UDC 004.932

DOI: 10.15587/1729-4061.2022.262509

EFFECTIVENESS OF THE USE OF ALGORITHMS AND METHODS OF ARTIFICIAL TECHNOLOGIES FOR SIGN LANGUAGE RECOGNITION FOR PEOPLE WITH DISABILITIES

Aigulim Bayegizova
Candidate of Physical and Mathematical Sciences,
Assistant Professor
Department of Radio Engineering,
Electronics and Telecommunications***

Gulden Murzabekova
*Corresponding autho*r
Candidate of Physical and Mathematical Sciences,
Associate Professor
Department of Computer Sciences**
E-mail: g.murzabekova@kazatu.kz

Aisulu Ismailova
PhD, Associate Professor*

Ulzada Aitimova
Candidate of Physical and Mathematical Sciences,
acting Associate Professor*

Ayagoz Mukhanova
PhD, Associate Professor
Department of Information Systems***

Zhanar Beldeubayeva
PhD, Senior Lecturer*

Aliya Ainagulova
Candidate of Technical Sciences, Senior Lecturer*

Akgul Naizagarayeva
Master of Engineering*

*Department of Information Systems**
**S. Seifullin Kazakh Agrotechnical University
Zhenis ave., 62, Nur-Sultan, Republic of Kazakhstan, 010011
***L. N. Gumilyov Eurasian National University
Satpayev str., 2, Nur-Sultan, Republic of Kazakhstan, 010008

According to WHO, the number of people with disabilities in the world has exceeded 1 billion. At the same time, 80 percent of all people with disabilities live in developing countries. In this regard, the demand for the use of applications for people with disabilities is growing every day. The paper deals with neural network methods like MediaPipe Holistic and the LSTM module for determining the sign language of people with disabilities. MediaPipe has demonstrated unprecedented low latency and high tracking accuracy in real-world scenarios thanks to built-in monitoring solutions. Therefore, MediaPipe Holistic was used in this work, which combines pose, hand, and face control with detailed levels.

The main purpose of this paper is to show the effectiveness of the HAR algorithm for recognizing human actions, based on the architecture of in-depth learning for classifying actions into seven different classes.

The main problem of this paper is the high level of recognition of the sign language of people with disabilities when implementing their work in cross-platform applications, web applications, social networks that facilitate the daily life of people with disabilities and interact with society. To solve this problem, an algorithm was used that combines the architecture of a convolutional neural network (CNN) and long short-term memory (LSTM) to study spatial and temporal capabilities from three-dimensional skeletal data taken only from a Microsoft Kinect camera. This combination allows you to take advantage of LSTM when modeling temporal data and CNN when modeling spatial data.

The results obtained based on calculations carried out by adding a new layer to the existing model showed higher accuracy than calculations carried out on the existing model

Keywords: neural network model, convolutional neural network, LSTM module, convolution, sign language

Received date 24.05.2022
Accepted date 24.07.2022
Published date 31.08.2022

***How to Cite**: Bayegizova, A., Murzabekova, G., Ismailova, A., Aitimova, U., Mukhanova, A., Beldeubayeva, Z., Ainagulova, A., Naizagarayeva, A. (2022). Effectiveness of the use of algorithms and methods of artificial technologies for sign language recognition for people with disabilities. Eastern-European Journal of Enterprise Technologies, 4 (2 (118)), 25–31. doi: https://doi.org/10.15587/1729-4061.2022.262509*

1. Introduction

In the modern world, it is relevant to use the capabilities of artificial intelligence in simplifying the daily life of people with disabilities and ensuring their place in society. The difficulty of detecting sign language using machine learning is partly due to the complex nature of sign language: in addition to spoken language, the transmitted information is divided into three separate channels, namely hands, face and upper body, so that only all their information contains all the information. In this construction, hands tend to convey words themselves, while the face and upper body tend to convey grammatical and temporal information. These exceptions require special processing of all three channels, since each channel differs from the others in detail, scale and range of motion. Only with the help of efficient preprocessing algorithms and computer vision can the basic model be sufficiently simplified to provide real-time viability, appropriate accuracy and simple configuration of sign language detection equipment.

