Sign Language Recognition Method

Subjects: Computer Science, Artificial Intelligence

Contributor: Nurzada Amangeldy , Saule Kudubayeva , Akmaral Kassymova , Ardak Karipzhanova , Bibigul Razakhova , Serikbay Kuralov

Technologies for pattern recognition are used in various fields. One of the most relevant and important directions is the use of pattern recognition technology, such as gesture recognition, in socially significant tasks, to develop automatic sign language interpretation systems in real time. More than 5% of the world's population—about 430 million people, including 34 million children—are deaf-mute and not always able to use the services of a living sign language interpreter. Almost 80% of people with a disabling hearing loss live in low- and middle-income countries. The development of low-cost systems of automatic sign language interpretation, without the use of expensive sensors and unique cameras, would improve the lives of people with disabilities, contributing to their unhindered integration into society.

sign language hand shape palm definition model

1. Introduction

Globally, 432 million adults and 34 million children need rehabilitation for "disabling" hearing loss. It is estimated that by 2050, more than 700 million people—or 1 in 10 people—will have a disabling hearing loss. The prevalence of hearing loss increases with age, with over 25% of people over 60 years of age suffering from a disabling hearing loss [1].

Recently, more attention has been paid in the world to improving the quality of life of people with disabilities. Necessary conditions for movement, training, and interaction with the public for people with disabilities are being created; special hardware, software, scientific and technical products are being developed, and various social state programs and programs of inclusive education are being implemented. For scientists in countries around the world, creating a barrier-free society for people with disabilities is one of the most important tasks.

In modern Kazakhstan, one of the most important directions of state policy concerns the equal right to education for all citizens. The prerequisite for ensuring accessibility of education is an inclusive environment. The modern task of inclusive education provides intellectual development, ensuring equal access to education for all levels of the population, taking into account their psycho-physiological and individual characteristics. The process of inclusive education is conditioned by normative legal documents, such as the Law of the Republic of Kazakhstan on Education [2] and "the concept of development of inclusive education in the Republic of Kazakhstan" [3], which strengthen the requirements for professional activities of teachers, the process of barrier-free training of people with disabilities, and access to software products.

Object recognition technologies are a key factor that can provide solutions to improve the quality of life for people with disabilities. The ability of a machine to understand human gestures, interpret deaf-mute people's intentions, and react accordingly is one of the most important aspects of human–machine interaction. At the same time, gesture recognition is quite a challenge, not only because of the variety of contexts, multiple interpretations, spatial and temporal variations, and complex non-rigid hand properties, but also because of different lighting levels and complex backgrounds. The first attempts to create various automated systems capable of perceiving the world like humans were made decades ago. Over time, greatly improved, these technologies have been widely used in many fields.

2. Sign Language Recognition Method

In order to find and determine the most appropriate and most optimal approach for the development of a full-fledged automatic sign language translation system, and for the development of appropriate digital devices based on gesture recognition methods to solve the previously identified social problem, work was carried out for the deaf and mute to analyze similar works based on machine learning methods (hereinafter ML).

It should be noted that there are a great number of types of tasks solved with ML, types of ML, and algorithms of ML models; therefore, a systematic special study of the application of ML for gesture recognition tasks is required. For a more complete characterization of the issue under consideration, the works of a number of researchers have been studied. The scope of machine learning applications is very diverse. A number of scientific directions that are used in gesture recognition tasks can be distinguished: classical learning [4][5][6][7][8], ensembles [9][10][11], neural networks, and deep learning [12][13][14][15][16][17][18][19][20][21][22][23][24][25][26].

Classical learning: the simplest algorithms, characterized by direct heirs of computing machines of the 1950s. They knowingly solved formal problems, such as finding patterns in calculations and calculating the trajectory of objects. Today, methods based on classical learning are the most common. They form the recommendation module on many platforms.

