applied sciences

MDPI

Article

A Real-Time Dynamic Gesture Variability Recognition Method Based on Convolutional Neural Networks

Nurzada Amangeldy [1], Marek Milosz [2,*], Saule Kudubayeva [1], Akmaral Kassymova [3], Gulsim Kalakova [4] and Lena Zhetkenbay [1]

[1] Department of Artificial Intelligence Technologies, Faculty of Information Technologies, L.N. Gumilyov Eurasian National University, Pushkina 11, Astana 010008, Kazakhstan; nurzadaamangeldy@gmail.com (N.A.); saulekudubayeva@gmail.com (S.K.); jetlen7@gmail.com (L.Z.)
[2] Department of Computer Science, Faculty of Electrical Engineering and Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland
[3] Higher School of Information Technologies, Faculty of Economics, Information Technology and Vocational Education, Zhangir Khan University, Zhangir Khan 51, Uralsk 090009, Kazakhstan; kasimova.akmaral2020@gmail.com
[4] Department of Physics, Mathematics and Digital Technology, A. Baitursynov Kostanay Regional University, Kostanay 110000, Kazakhstan; gulsim_1507@mail.ru
* Correspondence: m.milosz@pollub.pl

Abstract: Among the many problems in machine learning, the most critical ones involve improving the categorical response prediction rate based on extracted features. In spite of this, it is noted that most of the time from the entire cycle of multi-class machine modeling for sign language recognition tasks is spent on data preparation, including collection, filtering, analysis, and visualization of data. To find the optimal solution for the above-mentioned problem, this paper proposes a methodology for automatically collecting the spatiotemporal features of gestures by calculating the coordinates of the found area of the pose and hand, normalizing them, and constructing an optimal multilayer perceptron for multiclass classification. By extracting and analyzing spatiotemporal data, the proposed method makes it possible to identify not only static features, but also the spatial (for gestures that touch the face and head) and dynamic features of gestures, which leads to an increase in the accuracy of gesture recognition. This classification was also carried out according to the form of the gesture demonstration to optimally extract the characteristics of gestures (display ability of all connection points), which also led to an increase in the accuracy of gesture recognition for certain classes to the value of 0.96. This method was tested using the well-known Ankara University Turkish Sign Language Dataset and the Dataset for Argentinian Sign Language to validate the experiment, which proved effective with a recognition accuracy of 0.98.

Keywords: sign language recognition; CNN; multimodality; preprocessing

check for updates

Citation: Amangeldy, N.; Milosz, M.; Kudubayeva, S.; Kassymova, A.; Kalakova, G.; Zhetkenbay, L. A Real-Time Dynamic Gesture Variability Recognition Method Based on Convolutional Neural Networks. *Appl. Sci.* **2023**, *13*, 10799. https://doi.org/10.3390/app131910799

Academic Editor: Douglas O'Shaughnessy

Received: 24 May 2023
Revised: 20 July 2023
Accepted: 9 August 2023
Published: 28 September 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recently, increasing attention has been paid to improving the quality of life of people with disabilities. The necessary conditions for their mobility, training, and interaction with society are being created; special hardware, software, scientific, and technical products are being developed; and various inclusive social and education programs are being implemented [1]. For scientists worldwide, creating a barrier-free society for people with disabilities is one of the most crucial tasks [2].

Hearing loss has become a common problem globally. According to the World Health Organization [3], 432 million adults and 34 million children are projected to have some degree of hearing loss by 2050, and approximately 700 million of these people will need hearing rehabilitation. This increases the dependence on sign language, which is the primary communication language for persons with various levels of hearing disorders.

Just as spoken languages vary greatly from region to region, sign languages also reflect the unique cultural and linguistic characteristics of the communities in which they are used. Sign languages have developed and evolved differently in different parts of the world, and are influenced by the cultural traditions and languages of deaf and hearing people in those regions. For example, American Sign Language (ASL) [1,4] and British Sign Language (BSL) [5] have some similarities, but at the same time are different languages with their own grammar, syntax, and vocabulary. A similar situation is observed for sign languages around the world. Some countries, like Kazakhstan and other post-Soviet countries, were once under the influence of the Russian Empire. Until 1991, they were part of the Soviet Union, whose centralized language policy led to the spread of Russian Sign Language (RSL) in the Soviet republics [6,7]. Specialists should pay attention to sign languages, as it is necessary to consolidate their status as separate systems and prevent their extinction.

Sign languages used around the world are unique, but the way information is transmitted between deaf people all over the world is the same: it is transmitted spatially and visually, and it is perceived visually. An important property that was adopted from the visual areas of the human brain initially served as a prototype for the creation of neural networks, and later deep neural networks. Scientists all over the world who study and research gesture language recognition face the problem of continuous multimodal sign language interpretation, which consists of monomodal tasks of recognizing the configuration of the hand that performs the gesture, hand movements in space relative to the body of the speaker or camera, lip movements during the gesture demonstration, and emotions during the gesture demonstration.

Modern research results make it clear that machine learning methods [8–16] based on deep neural networks [8–15], in comparison with traditional classical approaches [7,16,17], which are based on linear classifiers (e.g., the support vector method (SVM) or hidden Markov classification (HMC)), can demonstrate good results in segmentation, classification, and recognition as static and dynamic sign language (SL) elements. Thus, with the help of two-threaded convolutional neural networks [2,9,10], it is possible to extract the spatiotemporal features of a gesture from full-color images (RGB format) and 3D frames (depth map) of video streams separately. In turn, a deep neural network (DNN) is used to perform segmentation and classification of the hand shape with multiple architectures and sizes of the input images [4,11]. In addition, it was revealed that the architecture of the long short-term memory (LSTM) neural networks [1,3,12,13,15,18–20], with the help of long and short-term memory, can extract the spatiotemporal characteristics of a gesture from sequences of previously annotated 2D regions with a gesture. These methods considered extract spatial and temporal information at different stages or separately. Thus, if there is a complex dynamic background component on the scene, concurrently extracting both the spatial and temporal components of a gesture will be an effective solution.

Thejowahyono et al. [8] used a DNN to recognize hand gestures in real-time for seeking help, achieving an accuracy of 98.79%. Jose-Jimenez and others [9] presented an American Sign Language translator based on convolutional neural networks (CNN) that consistently classify the letters a-e correctly and a model that correctly classifies the letters a-k in most cases. The network proposed by Sahoo [4] and others eliminates the need for the preprocessing steps for the Massey University (MU) color image dataset by using a DNN for segmentation and classification of the hand shape. The system was tested in real-time on a community of 10 people, with an average accuracy of 81.74% and processing speed of 61.35 frames per second. Reference [10] proposes a comprehensive end-to-end system that uses hand gestures with sign language digits for contactless user authentication, achieving a 99.1% accuracy on the test dataset with a model output rate of 280 ms per image frame. Finally, study [11] presents a machine learning-based system that recognizes Turkish sign language in real-time with an accuracy of 98.97% using a cascade voting approach with five single classifiers. These studies demonstrate the potential of deep learning models for improving people's safety, simplifying processes, and creating contactless solutions for authentication and communication.

Several research studies have proposed innovative systems for recognizing SL gestures in real time. One such system [12] uses electromyography (EMG) signals and a neural network with long-term and short-term memory to recognize hand movements in ASL. Another system [13] employs bidirectional long short-term memory (BLSTM) and multifunctional structures to recognize dynamic gestures in Hong Kong Sign Language (HKSL) using data collected from smartwatches. Zhou, Z and et al. [14] developed a framework called SignBERT, which uses bidirectional encoder representations from transformers (BERT) based on a deep learning models and multimodal inputs to recognize dynamic gestures for continuous sign language recognition (CSLR). This system was tested on several datasets, including a new set of HKSL data, showing advanced results in recognition accuracy. Additionally, an experimental system was developed to support deaf people in the office when applying for an identity card using Polish Sign Language (PSL) expressions. The system [15] uses a Kinect sensor and a feature vector inspired by linguistic research. It was tested using three commonly used methods for dynamic gesture recognition.