Understanding human behavior is necessary for effective interaction between artificial systems and people in the real world. This can be achieved by translating perceived behavioral signals and context descriptors into encoded behavior. However, due to the complex nature of human activity, it is still difficult. This applies to many factors, such as variability of activity classes, background noise, and similarities between different activity classes. Even though many studies on human behavior recognition (HAR) have been published, these tasks do not have human-like results, and this ability is still an open question. The key to the success of this task is to obtain distinctive spatial and temporal features for effective modeling of the spatiotemporal evolution of various actions. It is important to consider and incorporate low-dimensional, high-precision embedded methods using new technologies into existing models [1]. Therefore, it is relevant to recognize the sign language of people with disabilities and facilitate its use in society, social networks and applications using new technologies.

2. Literature review and problem statement

Object tracking algorithms [2] have a set of significant advantages, but the disadvantage is high computational complexity and the need to comply with certain conditions. It is advisable to combine these algorithms with detection methods to correctly track objects in a video sequence.

In [3], 8 patterns of attention were used in the spatial and temporal domains. When using data from multiple time series, as in the case of HAR, traditional CNN cannot be used directly because capturing the spatial correlation characteristics of the data is not enough to create an accurate recognition model. Therefore, skeletal-based HAR requires both temporal and spatial information.

Despite the successes achieved in this area, there are also a number of shortcomings [4] that limit the use of these technologies both in the mass segment and in highly specialized ones, for example, in military equipment.

During the analysis of object detection algorithms [5], it was found that they are easy to implement, but most of them have low efficiency and are strongly influenced by external factors, such as lighting, background, object size, etc. To improve the efficiency of these algorithms, it is necessary to combine them, which will improve the accuracy of object localization in the image.

In [6], the authors describe deep neural networks (DNNs). Deep neural networks have recently made great strides in various learning tasks and have also been used to classify environmental sounds. Although DNNs show their potential in the classification problem, they cannot make full use of temporal information. In this paper, the authors propose a neural network architecture for using sequential information.

In the work [7], a depth map was calculated by matching the left and right images with the SAD (Sum of Absolute Differences) algorithm. The Theo Pavlidis Algorithm, which visits only the boundary pixels, was used to find the contours. This method brings down computational costs.

The Support Vector Machine (SVM) classifier was used in [8]. The authors deviate from other traditional methods by not using hand markers such as gloves for gesture recognition.

The peculiarity of this work is to demonstrate the effectiveness of the developed algorithm for studying spatial and temporal possibilities using three-dimensional skeletal data taken only from the Microsoft Kinect camera. To test the proposed algorithm, some tools are proposed and discussed that ensure its effectiveness for continuous recognition of human actions in real time.

3. The aim and objectives of the study

The aim of the study is to improve the efficiency of gesture recognition systems based on neural network algorithms for real-time image analysis.

To achieve this aim, the following objectives were accomplished:

– to analyze the existing sign language recognition algorithm;

– to implement sign language recognition for people with disabilities using neural network technologies.

4. Materials and methods

MP Holistic is a real-time human activity monitoring technology. Simultaneous real-time perception of human posture, facial expressions and hand control on mobile devices, fitness and sports analysis, motion control and sign language recognition, augmented reality effects, etc. Impressive applications can be added, such as MediaPipe, an open-source database source code, designed specifically for complex receive channels using accelerated inputs (such as GPU or CPU), providing fast and accurate, but separate solutions to these problems. Combining them all into a final solution that is semantically relevant in real time is a particularly challenging task, requiring multiple dependent neural networks to run simultaneously.

The HAR framework includes two modules: a standalone process that is generated only by a skeletal 3D stream captured from an RGBD10 camera, and a model that is studied using a combined CNN and LSTM architecture to explore spatial and temporal correlation characteristics. The second module refers to an online process that is used to infer the action taken by the user in real time, and then the prediction uncertainty is calculated to decide on the reliability of the recognized action (Fig. 1).

Long short-term memory networks (LSTMs) are a type of recurrent neural network (RNN) that can learn and remember long-term dependencies. Long-term recall of past information is the default action. Execution algorithm for LSTM modules:

– forget small parts of the previous state;

– select and update cell state values;

– display certain parts of the cell state [9].

Fig. 2 shows how the LSTM works.