When learning without a teacher, the machine itself must find the right solution among the cluttered data and sort objects by obscure features. For example, the machine may be required to classify a particular gesture among a set of data. The K-mean method is used in teacherless learning. Gani et al. [4] have applied the K-means algorithm in 2D space to divide all pixels into two groups corresponding to the hands of the signer. The K-means algorithm begins by placing K points (centroids) at random locations in 2D space. It places only two centroids that correspond to the user's hands. Each pixel is assigned to the cluster with the closest centroid, and then the new centroids are calculated as the average of the pixels assigned to it. The algorithm continues until no pixel changes its affiliation to the cluster. If the distance between two centroids is less than a constant, they are combined into one.

When learning with a teacher, the machine has a instructor who knows which answer is correct. This means that the raw data is already marked (sorted) in the right way, and the machine only has to determine the object with the

right attribute or calculate the result. The scholars Sharma et al. [5] created a set of key descriptors with identical characteristics for each class of gesture image. This set is used to create a set of feature models for all training images. K-means clustering is performed to obtain K clusters with similar descriptors. Each image fragment is correlated with the closest cluster. Then for each image, all descriptors are compared to their nearest cluster, and a codeword histogram is created. A codeword dictionary is created using a set of feature histograms. In their approach, K is taken to be 150; i.e., 150 codewords are created for each image. Subsequently, different algorithms, including *k*-NN, were applied for classification. Arshad Malik et al. [6] proposed a system which captures the input data through a web camera without using any additional equipment, and then, using segmentation approach, the hand is separated from the background, and one can extract the necessary features from the image using principal component analysis (PCA). Finally, the gesture function is classified using K-nearest neighbors (*k*-NN). Anuradha Patil et al. [7] proposed a structure using Kinect with SVM, which is linear, and *k*-NN with weights. In general, their algorithm achieved moderate accuracy and speed in most conditions. Ramesh et al. [8] proposed an algorithm which consists of two steps: training and testing. In a training set of 50 different domains, video samples are collected. Each domain contains five samples, and each video sample is assigned a word class and stored in the database. The test sample is pre-processed using median filter, canny operator for edge detection, and HOG for feature extraction. The SVM receives the input data as HOG features and predicts the class label based on the trained SVM model. Finally, a textual description is generated in the Kannada language.

Works based on classical learning are limited in the number of recognized gestures, mostly applied to dactyl alphabets of gesture languages; that is, limited in the amount of tested data, in real-time work, to a small number of gestures. Many works used different algorithms for feature extraction and classification, and they are also often used for recognition of static gestures.

Ensembles: groups of methods that use several machine learning methods at once and correct each other's errors. They include such classifiers as Random Forest and XGBoost, which boost when algorithms are trained sequentially, with each one paying special attention to the errors of the previous one.

Qin et al. [9] proposed a method of gesture recognition based on the fusion of several spatial features. These spatial features describe the shape and distribution of gestures in the local space, and one performs feature filtering, preserving the features of the discriminant information to reduce the computational cost. Researchers have experimented with two large sets of gesture data and, compared to popular methods, the method effectively improves recognition quality. In the future, people will consider how to improve the features to make them more intelligible; for example, by using convolutional neural networking and other methods to automatically learn gesture characteristics. Su et al. [10] proposed a random forest-based Chinese Sign Language (CSL) sub-word recognition method using an improved decision tree to increase the probability of obtaining the correct result from each decision tree in random forests. Based on the recognition results of 121 frequently used CSL sub-words, the superior performance of the random forest method in terms of accuracy and reliability was tested. Results with a recognition accuracy of 98.25% were obtained. Su et al. [10] proposed a system in which they introduced a two-stage pipeline based on two-dimensional body connection positions extracted from RGB camera data. First, the system divides the signed expression data stream into meaningful word segments based on a frame-by-frame

binary random forest. Each segment is then converted into an image-like form and classified using a convolutional neural network. The proposed system is then evaluated on a data set of continuous Japanese gesture language sentence expressions with variations of non-manual expressions. By exploring a variety of data representations and network parameters, researchers can distinguish verbal segments of specific non-manual intonations from the underlying body joint motion data with 86% accuracy.