While these studies offer innovative systems for sign language gesture recognition, they are not without limitations. For example, using EMG signals is one of the systems [12] that may require invasive procedures, which may limit its applicability in real conditions. Another system [13] uses data from smartwatches, which may not capture all the nuances of sign language gestures. The use of Leap Motion technology in one of the systems [16] can also create problems with predicting the position of fingers. Leap Motion controllers also have problems with detecting some hand positions (e.g., vertical, in which one finger is covered by another). Finally, there is a problem in considering all the variations and the complexity of SL for a more accurate and reliable interpretation of SL in general.

The authors of [16] present a system that uses data from a Leap Motion (LM) device and the HMC algorithm to learn gestures, achieving an average gesture recognition accuracy of 86.1% with a standard deviation of 8.2%. The system was tested on ASL gestures and was also able to recognize them with a typing speed of 3.09 words per minute. However, the authors pointed out some issues with using LM technology for gesture recognition. For instance, when the user's fingers are not visible to the infrared (IR) cameras, LM may make mistakes in predicting their position. In such cases, the hand may be depicted with folded fingers when they are actually extended. Additionally, the position of the thumb when pressed against the palm or between other fingers is poorly defined, making it difficult to reliably identify the gesture.

The authors took a step forward in their study [6,7] by including information about the movements of the human body and head together with the movements of the hands for a multimodal analysis of human movements. By recording changes in a person's posture, 3D head movement, and hand movement in a vector matrix, the authors strive to obtain a total picture of the dynamic gesture being performed.

Approaches to gesture recognition [21–24] based on wearable electronic input devices, which are gloves worn on the hand and containing various electronic sensors that track the movements of the hand, demonstrate fairly good gesture recognition results. However, these approaches have significant disadvantages, such as the size and material of the gloves and the complexity of connection and configuration, making them unsuitable, for example, for the elderly or for people with skin diseases.

The authors of [25] have developed a multilingual approach in which hand movement modeling is also carried out using data that is independent of the target sign language through derivation of subunits of hand movement. The scientists tested the proposed approach by researching Swiss, German, and Turkish sign languages and demonstrated that SL recognition systems could be effectively developed using multilingual sign language resources.

By integrating information from different modalities, the system can better understand the nuances of SL and accurately translate them into written language, greatly improving the accuracy and efficiency of automatic SL interpretation systems and making them more accessible to a wide range of users. Continuing the idea proposed in reference [25], in

this paper, the authors propose a methodology for improving the prediction of categorical responses in SL recognition tasks by automatically collecting the spatiotemporal characteristics of gestures and grouping them by their form using the method of principal components, then selecting the optimal architecture that is suitable for all groups of gestures. This proposed method is also aimed at reducing the time spent on data preparation by automating data collection. The proposed method was tested on the popular datasets, including the Ankara University Turkish Sign Language Dataset (AUTSL) [26] and the Dataset for Argentinian Sign Language (LSA64) [27]. The experiment confirmed the system's effectiveness and multilingualism.

2. Purpose and Summary of the Work

This work's primary focus is the development of multimodal systems [28] for automatic sign language translation.

Firstly, the authors strive to overcome the problems of recognizing sign language by integrating several modalities, such as hand movement, hand configuration, and body and head movements. Secondly, the authors try to cover various variations in the demonstration of gestures and group them into four separate categories (described further) for a more accurate and reliable interpretation of sign language. Thirdly, the authors develop a preprocessing module that allows one to collect gesture properties from video files or in real-time automatically. This makes the technique applicable to any sign language and allows for further experiments. Fourth, the authors create their own extensive Kazakh Sign Language (KSL) dataset consisting of 2400 training, test, and validation samples, which helps solve the problem of limited datasets in gesture recognition research. Finally, the authors test their architecture on various public data sets, contributing to ongoing efforts to improve gesture recognition technologies.

The presented work makes it possible to answer the following formulated research questions:

(1) Is it possible in practice to use technologies that integrate hand, body, and head movements to improve sign language recognition?
(2) Is it possible to group gestures for a more accurate and reliable interpretation of sign language?
(3) Is it possible to create preprocessing software that allows for gesture properties to be automatically collected from video files or in real time from a video camera system?
(4) Is the created architecture of the real-time dynamic gesture variability recognition system based on CNNs effective for different data sets?

3. Materials and Methods

The authors propose an approach that includes several modules for dynamic gesture recognition, including a video preprocessing module for extracting key points from recorded video sets, a video recording module for extracting key points in real-time, a model creation module, and a real-time gesture demonstration system.

The proposed method with these modules is presented in Figure 1. It consists of preprocessing modules (for video files and real-time videos), a module for training and testing CNNs, and a developed model.

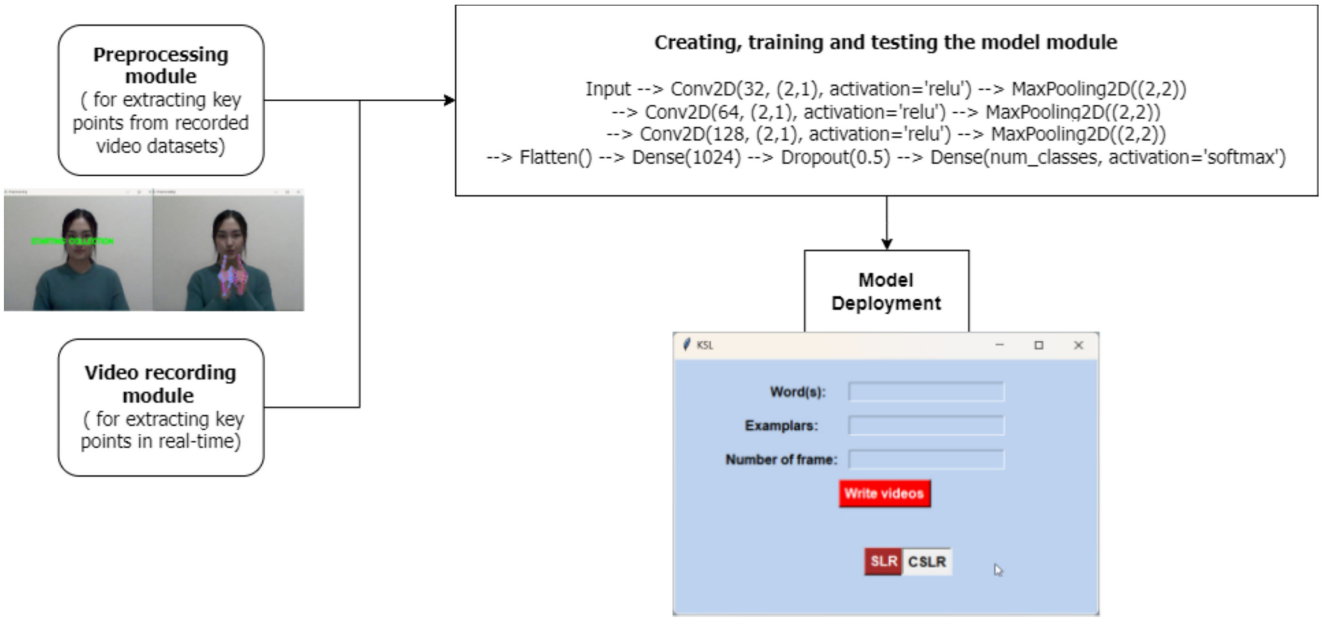

Figure 1. The proposed method.

3.1. Preprocessing Module

The MediaPipe Hand tool [29], an open-source software tool created by Google, is a comprehensive hand tracking and gesture recognition tool. It uses a set of 21 connections for each hand, including four for each finger, one at the center of the palm, and one at the wrist. By analyzing the movement and position of these connections, the tool can accurately track the rotation and movement of the hand in three-dimensional space. In addition to tracking the hand, the tool can also track the body skeleton, including the hand, pose, and head relative to the camera. This allows one to analyze dynamic gestures that involve movements of the entire body, not just the hands, using a 2D camera.

