Article

Multi-Stream Graph-based Deep Neural Networks for Skeleton-based Sign Language Recognition

Abu Saleh Musa Miah , Md. Al Mehedi Hasan , Si-Woong Jang , Hyoun-Sup Lee * , Jungpil Shin *

Posted Date: 8 May 2023

doi: 10.20944/preprints202305.0467.v1

Keywords: Sign Language Recognition (SLR); Large Scale Dataset; American Sign Language; Turkey Sign Language; Chinese Sign Language; AUTSL; CSL

Article

Multi-Stream Graph-Based Deep Neural Networks for Skeleton-Based Sign Language Recognition

Abu Saleh Musa Miah [1], Md. Al Mehedi Hasan [2], Si-Woong Jang [3], Hyoun-Sup Lee [4,*] and Jungpil Shin [1,*],

[1] School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Fukushima 965-8580, Japan; jpshin@u-aizu.ac.jp; d8231105@u-aizu.ac.jp
[2] Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology(RUET), Rajshahi, Bangladesh; mehedi_ru@yahoo.com
[3] Department of Computer Engineering, Dongeui University, Busanjin-Gu, Busan, 47340, Korea; swjang@deu.ac.kr
[4] Department of Applied Software Engineering, Dongeui University, Busanjin-Gu, Busan, 47340, Korea;lhskmj@deu.ac.kr
* Correspondence:Hyoun-Sup Lee (lhskmj@deu.ac.kr), Jungpil Shinjp (jpshin@u-aizu.ac.jp)

Abstract: Sign Language Recognition (SLR) aims to bridge speech-impaired and general communities by recognizing signs from given videos. Researchers still face challenges developing efficient SLR systems because of the video's complex background, light illumination, and subject structures. Recently many researchers developed a skeleton-based sign language recognition system to overcome the subject and background variation of hand gesture signs. However, skeleton-based SLR is still under exploration due to the lack of information and annotations on hand key points. More recently, researchers included body and face information with the hand gesture for the SLR, but their performance and efficiency are unsatisfactory. We proposed a Multi-Stream Graph-based Deep Neural Network (SL-GDN) for a skeleton-based SLR system to overcome the problems. The main purpose of the proposed SL-GDN approach is to improve the efficiency and performance of the SLR system with a low computational cost based on the human body pose in the form of 2D landmark locations. In the procedure, firstly, we constructed a skeleton graph based on the selected 27 whole-body key points among 67 key points to solve the inefficiency problems. Then we proposed multi-stream SL-GDN to extract features from the whole-body skeleton graph for four streams. Finally, we concatenated the four different features and applied a classification module to refine the feature and recognize corresponding sign classes. Our data-driven and graph construction method increases the system's flexibility and brings high generability to adapt various data samples. We used three large-scale benchmark SLR datasets to evaluate the proposed model: WLASL, AUTSL and CSL. The demonstrated performance accuracy table proved the superiority of the proposed model, and we believe this will be considered a great invention in the SLR domain.

Keywords: sign language recognition (SLR); large scale dataset; American sign language; Turkey sign language; Chinese sign language; AUTSL; CSL