Kenshimov et al. [11] proposed a system of dactylic alphabet recognition of Kazakh Sign Language based on SVM, Extreme Gradient Boosting, and Random Forest. The Kazakh Sign Language dactyl alphabet has 42 letters, but in their work 31 classes were distinguished; that is, two-handed and dimaic gestures were not included, in contrast to the system.

Ensemble methods are a machine learning paradigm in which multiple models (often called weak learners or baseline models) are trained to solve the same problem and combined to improve performance. The basic hypothesis is that if people combine weak learners correctly, people can obtain more accurate and/or reliable models.

Neural networks and deep learning: the most complex level of AI learning. Neural networks simulate the work of the human brain, which consists of neurons constantly forming new connections with each other. They can be conventionally defined as a network with many inputs and one output. Neurons form layers through which a signal sequentially passes. All this is connected by neuronal connections, or channels, through which data are transmitted. Each channel has its own “weight”—a parameter that affects the data it transmits.

The AI collects data from all inputs, evaluating their weight according to given parameters, and then performs the desired action and outputs the result. At first, it is random, but then, through many cycles, it becomes more and more accurate. A well-trained neural network works like a normal algorithm, or more accurately.

The real breakthrough in this field has been deep learning, which trains neural networks at multiple levels of abstraction.

Deep neural networks are the first to learn how to recognize gestures, one of the most complex objects for AI. They do this by breaking them into blocks, identifying the dominant lines in each, and comparing them to other images of the desired object [11][12][13][14][15][16][17][18][19][20][21][22][23][24][25][26].

Recurrent neural networks [15][27] are mainly used for text and speech recognition. They identify sequences in them and associate each unit—a letter or sound—with the rest.

For machine learning algorithms, it is important that the data arriving to the input of the algorithm can accurately describe the properties of the object, provide accurate information about the object, and the volume of incoming input data. Because the amount of incoming data depends on the speed of data processing, the requirement for machine performance and the accuracy of object recognition depends directly on the accuracy of the input data. The MediaPipe technology used in the work allows one to solve these problems, and at the input of the artificial

neural network it provides only the coordinates of 21 points and the trajectory of change of each point. In addition, it removes the load on the algorithm used.

In their study, Nafis and Ayas Faikar used the wrist position recommended by MediaPipe. The Shift-GCN model included modification of the moving weight of the main points of the obtained palmar joints. The study used the values of the main points of the hand as a data set [17]. Caputo and Ariel created the SHREC 2021: Track system, which recognizes hand gestures based on the hand skeleton. With Leap Motion, they created many datasets for 18 character classes, and the datasets were learned and recognized by the ST-GCN model. The researchers wrote that there were some errors in the recognition of dynamic gestures [28].

Halder et al. proposed a method based on the open-source MediaPipe platform and a machine learning algorithm. The model is lightweight and can be adapted to a smart device with American, Indian, and Turkish sign languages serving as the data set. The reliability and accuracy of the proposed models are estimated at 99% [29]. The algorithm proposed by Gomase and Ketan using Mediapipe and recognition using computer vision was partially successful, and accurate at an average of 17 frames per second, with an average accuracy of 86 to 91% [30]. In their papers, Alvin and Arsheldy proposed an American Sign Language recognition system based on Mediapipe and K-mean. Thus, Mediapipe is one of the most advanced real-time gesture recognition technologies [31].

Chakraborty proposed a methodology for classifying English alphabets rendered by various Indian Sign Language (ISL) hand gestures using the Mediapipe Hands API launched by Google. The purpose of using this API is to find the 21 significant points in each hand along with their x, y, and z coordinates in 3D space. Due to the lack of a proper dataset available on the internet for ISL, at the very beginning, they created a dataset of 15,000 per English character, each consisting of the coordinates of 21 points recognized by the Mediapipe Hands API [32].