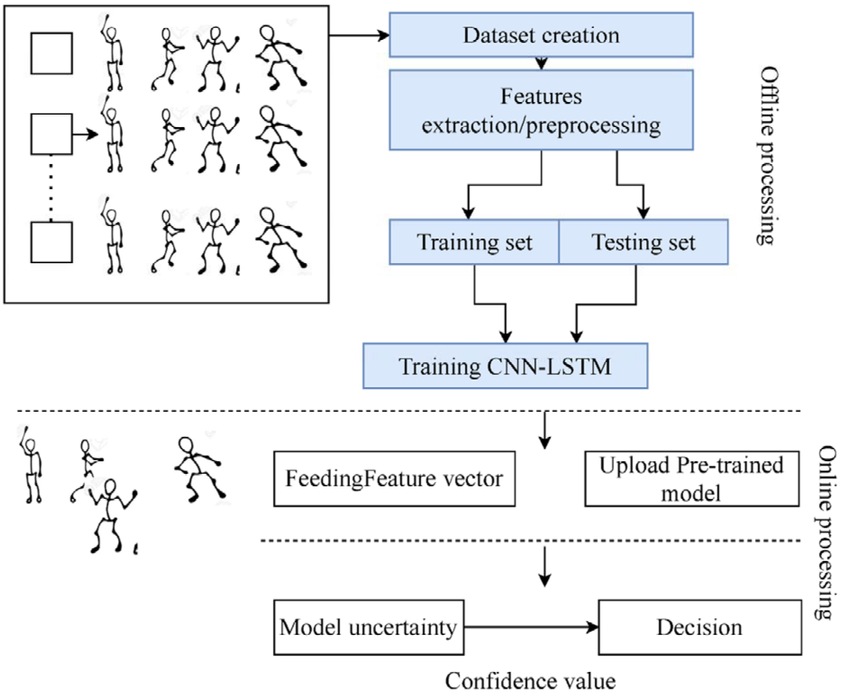


Fig. 1. HAR structure

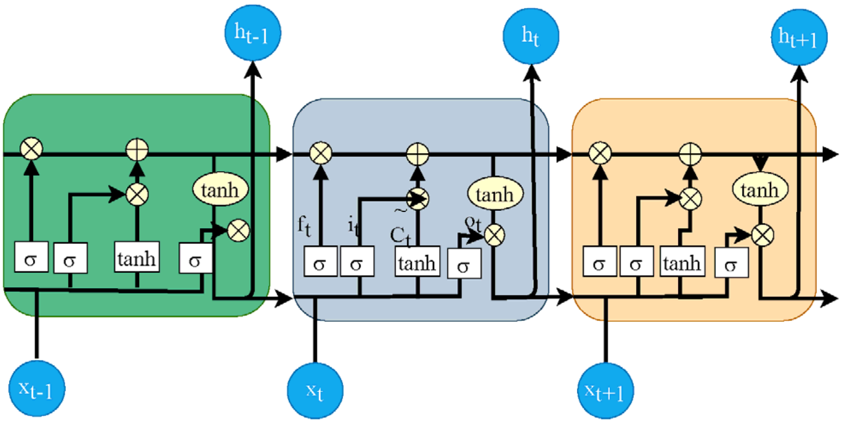


Fig. 2. Model diagram of the LSTM

Recurrent neural networks (RNN) have connections that form directed loops that allow input from the LSTM output into the current phase. The output of the LSTM becomes the input of the current phase and can remember previous inputs thanks to its internal memory. RNN is commonly used for image recording, time series analysis, natural language processing, handwriting recognition, and machine translation [10]. RNN neural network execution algorithm:

– the output at time *t*–1 is transferred to the input at time *t*;

– similarly, the output at time *t* is transferred to the input at time *t*+1;

– RNN can process the input of any length;

– the calculation takes into account historical information, and the sample size does not increase with the size of the input data [10].

An open RNN looks like Fig. 3.

Like RNNs, LSTMs have repetitive connections, so the neuron's activation state from the previous time step is used as the context to generate the output. But unlike other RNNs, LSTM has a unique formula that avoids other RNN training and scaling issues. As well as the impressive results that can be achieved with the task of recognizing sign languages are the reasons for the popularity of this technology.

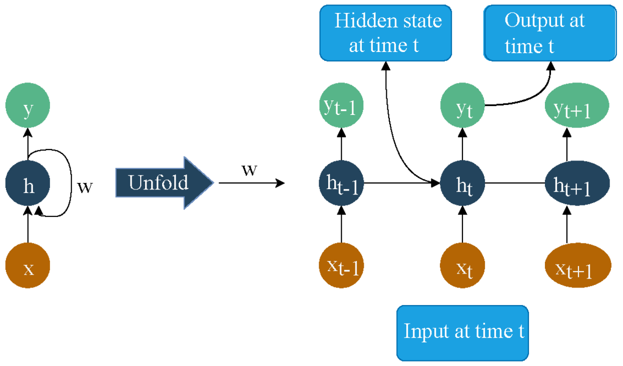


Fig. 3. Scheme of RNN operation

5. Results of the application of neural network technologies for recognition of the sign language of people with disabilities

5. 1. Analysis of the existing sign language recognition algorithm

Using the LSTM module, a feature vector was constructed for each skeletal joint, including 3D Cartesian positions and three Euler rotation angles in 3D space (Fig. 4).

