Sign language recognition through Leap Motion controller and input prediction algorithm

D G Enikeev¹, S A Mustafina¹

¹ Bashkir State University, 32, Zeki Velidi, Ufa 450076, Russian Federation
E-mail: cooldeepband@gmail.com

Abstract. The sign recognition systems are aimed to help deaf people communicate with society. In this paper we proposed our own concept of sign language recognition, which is based on a co-operative deep learning neural network, a text input prediction algorithm and a feedback from the user. We have pointed out the complexity of the Russian sign language and conceived the fingerspelling recognition. The method utilizes the natural properties of fingerspelling in order to increase the accuracy and recognition performance by predicting the ongoing letter. We also provide a detailed review of data acquisition in the related works. From a hardware perspective, we suggest using Leap Motion controller.

1. Introduction
According to the WHO, about 5% of population, or 466 million people have a disabling hearing loss. One of the significant impacts on a deaf person’s life is a difference in social communication. The sign language is a major tool for sharing thoughts, ideas, feelings in deaf community. It is made by special hand movements, postures and facial expressions. Due to its complexity and high specialty, the sign language is hard to understand, which erects a communication barrier for deaf people. This affects their mental health, education and economic area of life. The government answer to the problem is a human sign language interpreter, though the number of these specialists does not cover the demand. As an example, there are only 1138 interpreters handling over 300 thousand speech and hearing disabled people. Compared to a spoken language, the sign language is much more difficult to learn. For many deaf children who need to learn the sign language, the lack of speech practice that accompanies hearing children from their birth (even not yet able to speak, they already learn to understand speech), leads to a slowdown in the development of the verbal-logical level of thinking. Therefore, the education system should provide a variety of simple and yet efficient methods of learning the sign language.

For people whose hearing is impaired, technological support is a must. The advancement in visual, multimedia and machine learning technology makes this problem feasible. The use of modern visual technology overcomes the above listed difficulties. The human – computer interaction must provide an interface between the hearing and deaf society, narrowing the communication gap. One of the approaches is sign language recognition (SLR). The SLR systems are purposed to capture the gestures and other features of the sign language and using computations translate them into text, audible speech or executable commands.

Also, the SLR could be used as an interactive learning method. A system with a high recognition rate serves as a benchmark of how precisely the sign was performed. A learning
person will have a 3d animation of a gesture, and afterwards the system will guide him through
the movements and amend them if they did not fit into the correct movement track and hand
postures.

The sign language in Russia contains two autonomous systems: the Russian Sign Language
(RSL) and the Russian Manually Coded Sign Language (RMCSL). The RSL is used in everyday
communication and has its own vocabulary and grammar. The RMCSL is used in a formal
setting, in which gestures accompany the speaker’s oral speech. The gestures in the RMCSL
act as word equivalents, and they come in the same order as the words in the corresponding
sentence. It is a secondary sign system, which is assimilated on the basis of speech learning by
a deaf child. A standalone type of the sign language is fingerspelling, where every letter has an
assigned fingers static posture or a dynamic gesture. Using a set of fingerprints, the speaker
follows the grammar of the verbal language. Thus, fingerspelling is a kind of kinetic form of the
verbal speech, communication. It is commonly used to transmit words for which there is no sign
equivalent, such as proper nouns.

2. Data Acquisition
The SLR systems utilize hardware data collection and software recognition algorithms. This
article focuses on the hardware part. From the hardware perspective, the SLR systems divide
into two major implementations: visual-based and the sensor gloves.

2.1. Sensor-based approach
One of the first approaches of the SLR is based on the use of a pair of sensor gloves.
Their structure is a combination of the instrumented gloves and an array of sensors like an
accelerometer, flex and proximity sensors. The main advantage of the sensory method is the
direct output of the necessary hand data (bend angle of the fingers, palm position, etc.) in digital
values, which eliminates converting raw data into the characteristics. The direct measurements
exclude any environmental impact on the acquisition process. The basic complexity of the glove-
based SLR system is described by degrees of freedom. It depends on the number of sensors
and their spacing. [1] uses 5DT data glove providing 14 data channels of fingers flexion and
abduction between them. Adding the accelerometers [2] gains movement and spatial information.
The Surface electromyographic (sEMG) sensors measure the muscle activity and are able to
distinguish different hand postures. [3] combined the sEMG and accelerometers to provide
two-component picture of kinematic information and hand shapes.