The listed studies [28][29][30][31][32] have significantly contributed to the development of multimodal gesture corpora. However, the problem of calculating the frame depth map for gesture recognition using the Mediapipe technology is still relevant. The image data is first acquired with a simple camera. The coordinates of the human palm joints are then computed using the BlazePalm single finger detector model, which is available in MediaPipe. In Kazakh Sign Language, SVM was used for multiple classifications of various numbers and letters. The novelty of the proposed algorithm is that, firstly, there are no full-fledged systems for recognizing the Kazakh dactyl alphabet today, and secondly, a unified draw_people functionality was integrated into the system for recording and demonstrating gestures in real time. Moreover, users can configure the frame depth map mode in that mode, contributing to the achievement of superior results.

The task of creating universal multimodal gesture corpora arises due to the solution of several unimodal subtasks: recognizing hand gestures, identifying the movement of the body and head, and recognizing facial emotions. The listed tasks are fraught with problems of spatial and temporal variations and complex non-rigid properties of hands, different levels of illumination, and complex backgrounds. Recognizing hand gestures is a semiotic task in which the dactyl alphabet and tracing speech are used. The dactyl alphabet is used for the introduction and transmission

of the sound of new words (for example, proper names), for which there are no ready-made means of sign language. Tracing signed speech is a secondary sign system that traces the sounding language's linguistic fabric.

S. Zhang et al. [33] propose a sign language recognition structure that combines RGB-B input and two-stream space–time networks. The ARS approach covers key information for aligned multimodal input data and effectively eliminates redundancy. The local focus of the hand optimizes the input of the spatial network. In addition, the D-shift network generates depth movement features to investigate depth information efficiently. Subsequently, convolution fusion is performed to merge the two feature streams and improve recognition results. Yu et al. [34] proposed the SKEPRID system, a repeated recognition method resistant to significant body posture and lighting changes. By including information about the skeleton, they reduced the influence of various poses and developed a set of light-independent features based on the skeleton, significantly increasing the accuracy of repeated recognition.

Luqman et al. [35] presented a new multimodal video database for sign language recognition. Unlike existing databases, the database focuses on signs that require both manual and non-manual articulators, which can be used in various studies related to sign language recognition. Two cases were considered for sign-dependent and sign-independent modes using manual and non-manual signs. In the first case, researchers used color and depth images directly, while in the second, researchers used optical flow to extract more relevant features related to the signs themselves and not to the signatories. The best results were obtained using MobileNet-LSTM with transfer training and fine-tuning: 99.7% and 72.4% for the "sign-dependent" and "sign-independent" modes, respectively. Kagirov et al. [36] also presented the Russian multimedia database of the Russian sign language. The database includes lexical units (individual words and phrases) from the Russian sign language within one thematic area, called "food in the supermarket", and was collected using the MS Kinect 2.0 device with Full HD video modes and depth maps. They provide new opportunities for a lexicographic description of the vocabulary of Russian sign language and expand research in the field of automatic gesture recognition.

D. Ryumin et al. [37] proposed an approach for the detection and recognition of 3D gestures of one hand for human–machine interaction. The logical structure of the system modules for recording the gesture database is described. The logical structure of the 3D gestures database is presented. Examples of frames demonstrating gestures in full high definition format, in map depth mode, and the infrared range are given. Models of a deep convolution network for recognizing faces and hand shapes are described. The results of automatic detection of the area with the face and the shape of the hand are given.

The works 44–48 listed above aim to calculate depth using a D-shift network that generates depth movement features, or a Kinect camera that provides depth information to increase gesture recognition. However, researchers calculate the depth of the frame using a simple draw_people functionality, which also contributed to the improvement of the indicator without using additional models or special cameras.