1. Introduction

Sign language is a spatial kind of visual language based on dynamic gesture movement including hand, body and facial gesture expression [1–7]. This is the language for those community who does not speak or hear anything spatially for the deaf and speech-impaired people. Because of the difficulties and various complexity of sign language, such as considerable time for understanding and utilizing general people are not eager to learn this language to establish communication with those specialized disabled people. In addition, teaching this language to the general people to communicate with the minor community is not practical and feasible. Moreover, there are no international common versions of this langue but it is affected by people's several languages such as Bangla [3], Turkey [8], Chinese [9], English [10], and culture [1,4,11]. To establish effective communication between the general people and the deaf community it is needed a language translator but it is rare to find expert sign, language interpreters. Researchers think automatic sign language recognition(SLR) can only solve this problem [3–6]. Researchers have been working to develop an SLR system with the help of computer vision [3,4,12] sensor-based methods [5,6,12–15] and artificial intelligence to ease communication way for deaf and hearing impaired people. Many researchers performed well with conventional hand gestures and action recognition [1–3,8]. However, there have major differences between the conventional hand gesture and SLR where the main difference is that SLR is more challenging to recognise. The main difficulty in the SLR is that it requires delicate and global hand gestures to carry the specific meaning and express the emotion of the person. In addition, right-hander, left-hander, body shape, localism, and speed can change the meaning of the gesture sign. Similar signs can produce different meanings because of their complexity. Collecting a large number of data from many signers can be solved this problem, but it will be highly expensive. Many researchers used various handcraft feature extraction methods with machine learning to classify sign language images. Most of them used SIFT [16] and HOG [17] hand-crafted feature extraction and then employed SVM and KNN to classify them [9,18,19]. Many researchers employed segmentation and semantic detection-based model to recognize the SLR by following two stages [20–22]. The main drawback of this method is that they may face difficulties in producing a good performance for the video or large-scale dataset. Due to the updated device and storage capacity most of the researchers are interested to work with time series and general large video datasets. The deep learning-based approach achieved significant improvement in overcoming the large-scale time series and general video datasets. Researchers focus on the deep learning-based model and applied RNN, LSTM and 3D CNNS to recognize pixel-based SLR [23–26]. Also, this method produced a good performance, but the spatial and temporal features combination is more efficient than this. To achieve this recently, many researchers employed an attention-based model to recognize actions and hand gestures [1,2]. To improve the performance some researchers combine the attention module with the local motion information [27,28]. Also, image-based sign language recognition can be generated good performance but still faces difficulties because of the computational complexity, complex background, light illumination, and partial occlusions. To overcome these challenges more recently, many researchers proposed a skeleton-based SLR system that mainly uses specific skeleton points instead of the pixels of the images [29–33]. The main advantage of the skeleton-based SLR system is that it can increase attention and has strong adaptability to complicated backgrounds and dynamic circumstances. However, there are still some deficiencies in extracting the skeleton points for the SLR because of the inefficiency of the ground truth skeleton annotation. Also, many motion capture systems such as Microsoft kinetic, Microsoft Oak-D, Intel RealS sense and other systems only provide the main body coordinates and their skeleton annotation but it is difficult to get the skeleton annotation for the gestures [34]. Shin et al. extracted the 21 hand key points with the mediapipe system from the American sign language dataset. After extracting the distance and angular features, they apply SVM for recognition [35]. The only hand skeleton information is sometimes insufficient to correctly carry the exact meaning of the sign because of the lackings of emotion and expression of the body. More recently, researchers have thought that a full-body skeleton is more efficient for SLR systems [36]. Xia et al. extracted the hand skeleton and body skeleton with different approaches, and they achieved good performance with the RNN-based model [37]. This work's main problem is the unreliable hand key points, and RNN can not produce a performance for the dynamics skeletons. Perez et al. extracted 67 key points, including the face, body and hand gestures, using a special camera, and finally, they achieved good performance with the LSTM [38]. Jiang et al. applied a different approach with a multimodal dataset, including full-body skeleton points and achieved a good performance accuracy [8]. They also think of reducing the number of skeletons to increase the model's efficiency. The main problem is that their method did not seem able to achieve good performance and the generalization property for the SLR compared to the existing systems. In addition, researchers focused on the skeleton-based SLR because of the high complexity of the pixel-based system. With the full-body skeleton, we face almost the same computational complexity problems. We proposed Multi-Stream Graph-based Deep Neural Networks (SL-GDN) to recognize sign language using potential skeleton points of whole body information to overcome the challenges. In the study, we designed a new skeleton graph for SLR, which included the spatial and temporal features using the graph and neural network to model the dynamics embedded.

Our major contributions to the work are given below:

- We constructed a skeleton graph for Large Scale SLR with selected 27 key points among the whole body key points. The main purpose of this graph is to construct a unified graph to dynamically optimize the nodes and edges based on the different actions. Due to the minimum number of the skeleton key points being selected among the whole body, the computational complexity can be solved to increase the model's efficiency.
- We extracted the hybrid feature by combining the Graph-based SL-GDN and Generale Neural network features from the multiple streams. After concatenating the feature, we used a classification module to refine the concatenated feature and prediction.
- To evaluate the model, we used three large-scale datasets with four modalities: joint, joint, bone and bone. Our model generated a high performance compared to the existing system.

The presented work is organized as follows: Section 2 summarizes the existing research work and problems related to the presented work, Section 3 describes the benchmark and proposed Korean sign language datasets, and Section 4 describes the architecture of the proposed system—Section 5 evaluation performance. In Section 6, draw the conclusion and future work.