2.2. Visual-based approach
The visual method of recognizing the sign language has a dominant role in the researches. The
visual SLR systems are the most intuitive as they minimize the period of the user adaptation
to the interface. At the same time, the developers face problems with computer vision. In the
visual approach, gesture information is presented as a set of image frames. The input to the
system is provided using video devices. Depending on their design, they can be divided into
three groups:

- single - cameras of smartphones, video and web cameras;
- stereo and multicamera systems - an assembly of monocular cameras;
- specialized IR cameras.

Most visual SLR systems perform three common operations:

- **Extraction of key frames.** A video camera is capable of capturing frames in a range
  between 15 and 120 frames per second. The sign language users’ average speed of the
spelling per minute is limited by the finite hand movement speed, necessity to lock it for a moment and time of the transient process between the signs. [4] shows the spread of signing speed between 50 to 120 signs per minute, [5] sets the comfortable signing rate of an adult to 300 letters per minute (or 60 words). The overhead between the signing and camera speed is computationally expensive and leads to a significant increase of the recognition delay. Therefore, it is crucial to process the most informative frames.

- **Preprocessing.** During the preprocessing stage the key image frames must be filtered in order to curate useful information and reduce the distracting noise, a static background and unwanted objects.

- **Feature extraction.** After that, processed images are fed into the algorithms which detect fingers, palm, face and their configuration, orientation, etc. The stage converts the image into an array of data characterizing the input sign.

The overwhelming majority of investigations use a single still digital camera. They study isolated sign recognition, where all signs are performed separately. [6] uses a single web camera as the most common visual acquisition device. [7] explores an FPGA unit in the SLR system with a single camera. In most cases, the capturing process is made on a uniform background to simplify the preprocessing stage. The multiple camera approaches are aimed to capture the three-dimensionality of signs, like depth cameras [8], and stereo pair [9]. [10] utilizes three cameras for front facing, side hand movement capturing and also for the face expressions of the signer. The development of the novel human-machine interactions brought specialized recognition systems like Microsoft Kinect and Leap Motion. The devices’ purpose is to give the user an opportunity to interact with virtual objects using his body. The Kinect is a 3D IR camera sensor which can use the depth information of a person to capture his skeletal data and movement. [11] presents a leverage of a video and a Kinect skeletal data SLR. Leap motion controller (LMC) is a sensor for hand and finger movements in a 3D space. It reports fingers and palm positions, orientation and speed. [12] uses a single LMC in American sign language letters and digits recognition.

### 3. Proposed work

The concept of the RSL recognition system proposed relies on three basic ideas: the Leap motion controller as a hardware data acquisition method, fingerspelling and further RMCSL aimed recognition, multiplexing a deep learning neural net and a predictive text input algorithm.

#### 3.1. Leap Motion controller data acquisition

The Leap Motion Controller (LMC) is a small sensor designed to track hands in real time and building their models in API. The controller transfers the captured hands into a 3D space, replacing them by skeletal representations. It is a compact rectangle (dimensions), with a capturing window with a USB interface. The stereoscopic pair of lenses captures the working area illuminated by the three IR range LEDs. The data is represented by a set of left and right cameras images in a grayscale format and transferred to the PC software driver. The Leap Motion Orion software kit rebuilds the 3D scene and performs preprocessing. The internal algorithms detect hands and small pointer-like objects. They represent the captured scene into the Euclidian coordinate system with the center in the device itself. User’s hands are represented as a digital model – the skeletal representation brought by Leap Motion. It has all physical dimensions, position, rotation and velocity. The LMC key benefits are follows:

- **High output rate and accuracy.** The LMC’s sample rate depends on many criteria, such as the USB interface workload, PC computing power and environmental conditions. Internal benchmarking tests have showed that the camera framerate imposes the average 89 frames
per second [13]. It allows one to make smooth samples and detect vast movements. Several investigations were made to measure the controller’s accuracy. [14] conducted experiments for measuring the accuracy of determining the coordinates of a robotic arm in space. In statics, the device can provide an average accuracy of 0.7 mm, while during the dynamic pattern movement it decreases to 2.5 mm. In a real context [15], the LMC had the static measurements with a human in a role of a subject. The results showed that the tracking accuracy was within 4–5 mm, which is comparable with a human hand tumor.