References

1. World Health Organization (WHO). Deafness and Hearing Loss. Available online: http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (accessed on 2 August 2022).
2. Law of the Republic of Kazakhstan “On Education”. Available online: https://adilet.zan.kz/kaz/docs/Z070000319_ (accessed on 2 August 2022).
3. The Concept of Development of Inclusive Education in Kazakhstan. Available online: https://legalacts.egov.kz/application/downloadconceptfile?id=2506747 (accessed on 2 August 2022).
4. Gani, E.; Kika, A. Albanian Sign Language (AlbSL) Number Recognition from Both Hand’s Gestures Acqu4red by Kinect Sensors. Int. J. Adv. Comput. Sci. Appl. 2016, 7.
5. Sharma, A.; Mittal, A.; Singh, S.; Awatramani, V. Hand Gesture Recognition using Image Processing and Feature Extraction Techniques. Procedia Comput. Sci. 2020, 173, 181–190.
6. Malik, M.S.A.; Kousar, N.; Abdullah, T.; Ahmed, M.; Rasheed, F.; Awais, M. Pakistan Sign Language Detection using PCA and KNN. Int. J. Adv. Comput. Sci. Appl. 2018, 9.
7. Patil, A.; Tavade, C.M. Performance analysis and high recognition rate of automated hand gesture recognition though GMM and SVM-KNN classifiers. Int. J. Adv. Trends Comput. Sci. Eng. 2020, 9, 7712–7722.
8. Kagalkar, R.M.; Gumaste, S.V. Mobile Application Based Translation of Sign Language to Text Description in Kannada Language. Int. J. Interact. Mob. Technol. 2018, 12, 92–112.
9. Qin, M.; He, G. Gesture Recognition with Multiple Spatial Feature Fusion. In Proceedings of the 2016 4th International Conference on Machinery, Materials and Computing Technology, Changsha, China, 18–20 March 2016; Atlantis Press: Amsterdam, The Netherlands, 2016.
10. Su, R.; Chen, X.; Cao, S.; Zhang, X. Random Forest-Based Recognition of Isolated Sign Language Subwords Using Data from Accelerometers and Surface Electromyographic Sensors. Sensors 2016, 16, 100.
11. Kenshimov, C.; Buribayev, Z.; Amirgaliyev, Y.; Ataniyazova, A.; Aitimov, A. Sign language dactyl recognition based on machine learning algorithms. East.-Eur. J. Enterp. Technol. 2021, 4, 58–72.
12. Koller, O.; Zargaran, S.; Ney, H.; Bowden, R. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. Int. J. Comput. Vis. 2018, 126, 1311–1325.
13. Koller, O.; Zargaran, S.; Ney, H.; Bowden, R. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In Proceedings of the British Machine Vision Conference 2016, York, UK, 19–22 September 2016.

14. Raj, H.; Duggal, A.; Uppara, S. Hand Motion Analysis using CNN. Int. J. Soft Comput. Eng. 2020, 9, 26–30.
15. Bendarkar, D.S.; Somase, P.A.; Rebari, P.K.; Paturkar, R.R.; Khan, A.M. Web Based Recognition and Translation of American Sign Language with CNN and RNN. Int. J. Online Biomed. Eng. 2021, 17, 34–50.
16. Rahim, A.; Islam, R.; Shin, J. Non-Touch Sign Word Recognition Based on Dynamic Hand Gesture Using Hybrid Segmentation and CNN Feature Fusion. Appl. Sci. 2019, 9, 3790.
17. Nafis, A.F.; Suciati, N. Sign Language Recognition on Video Data Based on Graph Convolutional Network. J. Theor. Appl. Inf. Technol. 2021, 99.
18. Adithya, V.; Rajesh, R. A Deep Convolutional Neural Network Approach for Static Hand Gesture Recognition. Procedia Comput. Sci. 2020, 171, 2353–2361.
19. Ahuja, R.; Jain, D.; Sachdeva, D.; Garg, A.; Rajput, C. Convolutional Neural Network Based American Sign Language Static Hand Gesture Recognition. Int. J. Ambient Comput. Intell. 2019, 10, 60–73.
20. Rahim, A.; Shin, J.; Yun, K.S. Hand Gesture-based Sign Alphabet Recognition and Sentence Interpretation using a Convolutional Neural Network. Ann. Emerg. Technol. Comput. 2020, 4, 20–27.
21. Bastwesy, M.R.M.; Elshennawy, N.M.; Saidahmed, M.T.F. Deep Learning Sign Language Recognition System Based on Wi-Fi CSI. Int. J. Intell. Syst. Appl. 2020, 12, 33–45.
22. Xiao, Q.; Chang, X.; Zhang, X.; Liu, X. Multi-Information Spatial–Temporal LSTM Fusion Continuous Sign Language Neural Machine Translation. IEEE Access 2020, 8, 216718–216728.
23. Hossain, B.; Adhikary, A.; Soheli, S.J. Sign Language Digit Recognition Using Different Convolutional Neural Network Model. Asian J. Res. Comput. Sci. 2020, 16–24.
24. Sai-Kumar, S.; Sundara-Krishna, Y.K.; Tumuluru, P.; Ravi-Kiran, P. Design and Development of a Sign Language Gesture Recognition using Open CV. Int. J. Adv. Trends Comput. Sci. Eng. 2020, 9, 8504–8508.
25. Mohammed, A.A.Q.; Lv, J.; Islam, S. A Deep Learning-Based End-to-End Composite System for Hand Detection and Gesture Recognition. Sensors 2019, 19, 5282.
26. Khari, M.; Garg, A.K.; Gonzalez-Crespo, R.; Verdu, E. Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks. Int. J. Interact. Multimed. Artif. Intell. 2019, 5, 22.
27. Jia, Y.; Ding, R.; Ren, W.; Shu, J.; Jin, A. Gesture Recognition of Somatosensory Interactive Acupoint Massage Based on Image Feature Deep Learning Model. Trait. Signal 2021, 38, 565–572.