Fig. 5 shows the shape of the feature vector. Let S be the number of data in the model, S=(3+3)×Njoint×Nframe, where Njoint is the number of links used, equal to 15. Nframe is the number of frames captured by the action, about 100. Therefore, according to the model, the feature S=6*15*30=2,700.

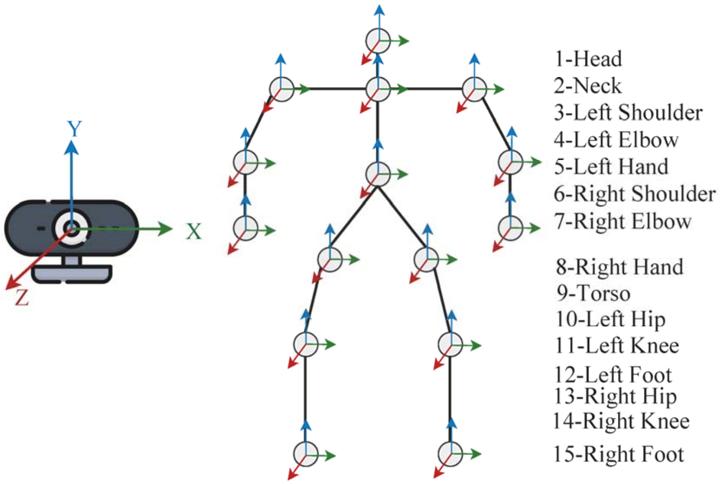


Fig. 4. Marking the points of the joints of the used human skeleton

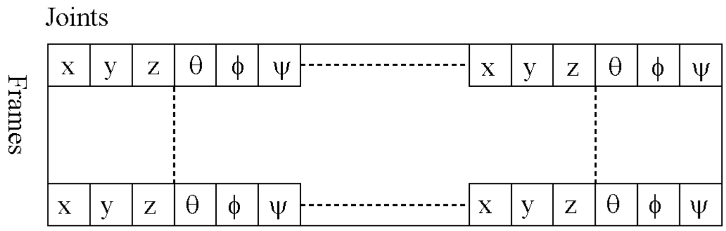


Fig. 5. Feature vector

Function preprocessing consists of the following steps:

– normalization – the process of entering ordinary or more familiar data into a range. The main benefit of recovery is that it eliminates intra-class discrepancies between data when the same action is performed by different people to match the height, limb length, orientation, and location of different users. Officially, it converts data from a range (low, high) to a new range (new low, new high);

– remodeling – data from the sensor can be obtained at irregular intervals or in different sizes (20, 40, 50, 60 frames). To properly use time-series data, it must be within a certain interval size. So some interpolation/extrapolation functions are used to resize the input image to the same number of frames per pattern. Fig. 6 shows some extrapolation methods where a linear model is used;

– the translation provides the same origin of the system for all received frames, for all used skeletal joints, a link was set on the sensor to view dynamic actions (Fig. 7);

– symmetry is necessary for actions such as "Good afternoon", "Hello", "How are you", etc. The idea is to consider a new model based on a rotated version of the original 3D skeletal data, such as human right and left-hand movements (shown in Fig. 8).

