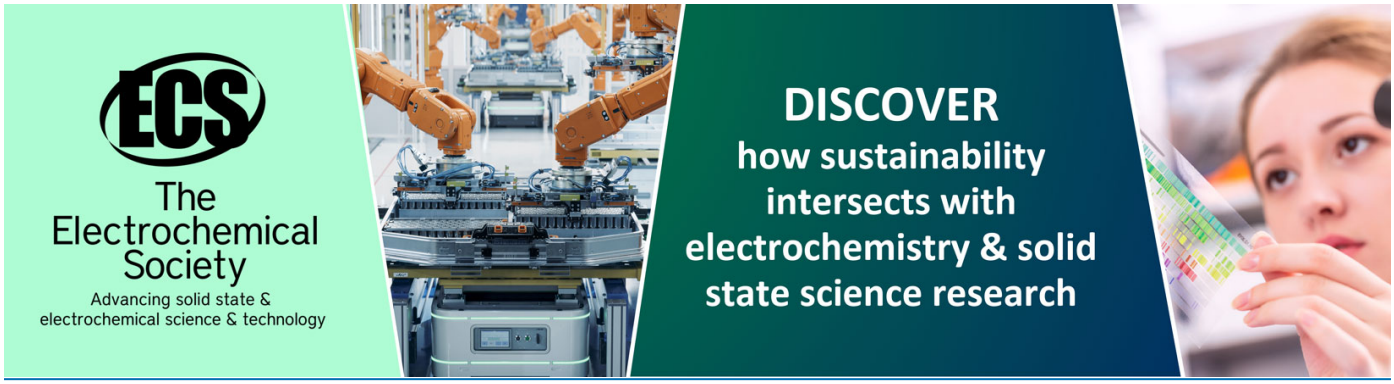
Development of a software module for recognizing the fingerspelling of the Russian Sign Language based on LSTM



Development of a software module for recognizing the fingerspelling of the Russian Sign Language based on LSTM

M G Grif[1], Y K Kondratenko[2]

[1] Novosibirsk State Technical University, 20, Karl Marx ave., Novosibirsk, 630073, Russian Federation
[2] Novosibirsk State University, 1, Pirogov str., Novosibirsk, 630090, Russian Federation
E-mail: grifmg@mail.ru, y.kondratenko@g.nsu.ru

Abstract. Russian Sign Language is a natural language that serves as a means of communication for people with hearing impairments. Currently, it is necessary to provide communication paths between hearing and deaf people, which requires the solution of several specific problems, one of which is gesture recognition. The paper presents a system for recognizing isolated static and dynamic gestures of the dactyl alphabet of the Russian Sign Language. The system is based on machine learning methods and is an LSTM with Mediapipe Hands as feature extractor. The network showed an F-measure value of 91% on test data.

1. Introduction

Sign languages are languages used for communication by people with hearing impairments. They are natural languages in which communication is achieved not through the articulation of sounds, but through gestures, changes in body position and facial expressions to create signs. The plane of expression of such languages is based on gesture-mimic basis. The functional and communication capabilities of these languages are not inferior to spoken languages. According to the 2010 All-Russian Population Census, the number of people who speak Russian Sign Language in Russia is 120.5 thousand people.

Russian Sign Language machine translation systems play an important role in solving the problem of ensuring communication between deaf and hearing. This is necessary due to the insufficient number of personnel in the field of sign language translation, as well as due to the not always desirable mediation in the communication of deaf and hearing citizens (for example, in matters of medicine and personal relations).

Deaf people use two types of sign systems: sign language and the dactyl alphabet. The dactyl alphabet (fingerspelling) is an auxiliary system of the Russian Sign Language, a system of hand signs corresponding to letters, an alphabet reproduced by hands (figure 1). The Russian fingerspelling refers to one-handed, copying (i.e. tends to resemble a dactyl sign with a letter) and alphabetic (i.e. it is shown in letters, not syllables). There are 33 signs in Russian dactylology (the same number as letters in the Russian alphabet). Pronunciation is conducted according to the rules of Russian spelling. The fingerspelling is used to pronounce words that do not have a special sign in sign language (for example, toponyms and personal names), as well as in the case when it is necessary to clarify the meaning of a word.