28. Caputo, A.; Giachetti, A.; Soso, S.; Pintani, D.; D'Eusanio, A.; Pini, S.; Borghi, G.; Simoni, A.; Vezzani, R.; Cucchiara, R.; et al. SHREC 2021: Skeleton-based hand gesture recognition in the wild. Comput. Graph. 2021, 99, 201–211.
29. Halder, A.; Tayade, A. Sign Language to Text and Speech Translation in Real Time Using Convolutional Neural Network. Int. J. Res. Publ. Rev. 2021, 8, 9–17.
30. Gomase, K.; Dhanawade, A.; Gurav, P.; Lokare, S. Sign Language Recognition using Mediapipe. Int. Res. J. Eng. Technol. 2022, 9.
31. Alvin, A.; Husna-Shabrina, N.; Ryo, A.; Christian, E. Hand Gesture Detection for American Sign Language using K-Nearest Neighbor with Mediapipe. Ultim. Comput. J. Sist. Komput. 2021, 13, 57–62.
32. Chakraborty, S.; Bandyopadhyay, N.; Chakraverty, P.; Banerjee, S.; Sarkar, Z.; Ghosh, S. Indian Sign Language Classification (ISL) using Machine Learning. Am. J. Electron. Commun. 2021, 1, 17–21.
33. Zhang, S.; Meng, W.; Li, H.; Cui, X. Multimodal Spatiotemporal Networks for Sign Language Recognition. IEEE Access 2019, 7, 180270–180280.
34. Yu, T.; Jin, H.; Tan, W.-T.; Nahrstedt, K. SKEPRID. ACM Trans. Multimedia Comput. Commun. Appl. 2018, 14, 1–24.
35. Luqman, H.; El-Alfy, E.-S. Towards Hybrid Multimodal Manual and Non-Manual Arabic Sign Language Recognition: mArSL Database and Pilot Study. Electronics 2021, 10, 1739.
36. Kagirov, I.; Ivanko, D.; Ryumin, D.; Axyonov, A.; Karpov, A. TheRuSLan: Database of Russian sign language. In Proceedings of the LREC 2020—12th International Conference on Language Resources and Evaluation, Conference Proceedings, Marseille, France, 11–16 May 2020.
37. Ryumin, D.; Kagirov, I.; Ivanko, D.; Axyonov, A.; Karpov, A.A. Automatic detection and recognition of 3d manual gestures for human-machine interaction. ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 2019, XLII-2/W12, 179–183.