People Who Have Experienced A Stroke. *IEEE, In the International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 1–6. doi.org/10.1109/ISMSIT.2019.8932739.

Chung H., Chung Y., Tsai W., 2019: An efficient hand gesture recognition system based on deep CNN. *IEEE, In International Conference on Industrial Technology (ICIT)*, 853–858. doi.org/10.1109/ICIT.2019.8755038.

Xi C., Chen J., Zhao C., Pei Q., Liu L., 2018: Real-time Hand Tracking Using Kinect. *In the International Conference on Digital Signal Processing*, 37– 42. doi.org/10.1145/3193025.3193056.

Devineau G., Moutarde F., Xi W., Yang J., 2018: Deep learning for hand gesture recognition on skeletal data. *IEEE, In the International Conference on Automatic Face and Gesture Recognition (FG)*, 106–113. doi.org/10.1109/FG.2018.00025.

Premaratne P., Yang S., Vial P., Ifthikar Z., 2017: Centroid tracking based dynamic hand gesture recognition using discrete Hidden Markov Models. *Neurocomputing*, 228, 79–83. doi.org/10.1016/j.neucom.2016.06.075.

John V., Boyali A., Mita S., Imanishi M., Sanma N., 2016: Deep learning-based fast hand gesture recognition using representative frames. *In the International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–8. doi.org/10.1109/DICTA.2016.7797030.

Tekin B., Bogo F., Pollefeys M., 2019: H+ O: Unified egocentric recognition of 3D hand-object poses and interactions. *IEEE, In the Conference on Computer Vision and Pattern Recognition*, 4511–4520. arXiv:1904.05349.

Malik J., Elhayek A., Stricker D., 2018: Structure-aware 3D hand pose regression from a single depth image. *In the International Conference on Virtual Reality and Augmented Reality*, 3–17. doi.org/10.1007/978-3-030-01790-3_1.

Ryumin D., Kagirov I., Ivanko D., Axyonov A., Karpov A.A., 2019: Automatic Detection and Recognition of 3D Manual Gestures for Human–machine Interaction. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W12, 179–183. doi.org/10.5194/isprs-archives-XLII-2-W12-179-2019.

Ivanko D., Ryumin D., Axyonov A., Železný M., 2018: Designing Advanced Geometric Features for Automatic Russian Visual Speech Recognition. *In Proceedings of the International Conference on Speech and Computer (SPECOM). Lecture Notes in Computer Science*, 11096, 245–254. doi.org/10.1007/978-3-319-99579-3_26.

Hochreiter S., Schmidhuber J., 1997: Long short-term memory. *Neural computation*, 9(8), 1735−1780. doi.org/10.1162/neco.1997.9.8.1735.

Ryumina, E., Karpov, A., 2020: Facial expression recognition using distance importance scores between facial landmarks. *Graphicon, CEUR Workshop Proceedings*. 2744, 1-10. ceur-ws.org/Vol-2744/paper32.pdf.

Ji S., Xu W., Yang M., Yu K., 2010: 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 35, 221−231. doi.org/10.1109/TPAMI.2012.59.

Ryumina, E.V., Karpov, A.A., 2020: Comparative analysis of methods for imbalance elimination of emotion classes in video data of facial expressions. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 20 (5(129)), 683-691. doi.org/10.17586/2226-1494-2020-20-5-683-691.

Ryumin D., Kagirov I., Axyonov A., Pavlyuk N., Saveliev A., Kipyatkova I., Zelezny M., Mporas I., Karpov A. A., 2020: Multimodal User Interface for an Assistive Robotic Shopping Cart. *Electronics*,9(12), 1–25. doi.org/ 10.3390/electronics9122093.

Ryumin D., Ivanko D., Axyonov A., Kagirov I., Karpov A., Železný M., 2019: Human-Robot Interaction with Smart Shopping Trolley using Sign Language: Data Collection. *IEEE In International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 949–954. doi.org/10.1109/PERCOMW.2019.8730886.

Cao Z., Hidalgo G., Simon T., Wei S-E., Sheikh Y., 2018: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Conference on Computer Vision*

and Pattern Recognition (CVPR). arXiv preprint arXiv:1812.08008.

Oyedotun O., Khashman A., 2017: Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications*, 28(12), 3941–3951. doi.org/10.1007/s00521-016-2294-8.

Viola P., Jones M., 2004: Robust real-time face detection. *International Journal of Computer Vision*, 57(2), 137–154. doi.org/10.1023/B:VISI.0000013087.49260.fb.

Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C-Y., Berg A., 2016: SSD: Single Shot MultiBox Detector. *European conference on computer vision (ECCV)*, 21–37. doi.org/10.1007/978-3-319-46448-0_2.

He K., Zhang X., Ren S., Sun J., 2016: Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. arXiv:1512.03385.

Déniz O., Bueno G., Salido J., De la Torre F., 2011: Face recognition using histograms of oriented gradients. *Pattern Recognition Letter*, 32(12), 1598–1603. doi.org/10.1016/j.patrec.2011.01.004.

Chang C.C., Lin C.J., 2011: LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3), 1–27. doi.org/10.1145/1961189.1961199.

King D.E., 2015: Max-margin object detection. arXiv preprint arXiv:1502.00046.

Liu L., Shao L., 2013: Learning discriminative representations from RGB-D video data. *In Twenty-Third International Joint Conference on Artificial Intelligence*, 1493–1500. https://ueaeprints.uea.ac.uk/id/eprint/62412.

Tung P., Ngoc L., 2014: Elliptical density shape model for hand gesture recognition. *In Proceedings of the Fifth Symposium on Information and Communication Technology (ICTD)*, 186–191. doi.org/10.1145/2676585.2676600.

Nishida N., Nakayama H., 2015: Multimodal gesture recognition using multi-stream recurrent neural network //*In Image and Video Technology*, 682–694. doi.org/10.1007/978-3-319-29451-3_54.

Molchanov P., Yang X., Gupta S., Kim K., Tyree S., Kautz J., 2016: Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. *IEEE In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 4207–4215. doi.org/10.1109/CVPR.2016.456.

Elboushaki A., Hannane R., Afdel K., Koutti L., 2020: MultiD-CNN: A multidimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. *Expert Systems with Applications*, 139, 1–25. doi.org/10.1016/j.eswa.2019.112829.