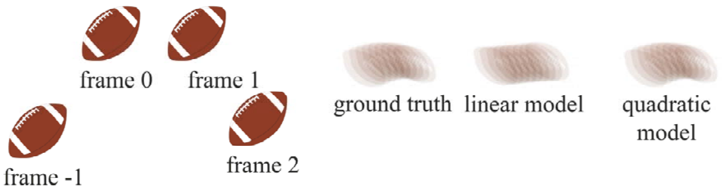


Fig. 6. Extrapolation methods

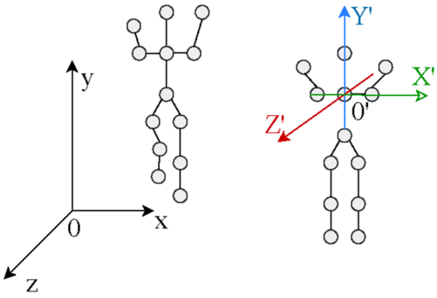


Fig. 7. Translation preprocessing

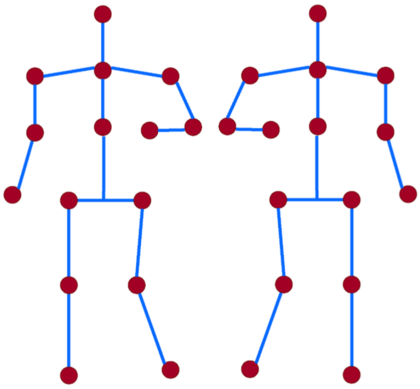


Fig. 8. Symmetry preprocessing

The difficulty in detecting sign language with machine learning is partly due to the complex nature of sign language: in addition to spoken language, the information transmitted is divided into three separate channels, namely hands, face, and upper body, so that only all of their information contains all information. In this design, hands tend to convey words themselves, while the face and upper body tend to convey grammatical and temporal information. These exceptions require special handling of all three channels, as each channel differs from the others in detail, scale, and range of motion.

5. 2. Implementation of sign language recognition for people with disabilities using a neural network technology

The proposed image classification model is a multilayer deep neural network consisting of long short-term memory (LSTM) and a convolutional neural network (CNN).

LSTM is a type of closed-loop RNN as opposed to traditional feed-forward neural networks. This feature of closed-loop connections makes the LSTM a kind of "general-purpose computer" that can compute anything that a Turing machine can do.

Fig. 9 shows the architecture of our proposed LSTM-CNN model. The images are fed through the input layer. This input layer is then passed to the batch normalization layer. The batch normalization layer applies a transformation to the previous layer that keeps the standard deviation of the activation function close to 1 and the activation mean close to 0, thereby normalizing it. We apply per-feature normalization so that each input feature map is normalized separately. The axis argument specifies the axis on which the normalization is performed. Statistics were applied to each batch to normalize the training data and used the average values calculated over the training period during testing. The output shape of the batch normalization layer is the same as the input shape, making it unsuitable for LSTM cells. You can use the reshape layer before the LSTM layer to reshape to the desired size. After the input layer is resized, it passes through the LSTM cell. Tanh, i. e., hyperbolic tangent, is used as the activation function of the LSTM cell. The LSTM cell has a dropout factor that helps avoid data overfitting. With these characteristics, the LSTM can remember the long dependency and shape of the input image for a particular model. The output of the LSTM layer is fed directly to the convolutional layer. A convolutional kernel is generated by a convolutional layer that produces a tensor of outputs by convolution with the input of the layer in one spatial (temporal) dimension. The convolutional layer extracts locally significant features. The Adjusted Linear Unit (ReLU) was used as the activation function in this convolutional layer. To prevent over-connectivity due to "fully connected" neurons in a CNN, a capture layer can be used after the convolutional layer.

The MH pipeline combines separate models for the pose, face, and hand components, each optimized for its specific area. However, due to their different specializations, inclusion in one component is not appropriate for others. The pose estimation template, for example, takes a low fixed resolution (256×256) video frame as input. However, if you need to cut out areas of the arms and face from this image to switch to the appropriate templates, the resolution of the image will be too low for accurate articulation. Therefore, MH is implemented as a multi-stage pipeline that processes different areas using image resolution depending on the region. MH first estimates the person's pose with the BlazePose pose detector and then with the base model.

To implement the program, MediaPipe Holistic was used, which tracks the human body as an efficient pre-processing step for the sign language detection pipeline. At the input, data from an RGB webcam with a resolution of SD 640×360 pixels was used. The result is 640×360=230,400 pixels per frame. Since each pixel carries three times as much information, each frame contains 691,200 integers of information. Shooting at 30 frames per second results in pop-ups with 20,736,000 units of information per second. For the effectiveness of the method, the following steps of action are performed:

– the frame is pre-processed using MediaPipe Holistic, it will be obtained from these 691,200 values;

– 33 position markers;

– 21 signatures on each hand;

– 468 bookmarks. There are 543 landmarks in total. With 3D coordinates for each orientation, this means a total of 543×3=1,629 pop-up information values in one frame, i. e. 48,870 per second;

– the source data is processed through the connected library during operation. By converting the result obtained by MH into array values, we can achieve a certain result;

– a special Numpy library is used to read them evenly;

– an empty array of poses is built in a common line of code;

– the base values obtained in the test variable through the loop are added;

– it is added to the array of pos of the set of points res.x, rus.y and res.z according to their symbols. The following words were chosen as initial actions: ['Thank you', 'How are you?', 'What time is it?'] (Fig. 10–12);

– video recording data is recorded by entering each selected word into the folder file. To write a word or a sentence, the number of videos is 30, which means that the action is performed from every angle, i.e., top, bottom, horizontal and so on.