- **Environment independency.** A major issue in the SLR video systems is hand segment extraction. It depends on the color of the background, the user’s skin and clothing. Due to invisible IR range, the software can easily compensate background and untrackable objects. The dependencies of the environmental lighting conditions are relieved because of the LED brightness adjustments. The user’s hands could be easily extracted from the background. On the other hand, any IR resources, such as daylight and domestic heaters can expose the whole scene, which degrades accuracy. The second unsuitable situation is complete darkness. The LMC uses the USB 2.0 standard with a limit of consumption of 1 A, thus the cameras resolution suffers from insufficient light conditions.

- **Extensive dataset.** The LMC API brings a large dataset, consisting of skeletal data, positions, velocity and rotation. The captured hand fits in an internal hand model. The hand model provides information on the type of the hand (left or right), coordinates of the center of the palm and speed in mm/s. If part of the arm is outside the recognition zone, the program builds a continuation from the anatomical dependencies. A probabilistic assessment of hand tracking displays how accurately the observed data match the internal model of the hand. Characteristics of fingers include the direction of fingers in the form of vectors, length of the finger, position of the tip and its speed, coordinates of the joints of the distal, middle, proximal and metacarpal phalanges. The internal recognition system provides a gesture features traction like a swipe and a tap. The geometrical representation of the sign allows one to process it through the statistical analysis, highlighting the particular qualities. The LMC API also provide a raw input stream from the IR cameras for the system adjustment or the utilization of the visual approaches.

### 3.2. Fingerspelling recognition

Dialogues in the RSL usually have a form of an exchange of fragmentary messages, with the first statement as the initiator of the conversation setting the topic. The subsequent statements of the participants in the dialogue build up and clarify the initial idea. Like in ordinary everyday dialogs, verbs are often omitted in the expressions in the RLS, the whole phrase is significantly reduced, sometimes acquiring a slightly recognizable form compared to its counterpart in refined literary language. The RSL statements have a very pronounced consistency: in the sign languages there are no special gestures to indicate “what is always present in a conversation”, for example, a head, arm, nose, etc. [16]. These values are expressed by indicating, for example, your hand. Thus, one of the most important tasks of the RSL recognition systems is the generation of a coherent text based on keywords and expressions obtained from the “interlinear translation” of a sign utterance. Given the polysemy of gestures, it is necessary to carefully monitor the correspondence of the generated text to the current subject area. Proper RSL translation is a very difficult task, and we consider it excessive.

Real-time speech translation usually requires the use of continuous gesture recognition [17]; video stream should be processed with the least delay. This task is more computationally loaded and realistic than the determination of individual features. The quality of recognition in real time is affected by the effects of epenthesis (insertion of additional features in characters), articulation of words, the influence of the word order on the meaning of phrases, use of slang, uneven speed, etc. We suggest using fingerspelling recognition as the simplest yet effective interface solution.
Fingerspelling is a finite language system, which could be used as a communication method. It is simple, predominantly based on static signs and easy to perform. The database of fingerspelling is easy and able to include different samples as compared with the rich RSL dictionary. We highlight that the main purpose of the SLR systems is providing an applicable communication interface. The RMSCL is the next step of designing such systems, because it follows the Russian language linguistic rules.

3.3. Deep learning network and predictive text input methodology

The proposed system is represented at Fig. 1.

It consists of several blocks:

1. A user is faced with the data acquisition device – LMC. It provides data stream through the API.
2. A multiplexing stream software. It analyzes the data stream for the sign and control gestures and passes them to the corresponding block.
3. A deep learning neural network is used for fingerspelling recognition purposes. The network has an adjustable list of output candidates.
4. A control algorithm lets the user amend the recognition process.
5. An input text prediction block forms a short list of subsequent letters and the whole word.
6. A display for an output.