In the preprocessing module, key gesture points are extracted using the MediaPipe Holistic [30] pipeline. It is designed to overcome the limitations of individual models for pose and hand components. The MediaPipe Holistic pipeline starts with an assessment of a person's posture using the Blazes pose detector and a subsequent reference model. Then, using the found pose landmarks, two regions of interest (ROI) for each hand are determined, and a reframing model is used to improve the ROI. Then, the full-resolution input frame is cropped to these areas of interest, and hand models specific to the gesture recognition task are used to evaluate the corresponding landmarks.

In the final step, all the landmarks are combined with the pose model landmarks to obtain a total of 258 coordinates. These coordinates consist of 33 values for the pose (x, y, z) and 33 values for visibility, resulting in a total of $33 \times 4 = 132$. Additionally, there are 22 values (x, y, z) for each hand, accounting for a total of $22 \times 2 \times 3 = 126$ coordinates.

This multi-stage approach allows the MediaPipe Holistic pipeline to achieve a high level of accuracy in detecting human pose and hand movements. Let us break it down further:

Landmark Extraction: Initially, landmarks specific to the pose of a person, including body joints and facial features, are extracted using a pose estimation model. These landmarks provide information about the overall body position and orientation.

Pose Landmarks: The pose landmarks consist of 33 coordinates (x, y, z) and 33 visibility values. The (x, y, z) coordinates represent the spatial position of each landmark, while the visibility values indicate whether the landmark is occluded or visible.

Hand Landmarks: Hand landmarks are extracted separately for each hand using a hand tracking model. Each hand consists of 21 landmarks, resulting in a total of 22 values (including a wrist landmark). The (x, y, z) coordinates of these landmarks capture the hand's position, orientation, and finger movements.

Combining Landmarks: Finally, the extracted pose landmarks and hand landmarks are combined to form a comprehensive representation of the person's pose. By merging these landmarks, the system obtains a total of 258 coordinates (132 from pose landmarks and 126 from hand landmarks) that collectively describe the person's body posture and hand gestures.

This multi-step process enables the MediaPipe Holistic pipeline to accurately analyze and interpret human pose and hand movements, facilitating various applications such as gesture recognition, motion tracking, and augmented reality.

After collecting the critical points of the dynamic gesture, when viewing the captured data, it is observed that not all the connections were displayed; based on this, the motions were grouped according to the demonstration form to embrace the maximum possible forms of gesture demonstration, taking into account the following properties:

1. Palm orientation:
   - Palms facing the camera—gestures in which the palm faces the camera, and the fingers are visible to the viewer; for example, pointing or greeting.
   - Palms facing away from the camera—gestures in which the back of the palm faces the camera, and the fingers are not visible, for example, a demonstration of refusal or disapproval.
2. Localization:
   - Upper body—gestures involving movements of the arms and hands above the waist; for example, waving, clapping, or stretching the hand.
   - Lower body—gestures related to the movements of the legs and feet; for example, walking, running, or jumping.
3. Trajectory of movement:
   - Parallel to the camera—gestures in which the hand moves in the same plane as the camera; for example, waving or gesticulating horizontally.
   - Perpendicular to the camera—gestures in which the hand moves towards or away from the camera; for example, points or stretches.

Considering the above properties, four groups of dynamic gestures were selected, ten words were chosen from each group (i.e., 40 words—Table A1), and each word was recorded 50 times by demonstrators of different ages and genders. As a result, a training sample of 30,960,000 elements was obtained.

The principal component method was utilized to visualize the data, which comprised a total of 30,960,000 elements. The dataset consisted of 40 dynamic gestures (words), which were further divided into 2000 records. Each gesture was recorded in 50 instances, with each instance comprising 60 frames.

The gestures were characterized by 258 parameters, including pose coordinates and visibility values. The pose coordinates consisted of 33 values each for the x, y, and z coordinates, while the visibility values indicated the visibility of each pose point. This resulted in a total of $33 * 4 = 132$ pose parameters. Additionally, the hand coordinates included 22 values each for the x, y, and z coordinates for both hands, totaling $22 * 2 * 3 = 126$ hand parameters.

Each video file consists of 60 frames (Equation (1), and information about each frame is represented as a vector matrix (Equation (2) comprising 258 elements:

$$V = \{f_1,\ f_2, f_3 \dots f_{60}\} \tag{1}$$

$$f_1 = \{x_{p1},\ y_{p1}, z_{p1}, v_{p1} \dots\ x_{p33},\ y_{p33}, z_{p33}, v_{p33}\ ,\ x_{lh1},\ y_{lh1}, z_{lh1} \dots\ x_{lh21},\ y_{lh21}, z_{lh21}\ , x_{rh1},\ y_{rh1}, z_{rh1} \dots\ x_{rh33},\ y_{rh33}, z_{rh33}\ \} \tag{2}$$

where: *p*—pose, *v*—visibility, *rh*—right hand, *lh*—left hand.

Principal component analysis (PCA) was utilized to identify similar groups of observations (gestures) based on their principal component values. PCA allows for the discovery of patterns within high-dimensional data and representation of such patterns in a lower-dimensional space without significant loss of information. The purpose of the

PCA model was to examine predefined groups of gestures based on their demonstration form. The input data for PCA (Figure 2) consisted of the dimensions of the data, including the number of samples and variables, which were divided into four gesture clusters. Once 40 observations were obtained, the data were then inputted into the K-means algorithm to further divide the observations into four distinct clusters (as depicted by different colors in Figure 2: blue, purple, orange, and yellow). Each observation was assigned to the cluster with the closest mean value.

Group 1 (blue color in Figure 2, Figure 3a): gesture groups where the orientation of the palm is directed to the camera or the demonstrator, the connections are as readable as possible, and the trajectory of movement is parallel to the camera or static.

Group 2 (purple color in Figure 2, Figure 3b): gesture groups where the orientation of the palm is parallel to the camera, but the movement is perpendicular to the camera since the movement's trajectory for the camera is a sequence of the skeleton of the hand.

Group 3 (orange color in Figure 2, Figure 3c): gesture groups where the orientation of the palm is not directed at the cameras or the demonstrator, the connections are not all readable, and the movement is parallel relative to the camera. The orientation of the palm and the movement is perpendicular to the camera.

Group 4 (yellow color in Figure 2, Figure 3d): gesture groups in which the palm orientation and movement are perpendicular to the camera or interlocutor when demonstrating.

By grouping gestures in this way, the authors seek to account for the variations and complexities in gesture languages and to improve the accuracy and efficiency of recognition systems.