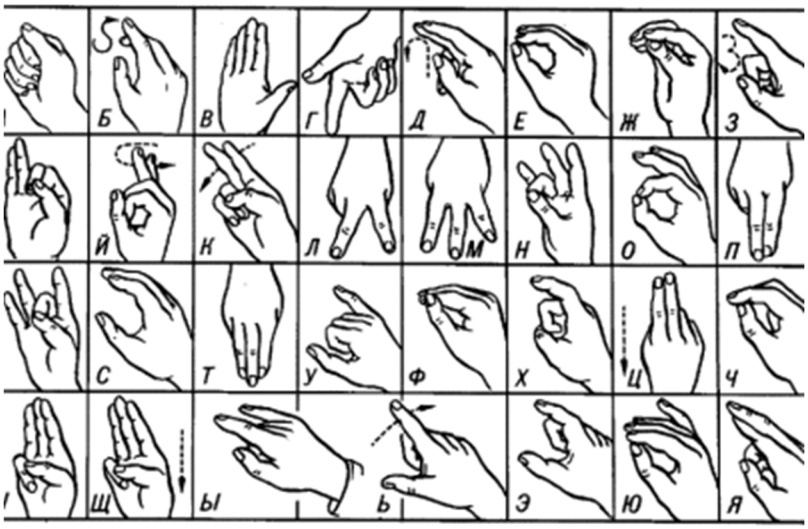


Figure 1. Russian dactylology

Before the 1960s, a gesture was considered an indivisible unit, but later three components that make up each gesture were identified: configuration, spatial position and movement. Nowadays, a characteristic such as the orientation of the hands in space relative to each other and the speaker's body and the non-manual component (facial expressions, articulation) are also highlighted. These parameters are allocated on the basis of the existence of minimal pairs of signs that differ only in one of these parameters and have different meanings.

However, the signs of the dactyl alphabet are specific, which makes it possible to ignore some of the described components. For example, the spatial position of the hand is not important - usually, gesturing with dactyl is carried out at the level of the gesturing person's chest with the hand bent at the elbow. However, the position can change without changing the semantic component. Since most of the Russian dactyl alphabet is represented by static gestures, the movement component is irrelevant for them. At the same time, for dynamic gestures, despite the fact that this component should be taken into account, the speed of movement and the nature of the gesture (for example, the movement is made abruptly or smoothly) do not matter, therefore movements can be characterized by a change of the configuration or position of the palm. Dactyl signs are reproduced only by hand, so the non-manual component also does not affect the semantic component of the message conveyed by the dactyl alphabet. Thus, the main component on the basis of which recognition can be carried out is the configuration of the palm.

Hand recognition is a challenging task for computers, because hands do not have high contrast patterns, overlap each other and close when moving. The solution to this problem involves the selection of a gesture or its components from a video sequence (or an image in the case of static gestures, when there is no need to process continuous hand movements) and their interpretation.

2. Related works

Existing gesture-recognition systems differ in several ways:

- Input data type. The data can be either a set of features obtained using special sensors or an image or video recording. An image usually represents an isolated static gesture, while video can represent both an act of performing an isolated gesture and a stream of speech in the dactyl alphabet.
- Data entry method. Data can be entered using a webcam, gloves with special sensors, cameras with depth sensors or motion controllers.
- Data processing and feature extraction.
- Way to identify and classify a gesture.

Over the past 15 years, a large number of works have appeared based on different approaches. For example, in the work of D. Warchoł et al. [1] for the Polish Sign Language, the Kinect sensor is used

for data input, while the work of B. Shi et al. [2], on the contrary, focuses on dactyl recognition from images obtained without the use of special technologies.

Machine learning methods are actively used. A common solution for both gesture recognition of sign languages in general and for fingerspelling recognition is to use convolutional neural networks. It is presented in the work of N. Aloysius [3]. At the same time, in studies devoted to the recognition of not only static but also dynamic gestures, methods based on the use of LSTM, a type of recurrent neural networks designed to work with temporary data, are gaining popularity. An example is the work of S.K.M. Lee and others [4], which recognizes 26 dactyls of American Sign Language with an accuracy of up to 99%. It is also not uncommon to use architectures that combine convolutional neural networks and LSTMs, for instance, the system described in the work of S. Aparna and M. Geeta [5].

Many of the existing works are devoted to the recognition of the dactyl alphabet of the American Sign Language. In systems aimed at recognition from two-dimensional images, this is due to the presence of relatively large data sets that can be used for training and testing machine learning systems (for example, MNIST-ASL, MUASL, and others).