2. Related Work

Sign Language Recognition (SLR) has achieved significant progress using various deep learning-based models which achieved good performance with the allowable computational power [1–4,21,22,26,39–43]. The existing SLR systems are still facing many difficulties in achieving good performance because of the inefficiency of the potential information, consider-able gesture for SLR and potential features. One of the common challenges is to capture global body motion skeleton and local arm, hand, and facial expressions simultaneously. Neverova et al. employed a ModDrop framework for initializing individual and gradual fusion modalities for capturing spatial information [39]. They achieved good performance for spatial and temporal information for multiple modalities. One of the drawbacks of their ideas is that they applied an augmented with audio which is not good for all time. Pu et al. employed connectionist temporal classification (CTC) for sequence modelling and a 3D convolutional residual network (3D-ResNet) for feature learning [26]. The employed LSTM and CTC decoder with jointly trained by a soft Dynamic Time Warping (soft-DTW) alignment constraint. Finally, they employed 3D-ResNet for training labels with lass and validated with RWTHPHOENIX-Weather and CSL datasets with 36.7% and 32.7-word error rate (WER) sequentially. Koller et al. employed a hybrid CNN-HMM model for combining the two kinds of features, such as the discriminative features of the CNN with the sequence features of Hid-den-Markov-Models (HMMs) [21]. They claimed they achieved good recognition accuracy for the three benchmark sign language datasets, which reduced 20% WER. Huang et al. proposed an attention-based 3D-convolutional neural net-works (3D-CNNs) for SLR aiming to extract the spatial-temporal feature and selected highlighted information with an attention mechanism [27]. Finally, they evaluated their model with the CSL and ChaLearn 14 benchmark dataset, where they achieved 95.30% accuracy with the ChaLearn dataset. Pigou et al. proposed a simple temporal feature pooling-based method that proved temporal information is more important as discriminative features for video classification-related research work [44]. They also focus on the recurrence information with temporal convolution, which can improve the significants of the video classification task. SINCAN et al. proposed a hybrid method combining an LSTM, Feature pooling and CNN method to recognize isolated sign language [24]. They included the VGG-16 pre-trained model with the CNN part and two parallel architectures for learning RGB and Depth information. Finally, they achieved 93.15% accuracy with Montalbano Italian sign language dataset.

Huang et al. applied a continuous sign language recognition approach to eliminate the preprocessing of temporal segmentation, namely Hierarchical Attention Network with Latent Space (LS-HAN) [28]. They mainly included two-stream CNN, LS and a HAN for video feature extraction, semantic gap bridging and latent space-based recognition, respectively. The main drawback of their work is that they mainly extracted pure-visual features, which are not good for capturing hand gestures and body movements. Zhou et al. proposed a holistic visual appearance-based approach and a 2D human pose-based method to improve the performance with large-scale sign language recognition [23]]. They also applied pose-based temporal graph convolution networks (Pose-TGCN) to extract the pose trajectories' temporal dependencies and achieved 66% accuracy for the 2000 words glosses. Liu et al. applied a feature extraction approach based on the deep CNN with stack temporal fusion layers with a sequence learning model Bidirectional RNN [45]. Guo et al. employed a hierarchical LSTM approach with word embedding, including visual content for SLR [46]. Firstly, spatial-temporal information is extracted by 3D CNN and then compacted into a visemes with the help of an online key based on the adaptive variable length. Their approach is not so much efficient for capturing motion information. The main drawback of image and video pixel-based work is to high computational complexity To overcome these drawbacks, researchers are thinking about the joint point instead of the full image pixels for the hand gesture and action recognition [47–49]. Various models have been used in skeleton-based gesture recognition among LSTM [33] and RNN [50] among them. Yan et a. applied a graph-based method, namely ST-GCN, for building a dynamics pattern for skeleton-based action recognition with Graph convolutional network(GCN) [33]. By following the previous task, many researchers have employed some modified versions of the ST-GCN to improve the performance accuracy for Hand Gesture and Hu-man activity recognition work. Li et all; employed an encoder and a decoder for extracting action-specific latent information [49]. They included two links to do this and finally employed GCN-based action structured GCN to learn temporal and spatial information. Shi et al. employed a two-stream-based GCN for action recognition [51], and a multi-stream GCN for action recognition [30]. In the multi-stream GCN, they integrated the GCN with a spatial temporal-based network to extract the more important joints and features from the all features. Zhang et al. proposed a decoupling GCN to recognize skeleton-based action recognition [29]. Song et al. proposed ResGCN integrating with Part-wise Attention (PartAtt) to improve the performance and computational cost of the skeleton-based action recognition [31]. But their main drawback is their performance is not so much higher than the existing ResNet performance. Amorin et al. proposed a human skeleton movement-based sign language recognition using ST-GCN, where they proposed to select the potential key points from the whole body key points. Finally, they achieved 85.0% accuracy with their dataset name ASLLVD [52]. The disadvantage of this work is that they consider only one hand with the body key point. Perez et al. extracted 67 key points, including the face, body and hand gestures, using a special camera, and finally, they achieved good performance with the LSTM [38]. In the same way, many researchers considered 133 points from the whole body to recognize sign language [8]. Jiang et al. applied a different approach with a multimodal dataset, including full-body skeleton points and achieved a good performance accuracy [8]. They also think of reducing the number of skeletons to increase the model's efficiency. The main problem is that their method did not seem able to achieve good performance and the generalization property for the SLR compared to the existing systems. In addition, researchers focused on the skeleton-based SLR because of the high complexity of the pixel-based system. With the full-body skeleton, we face almost the same computational complexity problems.