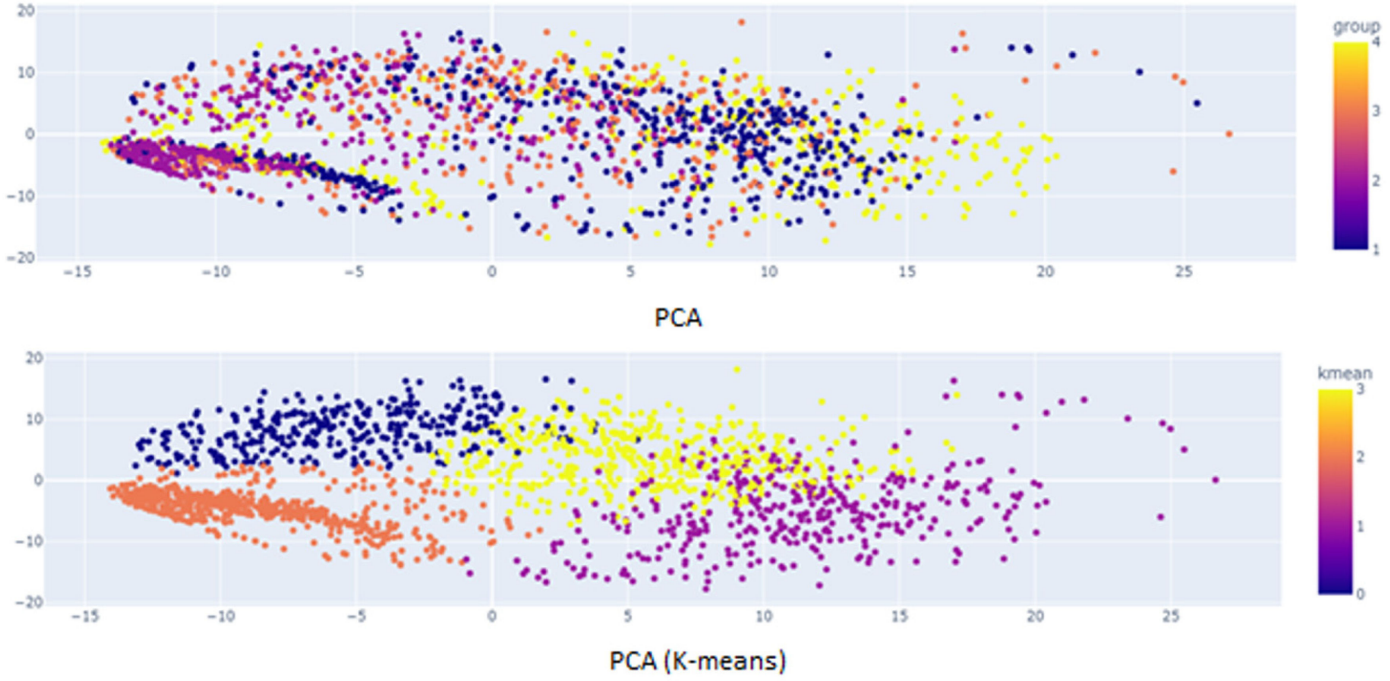

Figure 2. Clusters of gesture groups.

Figure 3. Samples of the 1st, 2nd, 3rd, and 4th groups of gestures (starting from the top; see text). Publication with the consent of the people shown in the figure.

3.2. Video Recording Module

The video recording module makes it possible to record one or more gestures in real-time, followed by data extraction for model training.

When recording a gesture, its name, the number of instances, and the total number of frames are indicated. The developed system is focused on the use of an ordinary everyday camera without additional devices, which makes it budget friendly. For recording, in this case, a Logitech WebCam C270 USB was used (Logitech, Lausanne, Switzerland), which provides video with a resolution of 1280 × 720 and a frequency of up to 30 frames per second. The system can work with any camera, providing a resolution higher than 480 × 640.

When capturing key points in real-time, the number of words, number of instances, and number of frames can be specified; hence, the resulting tensors are of different sizes for video files and for real-time recording.

One of the main features of the system is that it works in real time, which requires fast and efficient recording, analysis, and preprocessing of the video. To reduce the amount of

data to be processed, the recording begins from the moment of detection from the Media Pipe, resulting in 60 frames on average for each instance of the gesture. This ensures fast and efficient video data processing, allowing the system to provide results in real time.

3.3. Creating, Training, and Testing the Model Module

One of the popular approaches to recognizing dynamic gestures is using convolutional neural networks. The number of convolutional layers in a CNN may affect its ability to extract features from tensors. However, too many layers can lead to overtraining, so balancing the number of layers is necessary. The size of the convolution kernel is also crucial for extracting certain features, but too large a size can increase the number of parameters and lead to overfitting. The pooling size also affects the preservation of important information and needs to be adjusted. The number of fully connected layers affects the network's ability to classify gestures, but too many layers can also lead to overtraining. The number of neurons in thoroughly combined layers is also essential for classification and needs to be adjusted for optimal network performance. Considering the above, the CNN architecture was successfully selected, as shown in Figure 4.

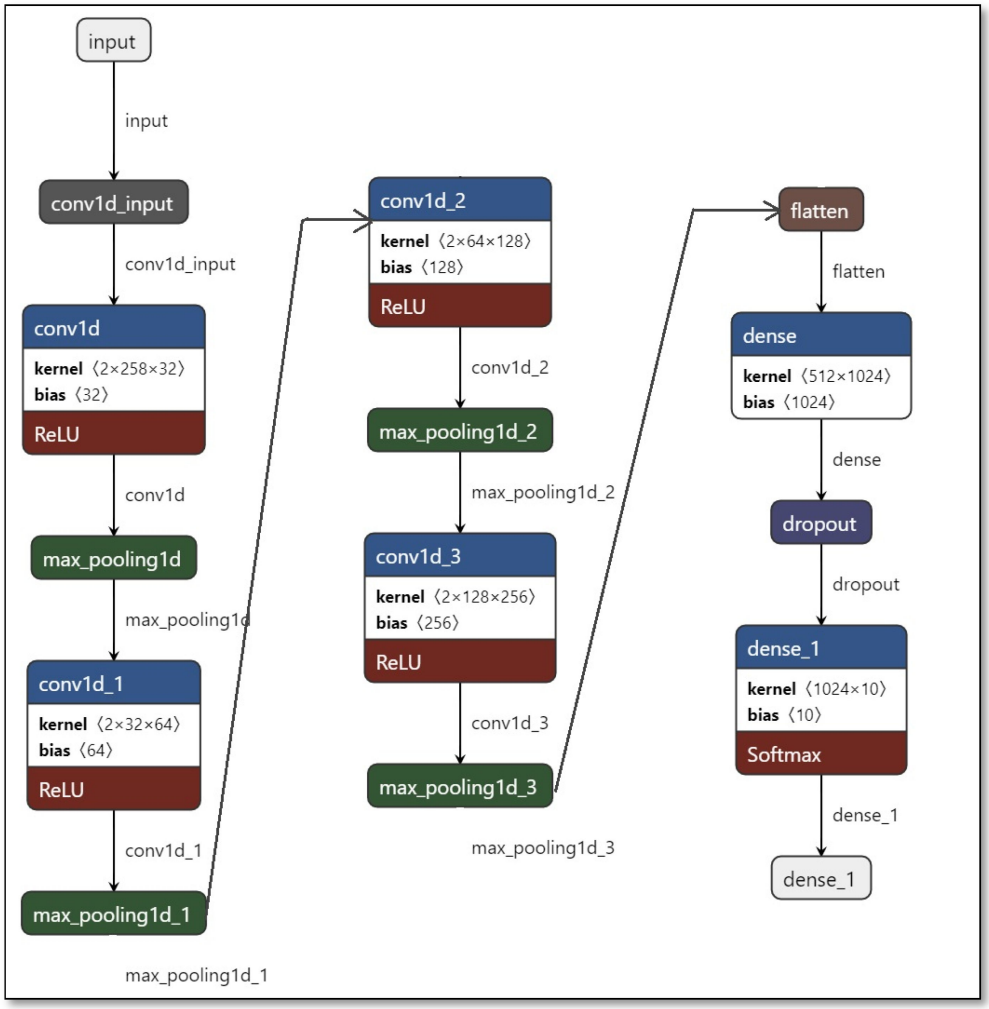

Figure 4. The proposed model's detailed architecture.

The model architecture shown in Figure 4 is a sequential neural network consisting of convolutional layers, max pooling layers, a flatten layer, and two dense layers. The input to the model is a sequence of data points, and the output is a prediction for the sign language gesture being performed.

The convolutional layers use filters to extract features from the input sequence, and the max pooling layers reduce the spatial size of the features while retaining the most important information. The flatten layer converts the 2D feature maps to a 1D feature vector, and the dense layers perform classification on the features.

The first dense layer has 1024 units and uses a dropout layer to prevent overfitting. The final dense layer has 40 units, which corresponds to the number of sign language gestures being recognized by the model. The total number of parameters in the model is 669,320, all of which are trainable.

The categorical cross-entropy loss function measures the distance between the actual and predicted probabilities of classes. For example, in the case of a classification problem with 10 categories, for each data element the model will output a probability vector of length 10, where each piece corresponds to the probability that the component belongs to a certain class. The prediction vector is compared with the actual class labels, represented as a one-ot-encoding vector. The loss function is minimized by optimizing the model parameters.

By plotting the training and validation accuracy ('train_accuracy', 'val_accuracy'), the model's performance can be tracked during training to determine whether the model is over-trained or under-trained. If the accuracy of training increases but verification does not increase, this may indicate that the model is being rebuilt. Plotting training and validation losses ('loss', 'val_loss') allows one to track the model's performance during training and determine whether the model is overly or insufficiently adapted. If training losses decrease and validation losses increase, this may indicate that the model is being rebuilt (Figure 5). If both losses are high, this may mean that the model is not sufficiently adapted and it needs to be trained for more epochs or with other parameters.