Fig. 13 shows the confusion, accuracy, and memory performance matrix according to the action table for the test data, which reaches 98.30 % accuracy. As shown in the matrix, most activities are classified correctly; especially the best results are obtained by the "Stop" action due to the large number of joints involved in this action.

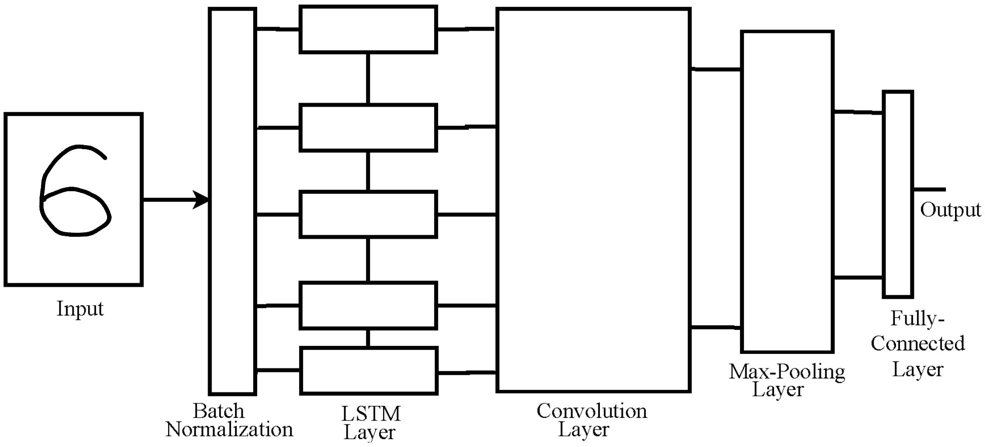


Fig. 9. LSTM-CNN model architecture

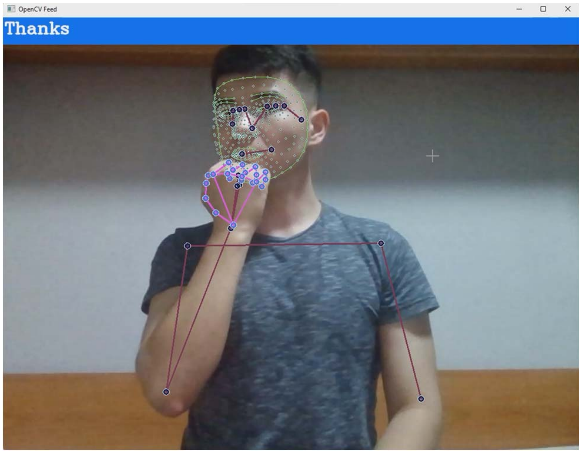


Fig. 10. Definition of "thank you"

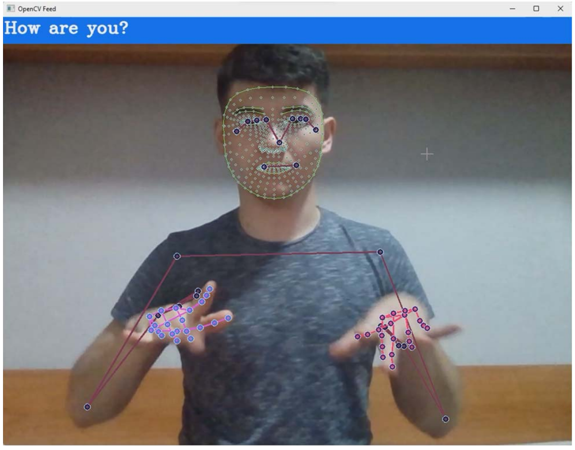


Fig. 11. Definition of "how are you?"