In 2013, at the IEEE international conference on the application of information and communication technologies, the Indian Sign Language dactyl recognition system was presented by V. Adithya, P. R. Vinod and others [6]. The input data is an image taken from a two-dimensional camera. Images are scaled, with a palm highlighted on each. The selection of the palm is based on the detection of skin color, then the entire image is converted into a binary color gamut, in which the palm is highlighted in white and everything else in black. The resulting images are processed using the distance map method, the data is converted into vector form. A neural network is used as a classification tool - a perceptron with two hidden layers. The achieved accuracy is 91%.

In 2020, H. Lukman and other researchers proposed a Gabor Filter-based recognition system for detection and convolutional neural networks for classifying American Sign Language and Arabic Sign Language Dactyl Alphabet gestures [7]. The system takes as input an image obtained from a two-dimensional camera. Image processing is carried out using the Gabor filter. The classifier is a convolutional neural network consisting of three convolutional layers, two layers and two fully connected layers. Recognition accuracy reaches 99%.

Studies devoted to the recognition of Russian Sign Language gestures are described in the works of A. Prikhodko, M. Grif and M. Bakaev [8], I. Makarov et al. [9]. The creation of recognition systems for Russian Sign Language gestures is complicated by the lack of sufficiently large data sets.

We can conclude that nowadays there is no sufficiently accurate and user-friendly system for recognizing the dactyl alphabet of the Russian Sign Language.

3. Description of the proposed model

This paper presents a system for recognizing isolated dynamic and static gestures of the dactyl alphabet of the Russian Sign Language. The system is based on machine learning methods.
The algorithm is implemented using the Python 3.9.2 object-oriented programming language.

3.1. Data set

A data set collected by researchers from Novosibirsk State Technical University was used to train and test the model. The set contains 13412 photographs and videos of 33 signs of the Russian dactyl alphabet (figure 2).

There were four gesturing people involved in creating the dataset. The photographs and videos are of a waist-length gesture performing a gesture of the dactyl alphabet with his right arm bent at the elbow at chest level.

In the process of work, we decided to expand the data set in order to achieve a higher accuracy of the model on real data, since the limited number of gestures and insufficient variety reduced the applicability of the model for examples close to real ones. Another 1566 images and videos collected on Youtube video hosting, as well as specially recorded videos for this purpose, were added to the data set.

3.2. Feature extraction

The cross-platform framework Mediapipe Hands for hand and finger tracking proposed by Google [10] was used as a tool for feature extraction. This solution is available for working with both a live video stream from a webcam and photo and video files. In the case of Python, MediaPipe is available as a Python module package. The model can be used to detect the palm and calculate 21 landmarks for each detected hand, where each landmark corresponds to a specific point on the palm and represents three coordinates - x, y and z (figure 2).

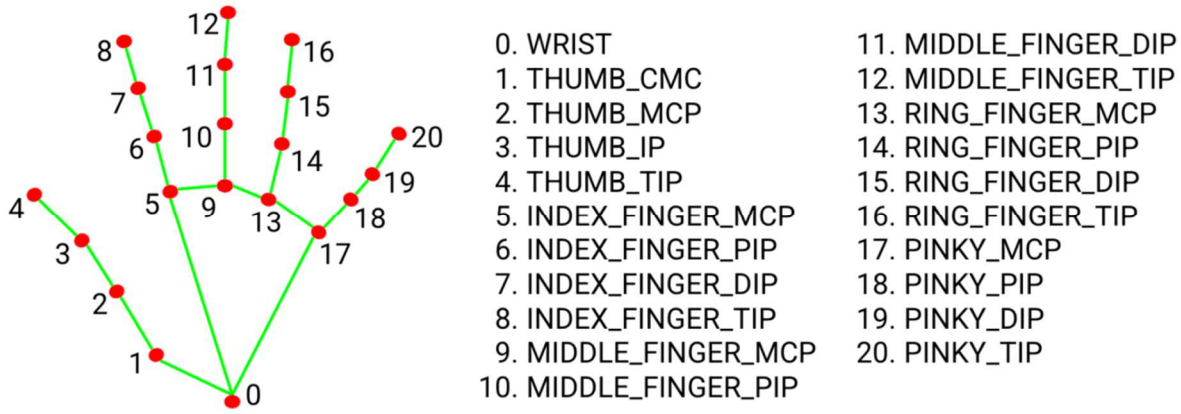


Figure 2. Mediapipe hand landmarks.