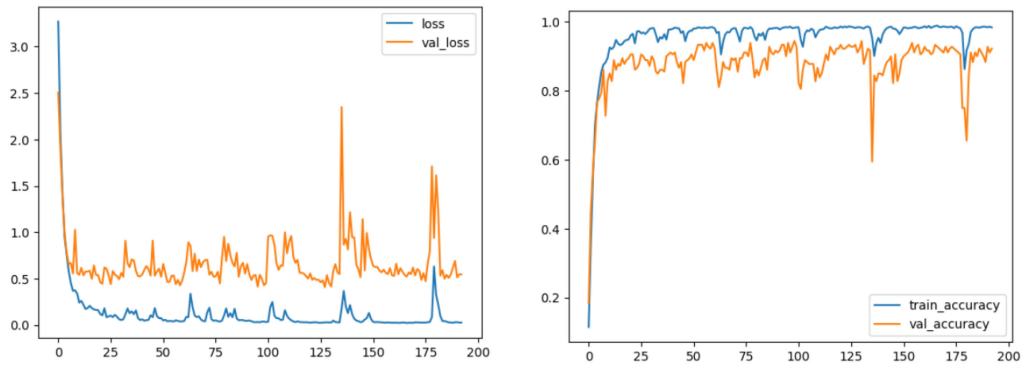

Figure 5. Model training and validation losses (**left**) and accuracy (**right**) of all groups.

Both metrics are essential for evaluating the performance of a machine learning model. Still, factual accuracy is more often used in multiclass classification tasks, and training accuracy is more often used to assess the performance of a model in the learning process.

The model was successfully trained with a 'train_accuracy' of 0.98 on a training sample consisting of a data array of 30,960,000 elements containing 40 dynamic gestures, which consists of 2000 records (each gesture is recorded in 50 instances, which consists of 60 frames), which are characterized by 258 parameters (pose coordinates: 33 x, y, z values and 33 visibility values, total $33 \times 4 = 132$, hand coordinates: 22 x, y, z values for both hands, total $22 \times 2 \times 3 = 126$). To verify the quality of the model and its ability to generalize to new data, its performance was evaluated on validation samples with an accuracy of 0.96 and test samples with an accuracy of 0.96.

3.4. Model Demployment Module

Deployment begins by initializing several variables, including an empty list to store a sequence of key points, a count of the number of frames processed, a list to store the recognized sentence, and a threshold value to screen out unlikely predictions.

The function then opens the video capture device (preprocessing or video recording) and runs a loop to read frames from the device. For each frame, the function uses the holistic model to detect and track hand landmarks and other body landmarks, then extracts the hand landmarks to use as input for the pre-trained model (2D CNN). If the hand landmarks are detected first, the function resets the sequence and frame counter.

Once the sequence reaches the desired length, the function feeds it into the pre-trained model to predict the gesture. If the prediction reliability is above the threshold value, the recognized gesture is added to the sentence. The function then draws the predicted gesture and sentence on a frame and displays them in a window (Figure 6).

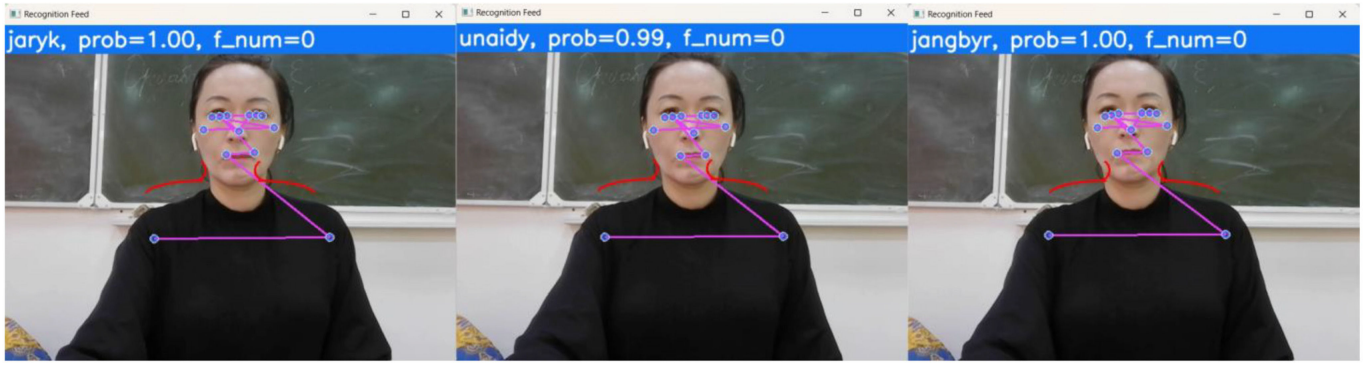

Figure 6. Dynamic gesture variability recognition system. Publication with the consent of the people shown in the figure.

4. Results

To achieve high accuracy in the classification of dynamic gestures, it is better to use a sufficiently large and diverse data set (which will cover various forms of demonstration as much as possible) for training and a complex model (Figure 7).

The first group of gestures (Figure 7)—where the palm orientation is directed towards the camera or the demonstrator, the connection points involved in the gesture are as readable as possible, and the trajectory of movement is parallel to the camera or static—showed the highest result for the training sample. The value of 'train_accuracy' was 0.98 for the test sample, the value of 'test_accuracy' was 0.98, and for the validation sample, the values of 'val_accuracy' was 0.96.

The second group of gestures (Figure 7) is where the orientation of the palm is parallel to the camera and the movement of the hand is perpendicular to the camera. This means that the hand is moving away from or towards the camera, and the palm is facing the camera. Since the movement of the hand is not parallel to the camera, it may be more challenging to capture the movement accurately. To accurately capture action, it may be necessary to use techniques such as motion capture or computer vision algorithms to track the position and direction of the hand in three-dimensional space. This may include using multiple cameras to capture the movement at different angles and triangulating the hand's position based on the images captured by each camera. But despite this, because the orientation of the palm is directed towards the camera, our model showed that in the result for the training sample, the value of 'train_accuracy' was 0.99; for the test sample, the value of 'test_accuracy' was 0.98, and for the validation sample, the value of 'val_accuracy' was 0.96.

In the third group of gestures (Figure 7), in which the orientation of the palm is not directed at the camera or the demonstrator, the connection points are not all readable, and the movement is parallel relative to the camera. This form of demonstration presents a challenge for accurately capturing and interpreting gestural movements. Since the orientation of the palm is not directed at the camera or the demonstrator, it can be challenging to see the exact position and configuration of the fingers. The model we proposed also successfully showed that the value of 'train_accuracy' was 0.98, for the test sample, the

value of 'test_accuracy' was 0.97, and for the validation sample, the value of 'val_accuracy' was 0.94.