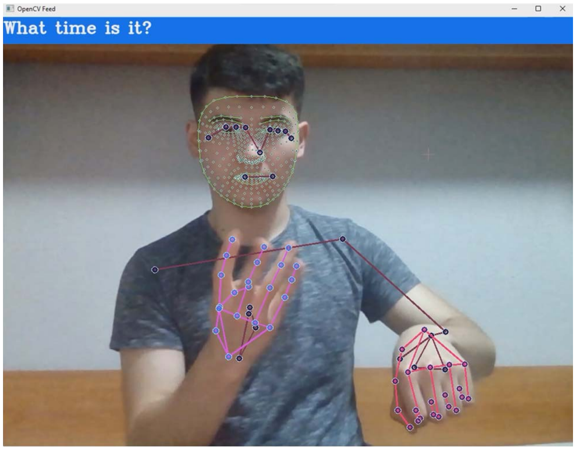
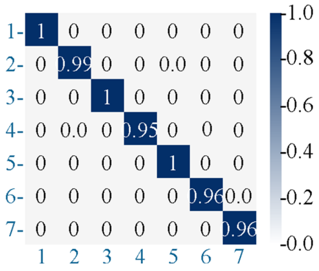


Fig. 12. Definition of "what time is it?"

Test Data	Precision	Recall
1 STOP	1.0	1.0
2 CALL	0.976	0.988
3 HELLO	0.977	1.0
4 COMING	1.0	0.952
5 GOING	0.97	1.0
6 POINTING	1.0	0.96
7 OTHER	0.957	0.957



	1	2	3	4	5	6	7
1-	1	0	0	0	0	0	0
2-	0	0.99	0	0	0.0	0	0
3-	0	0	1	0	0	0	0
4-	0	0.0	0	0.95	0	0	0
5-	0	0	0	0	1	0	0
6-	0	0	0	0	0	0.96	0.0
7-	0	0	0	0	0	0	0.96

Fig. 13. Model performance for test data

MH is a solution to this problem, offering a new posture topology for the modern human, opening up new areas of application. MH consists of a new pipeline with the optimized pose, face, and hand components, each running in real time, with minimal memory transfer between their end servers, and support for exchanging three components based on quality/speed matching.

6. Discussion of the results of using algorithms and methods of artificial technologies for the study of sign language recognition

When using data from multiple time series, as in the case with HAR, a traditional CNN cannot be used directly because capturing the spatial correlation characteristics of the data is not enough to create an accurate recognition model. Therefore, skeletal-based HAR requires both temporal and spatial information. This paper examines the efficiency of using only the human skeleton when passing the CNN output to the LSTM architecture for real-time spatiotemporal recognition (Fig. 1).

Only with efficient pre-processing algorithms and computer vision can the underlying model be sufficiently simplified to ensure real-time viability, adequate accuracy, and easy hardware setup for sign language detection (Fig. 4).

An important limitation of LSTM is memory. More precisely, the deterioration of memory. In the case of a large input time step of the LSTM model, one control can be stored. The fact that this LSTM module cannot store multiple controls indicates that its result is inaccurate.

As a result of the experiments, it was found that the developed algorithm reduces the number of errors during gesture recognition for people with disabilities in real time. In [11], the number of recognition errors in a video sequence was 6.6 %, in [12] – 8.2 %, and in [13] – 8 %. In this work, as a result of using LSTM-CNN, the number of errors was 1.7 %. The proposed algorithm has reduced the level of errors when recognizing gestures in the video sequence of a webcam and can be used to create natural human-machine interfaces, special systems for deaf and dumb people, and also be used to control software on a computer using gestures.

The advantage of this method is a two-layer deep LSTM RNN with a linear repeating projection layer. At every layer, LSTM outperforms sign language recognition performance. This architecture uses model parameters more efficiently than others considered, is faster to assemble, and outperforms a deep forward neural network with a large sequence of parameters.

A promising approach to this pre-processing technique is scoring, which traces specific points on the human body as 3D orientations, allowing full reconstruction of stick figure movements. Real-time posture assessment without complex multi-camera devices or physical body markers is a new field that is still in its infancy.

7. Conclusions

1. The data taken from the open-source MediaPipe has been pre-processed, which means that the data is divided into two directions. This paper has reviewed the LSTM RNN architecture for a wide range of modeling in sign language recognition. For sign language recognition, LSTM RNNs are more

efficient than conventional RNNs for acoustic modeling given medium-sized models learned on a single machine. The first distributed LSTM RNN training was presented using asynchronous stochastic gradient dip optimization on a large cluster of machines.