Figure 3. Feature extraction.

On all images and videos from the dataset, a hand was detected, a fragment containing the palm was cut out, and the coordinates of 21 landmarks were determined (figure 3). During preprocessing, the coordinates were normalized relative to the coordinates of the wrist point.

3.3. Neural network architecture

The recognition module is a neural network for multi-class classification. Each class corresponds to one of the signs of the dactyl alphabet. The network consists of six layers - three layers of bidirectional LSTM and three fully connected layers (figure 4). Through recursive links, LSTM layers can efficiently classify data that unfolds in time, which is essential for classifying dynamic gestures. The first, second and third layers contain 1024, 512 and 256 neurons, respectively, while fully-connected layers contain

128 and 64 neurons and have the relu activation function. The last fully-connected layer with softmax activation function contains 33 neurons.

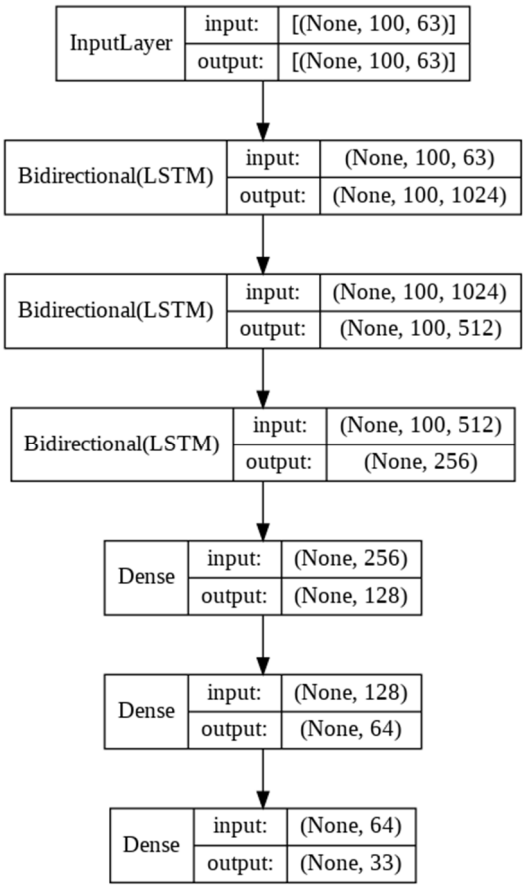


Figure 4. Neural network architecture.

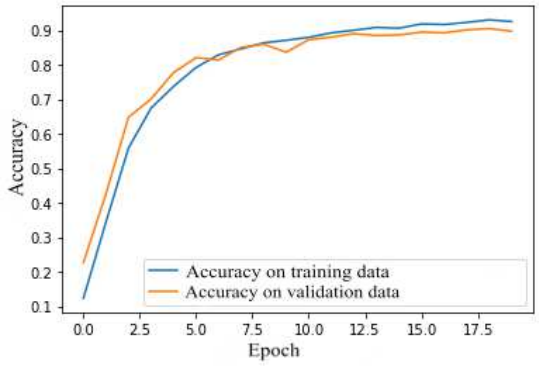


Figure 5. Graph of the values of accuracy on data during training.

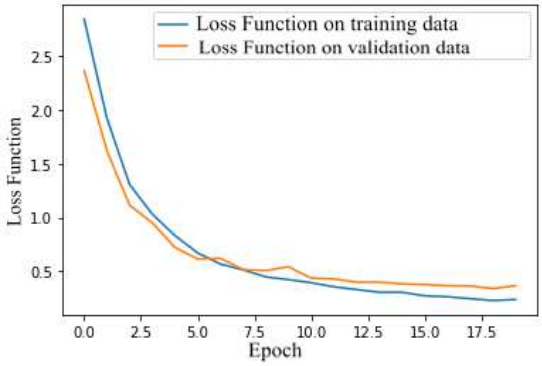


Figure 6. Graph of the values of loss functions on data during training.

4. Training

The size of the training sample was 7602 examples; the size of the validation sample was 1901 examples. The training lasted 20 epochs. To prevent overfitting, the Early Stopping method was used. The model accuracy reached 91% on the validation data (figure 5), the loss function value was 0.21 (figure 6).