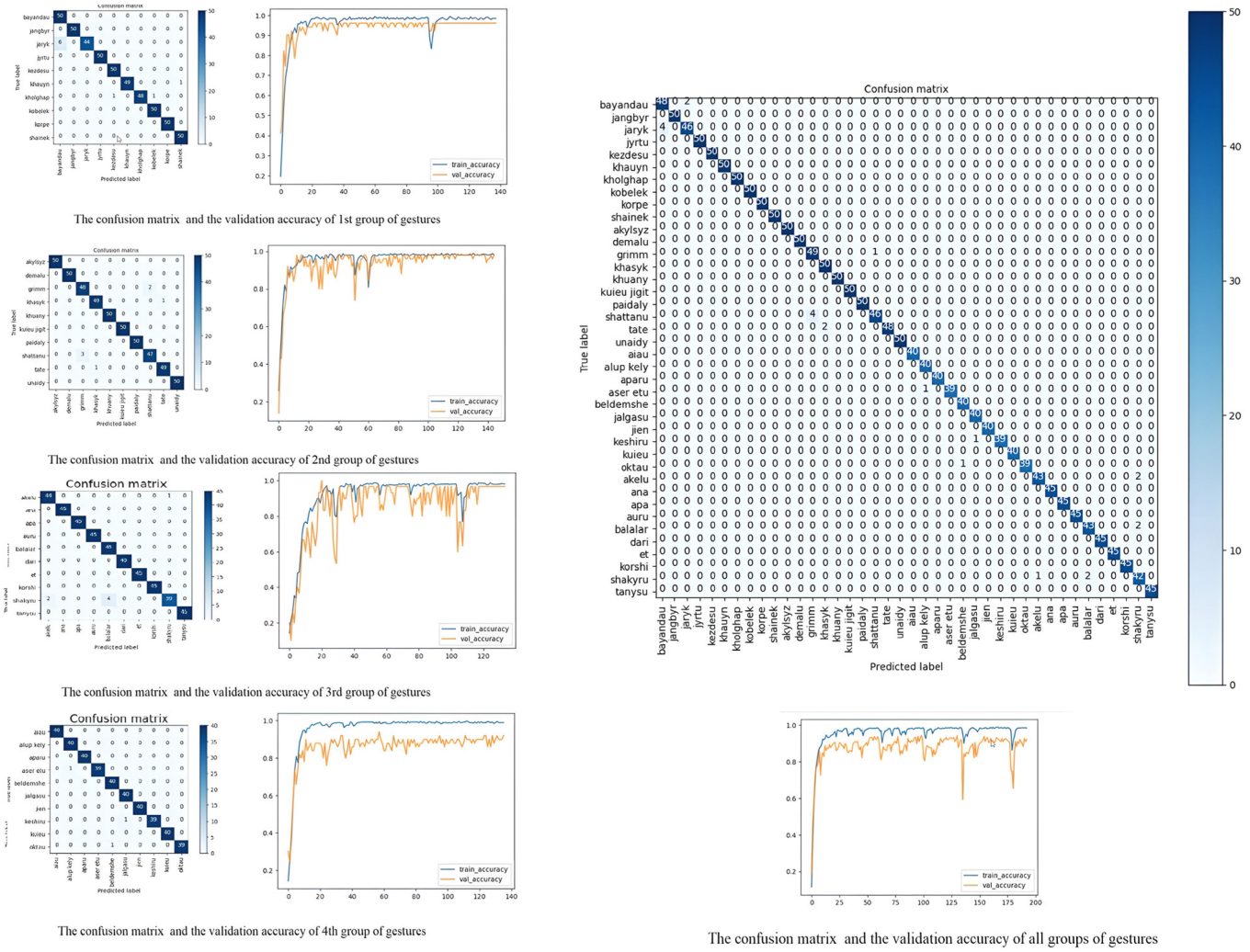

Figure 7. Confusion matrix of the dynamic gesture variability recognition model for all test groups (see Table A1 for Kazakh's words translations).

When the orientation of the palm and movement is perpendicular to the camera, it means that the hand is moving in the direction either towards the camera or away from it. Many gestures (Figure 6) belong to this category, but for the experiment, 10 gesture words were selected with these demonstration properties. These groups of sign words in the fourth group showed minor indicators for the validation sample, with the value of 'val_accuracy' being 0.92. For the test sample, the value of 'test_accuracy' was 0.97, and for the validation sample, the value of 'train_accuracy' was 0.98.

Based on the presented results (Figure 7), it can be concluded that the model that was used to recognize different groups of gestures has an average validation accuracy of 0.92. This means that, on average, the model is able to accurately classify gestures with a high level of confidence. However, it should be noted that some groups of gestures performed better than others, with some showing significantly higher accuracy than the average. The orientation of the palm and the trajectory of the movement were identified as important factors that can impact the accuracy of the model. Overall, the results suggest that the model has the potential to be a useful tool for gesture recognition, but further research and refinement may be needed to improve its performance on more challenging gestures.

The validation accuracy 0.92 is an important metric, as it provides an estimate of how well the model is generalizing to new, unseen data. A high validation accuracy indicates that the model is performing well and can be trusted to make accurate predictions for new data.

The proposed method for sign language recognition was tested on three different datasets (Table 1, and Figure 8): the Ankara University Turkish Sign Language Dataset (AUTSL), the Dataset for Argentinian Sign Language (LSA64), and Kazakh Sign Language (KSL).

Table 1. Comparison of the authors' method with other work (only RGB) on a test set of the AUTSL, LSA64, and KSL datasets.

Datasets	Precision	Recall	F1 Score	Accuracy
AUTSL	0.93	0.93	0.93	0.93
LSA64	1	1	1	1
KSL	0.98	0.98	0.98	0.98

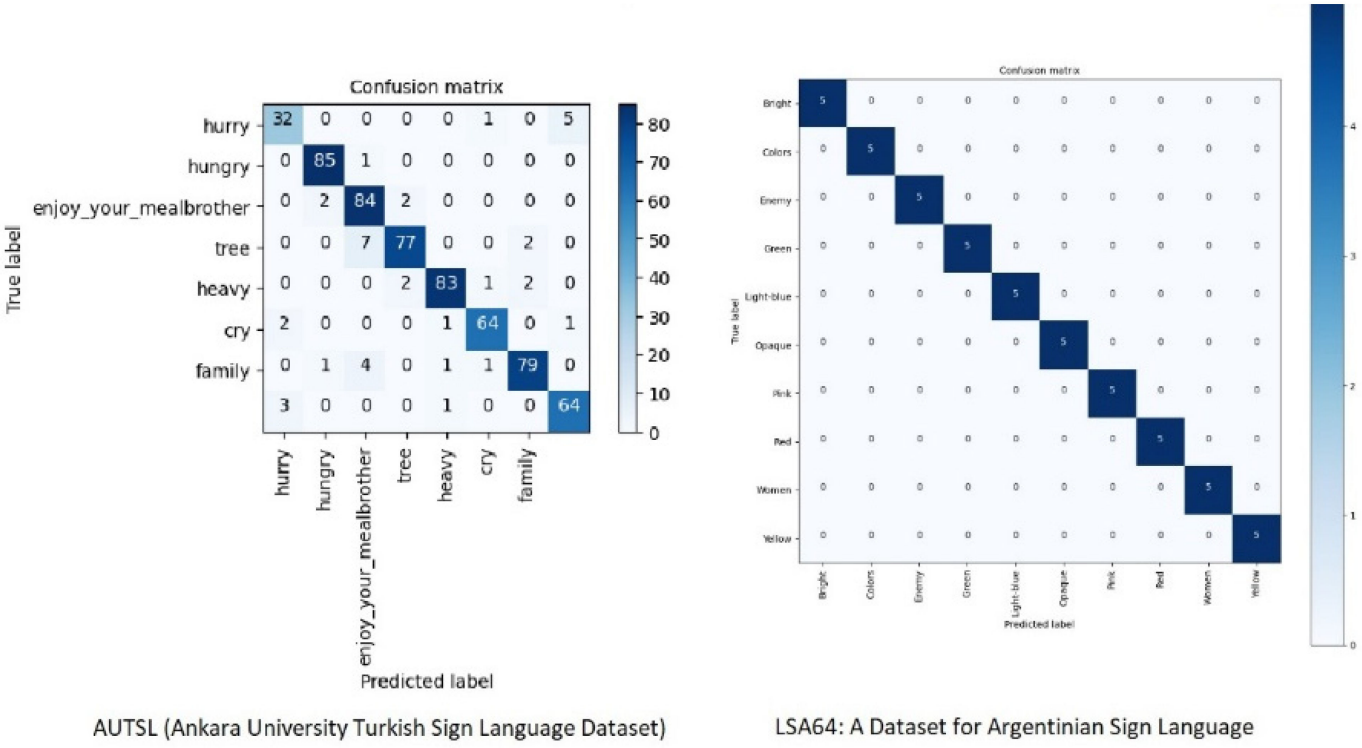

Figure 8. Confusion matrix of the AUTSL and LSA64 datasets (see Table A1 for Kazakh's words translations).

Table 1 presents a comparison of the performance of the proposed method for sign language recognition with other works based on three different datasets: AUTSL, LSA64, and the author's dataset: KSL.