2. The open-source database MediaPipe was used to recognize spatiotemporal data in real time. In real time, using a neural network, human posture, facial expressions, and hand control, fitness and sports analysis, motion control, and sign language recognition were implemented.

MediaPipe with built-in monitoring solutions has demonstrated unprecedented low latency and high accuracy tracking in real-world scenarios, which combines pose, hand, and face control with detailed levels. MediaPipe Holistic is interesting for sign language detection preprocessing, both in sign language identification preprocessing and usability for people with disabilities. As a result of calculations based on the existing model, on average, sign language is recognized with an accuracy of up to 87 %. When performing calculations by adding a new layer to this model, sign language is recognized with an accuracy of up to 98 %.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

References

1. Rastgoo, R., Kiani, K., Escalera, S. (2021). Sign Language Recognition: A Deep Survey. Expert Systems with Applications, 164, 113794. doi: https://doi.org/10.1016/j.eswa.2020.113794
2. Chuikov, A. V., Vulfin, A. M. (2017). Gesture recognition system. Vestnik UGATU, 21 (3 (77)), 113–122. Available at: https://cyberleninka.ru/article/n/sistema-raspoznavaniya-zhestov-na-osnove-neyrosetevyh-tehnologiy
3. Wang, M., Lyu, X.-Q., Li, Y.-J., Zhang, F.-L. (2020). VR content creation and exploration with deep learning: A survey. Computational Visual Media, 6 (1), 3–28. doi: https://doi.org/10.1007/s41095-020-0162-z
4. Murlin, A. G., Piotrovskiy, D. L., Rudenko, E. A., Yanaeva, M. V. (2014). Algorithms and methods for detection and recognition of hand gestures on video in real time. Politematicheskiy setevoy elektronniy nauchnyy zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta, 97, 626–635. Available at: https://www.elibrary.ru/item.asp?id=21527334
5. Rautaray, S. S., Agrawal, A. (2012). Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review, 43 (1), 1–54. doi: https://doi.org/10.1007/s10462-012-9356-9
6. Bae, S. H., Choi, I. K., Kim, N. S. (2016). Acoustic Scene Classification Using Parallel Combination of LSTM and CNN. Detection and Classification of Acoustic Scenes and Events. Available at: https://dcase.community/documents/workshop2016/proceedings/Bae-DCASE2016workshop.pdf
7. Lee, D.-H., Hong, K.-S. (2010). A Hand gesture recognition system based on difference image entropy. In 2010 6th International Conference on Advanced Information Management and Service (IMS), 410–413.
8. Chen, Y., Luo, B., Chen, Y.-L., Liang, G., Wu, X. (2015). A real-time dynamic hand gesture recognition system using kinect sensor. 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO). doi: https://doi.org/10.1109/robio.2015.7419071
9. Korchenko, A., Tereykovskiy, I., Karpinskiy, N., Tynymbaev, S. (2016). Neyrosetevye modeli, metody i sredstva otsenki parametrov bezopasnosti internet-orientirovannykh informatsionnykh sistem. Kyiv: TOV "Nash Format".
10. Top 10 Deep Learning Algorithms You Should Know in 2022. Available at: https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-algorithm
11. Liu, N., Lovell, B. C. (2003). Gesture classification using hidden markov models and viterbi path counting. In VII-th Digital image computing: techniques and Applications. Available at: https://www.researchgate.net/publication/37616560_Gesture_Classification_Using_Hidden_Markov_Models_and_Viterbi_Path_Counting
12. Phan, N. H., Bui, T. T. T., Spitsyn, V. G. (2013). Real-time hand gesture recognition base on Viola-Jones method, algorithm CAMShift, wavelet transform and principal component analysis. Upravlenie, vychislitel'naya tekhnika i informatika, 2 (23), 102–111. Available at: https://cyberleninka.ru/article/n/raspoznavanie-zhestov-na-videoposledovatelnosti-v-rezhime-realnogo-vremeni-na-osnove-primeneniya-metoda-violy-dzhonsa-algoritma
13. Tkhang, N. T., Spitsyn, V. G. (2012). Algoritmicheskoe i programmnoe obespechenie dlya raspoznavaniya formy ruki v real'nom vremeni s ispol'zovaniem surfcdeskriptorov i neyronnoy seti. Izvestiya Tomskogo politekhnicheskogo universiteta. Inzhiniring georesursov, 320 (5), 48–54.