5. Results and discussion

During the testing phase, the network showed an F-measure value with a micro-averaging of 91%. Testing was carried out on 2376 examples. Table 1 shows the results of recognition accuracy for 16 signs of the dactyl alphabet.

Table 1. Results

Sign	Precision	Recall	F1-measure	Support
а	0.72	0.82	0.77	51
б	0.82	0.80	0.81	70
в	1.00	0.98	0.99	61
г	1.00	0.97	0.99	111
д	0.50	0.29	0.36	7
е	0.88	0.83	0.85	110
ё	0.19	0.25	0.21	12
ж	0.89	0.98	0.93	100
з	0.29	0.71	0.42	7
и	0.83	0.91	0.87	111
й	0.33	0.43	0.38	7
к	0.61	0.69	0.65	16
л	0.82	0.89	0.86	90
м	0.94	0.91	0.92	107
н	0.95	0.95	0.95	132
о	0.96	0.95	0.96	115
overall	0.91	0.90	0.91	2376

The network showed the lowest recognition accuracy (less than 50% for at least one of the metrics used) for signs “д”, “ё”, “з”, “й”, “ц”, “щ”, “ъ” и “ь”. The low accuracy is probably caused by the heterogeneity of the data set and the insufficient number of examples for the data with dactyl, due to the fact that they are dynamic and can be fully represented only in the form of video recordings, which makes it difficult to collect data and store them.

For the rest of the signs, the network shows a high value of F1. As for the dactyls “в”, “г”, “н”, “о”, “т”, “у”, “ф”, “я” it exceeds 95%.

6. Conclusion

This paper considered the current situation in the development of sign language translation automation systems, described the existing methods and algorithms for recognizing sign languages. The analysis was carried out and it was revealed that at the moment there was no sufficiently accurate algorithm for recognizing Russian Sign Language gestures.

In this regard, an algorithm for the recognition of isolated dynamic and static signs of the dactyl alphabet of the Russian Sign Language was proposed. Our system showed good results on test data.

Acknowledgments
This paper was financially supported by the Russian Foundation for Basic Research (Grant No. 19-57-45006).

References

[1] Warchoł D, Kapuściński T and Wysocki M 2019 Recognition of fingerspelling sequences in Polish Sign Language using point clouds obtained from depth images *Sensors* **19(5)** p 1078
[2] Shi B, Rio D, Keane J, Brentari D, Shakhnarovich G and Livescu K 2019 Fingerspelling recognition in the wild with iterative visual attention *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)* pp 5399–408
[3] Aloysius N and Geetha MG 2020 A scale space model of weighted average CNN ensemble for ASL fingerspelling recognition *Int.l J. of Computational Science and Engineering* **22(1)** pp 154–61
[4] Lee CKM, Ng KKH, Chen C-H, Lau HCW, Chung SY and Tsoi T 2021 American sign language recognition and training method with recurrent neural network *Expert Systems with Applications* **167**
[5] Aparna C and Geetha MG 2020 CNN and stacked LSTM model for Indian Sign Language recognition *Machine Learning and Metaheuristics Algorithms, and Applications* pp 126–34
[6] Adithya V and Vinod PR, Gopalakrishnan U 2013 Artificial neural network based method for Indian Sign Language recognition *Proc. of the 2013 IEEE Conf. on Information and Communication Technologies (ICT)* pp 1080–5
[7] Luqman H, El-Alfy E-SM and BinMakhashen GM 2020 Joint space representation and recognition of sign language fingerspelling using Gabor filter and convolutional neural network. *Multimedia Tools and Applications* **80(7)** pp 10213–34
[8] Grif MG and Kugaevskikh AV 2020 Recognition of deaf gestures based on a bio-inspired neural network *Science Bulletin of the Novosibirsk State Technical University* **2-3(79)** pp 57–76
[9] Makarov I, Veldyaykin N, Chertkov M and Pokoev A 2019 American and Russian Sign Language dactyl recognition *Proc. of the 42nd Int. Conf. on Telecommunications and Signal Processing* pp 726–9
[10] Zhang F, Bazarevsky V, Vakunov A, Tkachenka A, Sung G, Chang C-L, et al 2020 MediaPipe Hands: on-device real-time hand tracking *Proc. of CVPR Workshop on Computer Vision for Augmented and Virtual Reality*