The results show that the proposed method has a higher precision, recall, F1 score, and accuracy than the other works for the KSL dataset. The precision, recall, F1 score, and accuracy values were all 0.98, indicating that the proposed method is highly accurate in recognizing signs in this dataset. For the AUTSL dataset, the precision, recall, F1 score, and accuracy values were all 0.93. This indicates that the proposed method performs well in recognizing signs in this dataset, although not as well as in the KSL dataset. For the LSA64 dataset, the precision, recall, F1 score, and accuracy values were all 1. This indicates that the proposed method performs perfectly in recognizing signs in this dataset (Figure 8).

Overall, the results suggest that the proposed method is effective in recognizing signs in multiple datasets and has a higher accuracy than other works in recognizing the signs in the KSL dataset.

Table 2 presents a comparison of the recognition results achieved using different gesture language recognition methods on the AUTSL dataset, including hybrid and ensemble architectures. The proposed method in the study achieved an accuracy of 0.98, which was slightly lower than the results achieved by other state-of-the-art methods, such as STF+LSTM and 3D CNN, CGN, and RGB-D, which achieved accuracies of 0.9856 and 0.9853, respectively.

Table 2. Comparison of the authors' method with other work (only RGB) on a test set from the AUTSL dataset.

Model	Accuracy
STF+LSTM [31]	0.9856
3D CNN, CGN, RGB-D [32]	0.9853
authors' method	**0.98**
ResNet. Transformer [33]	0.9292
CNN+FPM+BLSTM+Attention (RGB-D) [26]	0.6203

It is noteworthy that the proposed approach does not rely on complex hybrid architectures or ensemble models that demand significant computational resources. Despite this, the method achieves a reasonable level of recognition accuracy without the need for supplementary features such as text or lip data, or additional equipment like depth-sensing cameras or gloves. These findings suggest that gesture recognition can be accomplished using simpler architectures and only RGB images, indicating promising possibilities for practical applications.

The comparison table highlights the potential of more complex architectures and depth cameras for achieving higher accuracy in gesture recognition, as shown by some of the state-of-the-art methods presented. However, the proposed method is still effective in achieving acceptable recognition accuracy without using these more complex techniques, which suggests that it may be a useful and practical solution in certain contexts.

5. Discussion

The results presented in the text show that the proposed model for gesture recognition is effective for recognizing different groups of gestures with a high degree of accuracy. The analysis of the different groups of gestures showed that certain factors, such as the orientation of the palm and the trajectory of movement, play an important role in the accuracy of the model.

The high validation accuracy of 0.93 indicates that the model performs well and can be trusted to make accurate predictions on new data. This is an important metric for any machine learning model, as it shows how well it generalizes to new, unseen data. The results also suggest that the proposed model has the potential to be a useful tool for gesture recognition in sign language and can be applied to multiple sign languages, indicating its multilingualism.

Furthermore, the experiment conducted in the study was compared to other works on a test set from the AUTSL and LSA64 datasets, and the results showed that the proposed method had a higher precision, recall, F1 score, and accuracy.

This further supports the effectiveness of the proposed method for gesture recognition in sign language. Further development of the issues discussed in the article may include the refinement of the proposed method and experimental setup to improve its performance and accuracy. In addition, the system's applicability in real-world scenarios and its ability to work with various sign languages and dialects should also be studied.

6. Conclusions

The answer to all the formulized research questions above is: YES.

The proposed approach in this research combines several modalities, including hand movement, hand configuration, and body and head movements, to improve the accuracy of sign language recognition. The integration of information from different modalities allows the system to better understand the nuances of sign language and accurately translate it into written speech. This is a significant innovation in the field of sign language recognition and translation, as it improves the accuracy and efficiency of the system and makes it more accessible to a wide range of users.

Secondly, the authors have attempted to capture the maximum possible variations in the demonstration of gestures by grouping them into four different groups. This allows for a more comprehensive understanding of sign language and improves the accuracy of the recognition system.

Thirdly, the authors have developed a preprocessing module that can automatically collect gesture properties from a video file or in real-time, making the methodology universal for any sign language and open for further experiments.

In addition to proposing a new approach for automatic sign language translation, the authors have also made significant contributions to the recognition and preservation of Kazakh Sign Language (KSL). As KSL is a relatively new and less-studied sign language, its recognition as a separate sign language and its preservation is important. The authors have developed their own KSL dataset, which is a significant contribution to the field, as it provides ample training data for the proposed approach and can be used for further research and experimentation. This can help to improve the recognition and preservation of KSL, as well as other sign languages facing similar situations.

Overall, the proposed approach is effective and innovative, with several key contributions to the field of sign language recognition and translation. The authors have addressed several research questions and proved the effectiveness of the architecture by testing it on open datasets. These findings have the potential to make automatic sign language translation more accessible and efficient for a wide range of users.

7. Future Work

Firstly, by expanding the system we have proposed, it is possible to develop a continuous SL recognition system, which can then produce correctly translated sentences based on the semantic parsing of words. Secondly, the use of the single functionality (video recording system, processing of finished videos) offered in this system can become a prerequisite for the development of automatic SL translation systems for any sign language.

Author Contributions: Conceptualization, N.A. and S.K.; Methodology, M.M., S.K., A.K. and G.K.; Validation, G.K.; Investigation, N.A., M.M., S.K. and L.Z.; Data curation, N.A., A.K. and L.Z.; Writing—original draft, N.A. and M.M.; Visualization, A.K.; Supervision, N.A. and M.M.; Funding acquisition, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by the Lublin University of Technology Scientific Fund FD-20/IT-3/007.

Informed Consent Statement: Informed consent was obtained from all the subjects involved in this study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. List of Words Used in This Research

Table A1. List of Kazakh sign words and their translations to English.

No.	Kazakh Sign Word in Latin	Kazakh Sign Word in Cyrillic	Translation
1.	bayandau	Баяндау	Narrative
2.	jangbyr	Жаңбыр	Rain
3.	jaryk	Жарық	Light
4.	jyrtu	Жырту	Plow
5.	kezdesu	Кездесу	The meeting
6.	khauyn	Қауын	Melon
7.	kholghap	Қолғап	Gloves
8.	kobelek	Көбелек	Butterfly
9.	korpe	Көрпе	Blanket
10.	shainek	Шәйнек	Kettle
11.	akylsyz	Ақылсыз	Crazy
12.	demalu	Демалу	Rest
13.	grimm	Гримм	Grimm
14.	khasyk	Қасық	Spoon
15.	khuany	Қуану	Rejoice
16.	kuieu jigit	Күйеу жігіт	The groom
17.	paidaly	Пайдалы	Useful
18.	shattanu	Шаттану	Delight
19.	tate	Тәте	Aunt
20.	unaidy	Ұнайды	Like
21.	aiau	Аяу	Pity
22.	alup kely	Алып келу	Bring
23.	aparu	Апару	Drag
24.	aser etu	Әсер ету	Influence
25.	beldemshe	Белдемше	Skirt
26.	jalgasu	Жалғасу	Continuation
27.	jien	Жиен	Nephew
28.	keshiru	Кешіру	Forgive
29.	kuieu	Күйеу	Husband
30.	oktau	Оқтау	Loading
31.	akelu	Әкелу	Bring
32.	ana	Ана	Mother
33.	apa	Апа	Sister
34.	auru	Ауру	Disease
35.	balalar	Балалар	Children
36.	dari	Дәрі	Medicine
37.	et	Ет	Meat
38.	korshi	Көрші	Neighbor
39.	shakyru	Шақыру	The invitation
40.	tanysu	Танысу	Dating