The overall proposed recognition system consists of two interconnected blocks: deep learning neural network and text prediction algorithm. The input data from LMC API are fed into the N-dimensional layer of the neural network. The first letter goes through the prediction of all alphabet signs. The output of the system curates at the display. The output of the neural network goes into the input of the prediction algorithm. We address fingerspelling nature – it submits the linguistic rules and uses the words of the written language. The prediction algorithms are common for smartphone keyboards and have high prediction efficiency. [18] showed efforts reduction up to 70 percent. The presumptive profit of prediction usage is as follows:

- The prediction algorithm submits a list of words’ (and the following letters’) probability. It composes a list of suggested signs to the recognition block. The redundancy of comparable units accelerates the recognition process. In questionable output situations, the network could rely on the prediction result.
• Increasing the speed of bringing deaf person’s thoughts. The user has an opportunity to end the word earlier, because the prediction block has showed him his intended word. Potentially it could reduce the efforts of signing and increase the system efficiency by bypassing the other fingerspelling signs.

The system gives user a permission to correct recognition output, thereby building a feedback correction. Before signing, a user is given an instruction how to correct the SLR. We suggest using the common gesture actions (Fig. 2) – swipe up to confirm the predicted word, down for the system correction. In the correction menu the user can swipe left or right and choose between the alphabet in the order of the probabilistic output of the network. The manual tweaking of the system highlights the weak spots and gives feedback to develop a better approach.

Conclusion
The paper describes the importance of the SLR systems as an urgent interaction tool for deaf people socialization in the hearing community. We have studied the Russian sign language properties and translation issues. Fingerspelling is a trivial and slow method of communication, but it provides a simpler linguistic. The paper highlights that the main purpose of the SLR system is to give an operating interpreter the needs of a corresponding deaf community, rather than the inclusive SLR translator. The proposed system has a novel combination of the neural network and input prediction. It gives a user an alternative choice to correct the output. Also, we have reviewed the state of art data acquisition methods. The LMC is a compact IR sensor for finger tracking. It overcomes most video cameras weaknesses like environmental dependencies and feature extraction routine. Further study will focus on implementing system presented and benchmarking the overall accuracy.

Acknowledgments
The research was carried out within the state assignment of Ministry of Science and Higher Education of the Russian Federation (theme No. FZWU-2020-0027).

References
[1] Ramakant, Shaik N and Veerapalli L 2015 IEEE International Advance Computing Conference (Banglore) pp 639–43
[2] Vijayalakshmi P and Aarthi M 2016 International Conference on Recent Trends in Information Technology (New Jersey) p 1–6
[3] Li Y, Chen X, Zhang X, Wang K and Wang Z J 2012 IEEE Transactions on Biomedical Engineering 59 (10) 2695–704
[4] Vanjikumaran S and Balachandran G 2011 ICTer (Colombo) pp 39–44
[5] Grosjean F 1977 Perception & Psychophysics 22 (4) 408–13
[6] Vedak O, Zavre P, Todkar A and Patil M 2019 IRJET 6 (4) 1907
[7] Vargas L P, Leiner B, Torres C O and Mattos L 2011 J. Phys.: Conf. Ser. 274 012051
[8] Schlesinger I M and Namir L 1978 Sign language of the deaf: psychological, linguistic, and sociological perspectives (New York: Academic Press) p 339
[9] Uebersax D, Gall J, Van den Bergh M and Van Gool L 2011 IEEE ICCV Workshops (Barcelona) pp 383–90
[10] Campbell, Becker D, Azarbayejani A, Bobick A and Pentland A 1996 In Second Intl. Conf. on Face and Gesture Recogn. (Killington) pp 157–62
[11] Tsechpenakis G, Metaxas D and Neidle C 2006 Computer Vision and Image Understanding 104 (2-3) 140–56
[12] Chong T-W and Lee B-G 2018 Sensors Sensors (Basel) 18 (10) 3554
[13] Niechwiej-Szvedo E, Gonzalez D, Nouredanesh M and Tung J 2018 PLoS One 13 (3) e0193639
[14] Weichert F, Bachmann D, Rudak B and Fisseler D 2013 Sensors (Basel) 13 (5) 6380–93
[15] Valentini P P and Pozzuti E 2017 IJIDEpM 11 641–50
[16] Zayceva G L 1991 Dactylogy. Gesture speech: Textbook for high schools (Moscow: Proslyascheniye) p 12
[17] Forster J, Oberdörfer C, Koller O and Ney H 2013 *Pattern Recognition and Image Analysis* **7887** 89–99

[18] Rădescu R and Pupezescu V 2016 *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (Ploiesti)* pp 1–8