References

1. Abdullahi, S.B.; Chamnongthai, K. American Sign Language Words Recognition Using Spatio-Temporal Prosodic and Angle Features: A Sequential Learning Approach. *IEEE Access* **2022**, *10*, 15911–15923. [CrossRef]
2. Sincan, O.M.; Keles, H.Y. Using Motion History Images With 3D Convolutional Networks in Isolated Sign Language Recognition. *IEEE Access* **2022**, *10*, 18608–18618. [CrossRef]
3. Deafness and Hearing Loss. Available online: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (accessed on 22 July 2023).
4. Sahoo, J.P.; Prakash, A.J.; Pławiak, P.; Samantray, S. Real-Time Hand Gesture Recognition Using Fine-Tuned Convolutional Neural Network. *Sensors* **2022**, *22*, 706. [CrossRef]
5. Bird, J.J.; Ekárt, A.; Faria, D.R. British Sign Language Recognition via Late Fusion of Computer Vision and Leap Motion with Transfer Learning to American Sign Language. *Sensors* **2020**, *20*, 5151. [CrossRef]
6. Amangeldy, N.; Kudubayeva, S.; Razakhova, B.; Mukanova, A.; Tursynova, N. Comparative analysis of classification methods of the dactyl alphabet of the Kazakh language. *J. Theor. Appl. Inf. Technol.* **2022**, *100*, 5506–5523. Available online: http://www.jatit.org/volumes/Vol100No19/9Vol100No19.pdf (accessed on 18 April 2023).
7. Amangeldy, N.; Kudubayeva, S.; Kassymova, A.; Karipzhanova, A.; Razakhova, B.; Kuralov, S. Sign Language Recognition Method Based on Palm Definition Model and Multiple Classification. *Sensors* **2022**, *22*, 6621. [CrossRef] [PubMed]
8. Thejowahyono, N.F.; Setiawan, M.V.; Handoyo, S.B.; Rangkuti, A.H. Hand Gesture Recognition as Signal for Help using Deep Neural Network. *Int. J. Emerg. Technol. Adv. Eng.* **2022**, *12*, 37–47. [CrossRef] [PubMed]
9. Sosa-Jimenez, C.O.; Rios-Figueroa, H.V.; Rechy-Ramirez, E.J.; Marin-Hernandez, A.; Gonzalez-Cosio, A.L.S. Real-time Mexican Sign Language recognition. In Proceedings of the 2017 IEEE International Autumn Meeting on Power, Electronics and Computing, ROPEC 2017, Ixtapa, Mexico, 8–10 November 2017; pp. 1–6. [CrossRef]
10. Dayal, A.; Paluru, N.; Cenkeramaddi, L.R.; Soumya, J.; Yalavarthy, P.K. Design and Implementation of Deep Learning Based Contactless Authentication System Using Hand Gestures. *Electronics* **2021**, *10*, 182. [CrossRef]
11. Karacı, A.; Akyol, K.; Turut, M.U. Real-Time Turkish Sign Language Recognition Using Cascade Voting Approach with Handcrafted Features. *Appl. Comput. Syst.* **2021**, *26*, 12–21. [CrossRef]
12. Tateno, S.; Liu, H.; Ou, J. Development of Sign Language Motion Recognition System for Hearing-Impaired People Using Electromyography Signal. *Sensors* **2020**, *20*, 5807. [CrossRef] [PubMed]
13. Zhou, Z.; Tam, V.W.L.; Lam, E.Y. A Portable Sign Language Collection and Translation Platform with Smart Watches Using a BLSTM-Based Multi-Feature Framework. *Micromachines* **2022**, *13*, 333. [CrossRef] [PubMed]
14. Zhou, Z.; Tam, V.W.L.; Lam, E.Y. SignBERT: A BERT-Based Deep Learning Framework for Continuous Sign Language Recognition. *IEEE Access* **2021**, *9*, 161669–161682. [CrossRef]
15. Kapuscinski, T.; Wysocki, M. Recognition of Signed Expressions in an Experimental System Supporting Deaf Clients in the City Office. *Sensors* **2020**, *20*, 2190. [CrossRef]
16. Vaitkevičius, A.; Taroza, M.; Blažauskas, T.; Damaševičius, R.; Maskeliūnas, R.; Woźniak, M. Recognition of American Sign Language Gestures in a Virtual Reality Using Leap Motion. *Appl. Sci.* **2019**, *9*, 445. [CrossRef]
17. Du, Y.; Dang, N.; Wilkerson, R.; Pathak, P.; Rangwala, H.; Kosecka, J. American Sign Language Recognition Using an FM-CWWireless Sensor. In Proceedings of the AAAI 2020—34th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
18. Papastratis, I.; Dimitropoulos, K.; Daras, P. Continuous Sign Language Recognition through a Context-Aware Generative Adversarial Network. *Sensors* **2021**, *21*, 2437. [CrossRef]
19. Papastratis, I.; Dimitropoulos, K.; Konstantinidis, D.; Daras, P. Continuous Sign Language Recognition Through Cross-Modal Alignment of Video and Text Embeddings in a Joint-Latent Space. *IEEE Access* **2020**, *8*, 91170–91180. [CrossRef]
20. Zhou, H.; Zhou, W.; Zhou, Y.; Li, H. Spatial-Temporal Multi-Cue Network for Continuous Sign Language Recognition. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, 7–12 February 2020. [CrossRef]
21. Ahmed, M.; Zaidan, B.; Zaidan, A.; Salih, M.M.; Al-Qaysi, Z.; Alamoodi, A. Based on wearable sensory device in 3D-printed humanoid: A new real-time sign language recognition system. *Measurement* **2021**, *168*, 108431. [CrossRef]
22. Alrubayi, A.; Ahmed, M.; Zaidan, A.; Albahri, A.; Zaidan, B.; Albahri, O.; Alamoodi, A.; Alazab, M. A pattern recognition model for static gestures in malaysian sign language based on machine learning techniques. *Comput. Electr. Eng.* **2021**, *95*, 107383. [CrossRef]
23. Al-Samarraay, M.S.; Zaidan, A.; Albahri, O.; Pamucar, D.; AlSattar, H.; Alamoodi, A.; Zaidan, B.; Albahri, A. Extension of interval-valued Pythagorean FDOSM for evaluating and benchmarking real-time SLRSs based on multidimensional criteria of hand gesture recognition and sensor glove perspectives. *Appl. Soft Comput.* **2021**, *116*, 108284. [CrossRef]
24. Ahmed, M.A.; Zaidan, B.B.; Zaidan, A.A.; Alamoodi, A.H.; Albahri, O.S.; Al-Qaysi, Z.T.; Albahri, A.S.; Salih, M.M. Real-time sign language framework based on wearable device: Analysis of MSL, DataGlove, and gesture recognition. *Soft Comput.* **2021**, *25*, 11101–11122. [CrossRef]
25. Tornay, S.; Razavi, M.; Magimai-Doss, M. Towards multilingual sign language recognition. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 6309–6313. [CrossRef]

26. Sincan, O.M.; Keles, H.Y. AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods. *IEEE Access* **2020**, *8*, 181340–181355. [CrossRef]
27. Ronchetti, F.; Quiroga, F.; Estrebou, C.; Lanzarini, L.; Rosete, A. LSA64: A Dataset for Argentinian Sign Language. In Proceedings of the XXII Congreso Argentino de Ciencias de la Computación, CACIC 2016, San Luis, Argentina, 3–7 October 2016; pp. 794–803.
28. Ryumin, D.; Kagirov, I.; Axyonov, A.; Pavlyuk, N.; Saveliev, A.; Kipyatkova, I.; Zelezny, M.; Mporas, I.; Karpov, A. A Multimodal User Interface for an Assistive Robotic Shopping Cart. *Electronics* **2020**, *9*, 2093. [CrossRef]
29. Hand Landmarks Detection Guide. Available online: https://developers.google.com/mediapipe/solutions/vision/hand_landmarker (accessed on 8 January 2023).
30. MediaPipe Holistic. Available online: https://google.github.io/mediapipe/solutions/holistic (accessed on 8 January 2023).
31. Ryumin, D.; Ivanko, D.; Ryumina, E. Audio-Visual Speech and Gesture Recognition by Sensors of Mobile Devices. *Sensors* **2023**, *23*, 2284. [CrossRef]
32. Jiang, S.; Sun, B.; Wang, L.; Bai, Y.; Li, K.; Fu, Y. Skeleton Aware Multi-modal Sign Language Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 19–25 June 2012; pp. 3408–3418. [CrossRef]
33. De Coster, M.; Van Herreweghe, M.; Dambre, J. Isolated Sign Recognition from RGB Video using Pose Flow and Self-Attention. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 19–25 June 2021; pp. 3436–3445. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.