3. Dataset

We used three large-scale datasets in this study, which are shown in Table 1. Section 3.1, AUtSL Dataset, in Section 3.2 described the CSL dataset and in Section 3.3 described the WLASL dataset.

Table 1. Evaluated Dataset Description in the Study

Dataset Name	Language Name	Year	Signes	Subjects	Total Sample	Sample Sign
WLASL [53]	Amercian	2020	2000	119	21089	10.5
AUTSL [54]	Turkish	2020	226	43	38336	169.6
CSL [23]	Chines	2019	500	50	125000	250

3.1. AUTSL Dataset

Another sign language dataset is the Turkish Sign Language dataset (AUTSL), collected from diverse, challenging backgrounds, including real-life scenarios. To record the dataset, they used Microsoft Kinect V2, including RGB, depth and skeleton modalities [54]. This dataset was collected from 43 people considering 226 signs. They recorded 38336 video clips in total for the 226 signs with 30 frame speed, and they collected this from 20 different challenging backgrounds. In the background, they considered the camera field-of-view, increasing or decreasing the appearance by adding a new object or removing an object in the background. In addition, moving trees, various lighting conditions, sunlight, artificial light, people passing behind the signer, and some bright-dark areas or shadowed. In the selected signs, they considered the most used word in the turkey language, like push, wait, shoe, face, wait, help, danger, doctor, hospitals, building, and signs. Figure 1 shows the sample sign of the AUTSL Dataset.

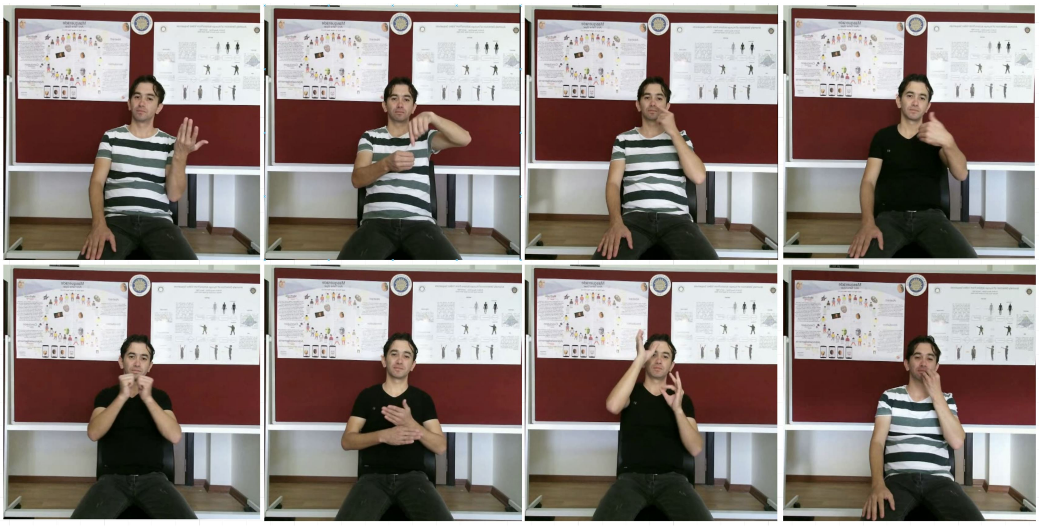

Figure 1. Sample of AUTSL Dataset.

3.2. CSL Dataset

CSL in the context of a large-scale Chinese Sign Language dataset, refers to a collection of videos and/or images of people signing in Chinese Sign Language, along with their corresponding transcriptions and annotations. There are 50 subjects that give the data with Depth, RGB and Skeleton modality. The video had 30 PFS and 1280*720 RGB resolution and 2 to 4 seconds duration videos [23]. They selected 500 different words for the labels and recorded 2 to 4 videos for each word. In total, they recorded 125000 total videos from 50 people with 30 FPS. These datasets are often used for training and evaluating machine learning models for sign language recognition, translation, and other tasks CSL, in the context of a large-scale Chinese Sign Language (CSL) dataset, refers to a collection of videos or images of people signing in CSL, along with their corresponding transcriptions and annotations. These datasets are important resources for research and development in sign language processing, including sign language recognition, translation, and other tasks. A large-scale CSL dataset typically includes many sign language samples recorded from a diverse group of signers, covering a wide range of signs and variations in signing styles. This diversity ensures that machine learning models trained on the dataset can generalize to real-world signing scenarios. The annotations in a CSL dataset can include information such as the signs being performed, each sign's start and end times, and the signing posture and movement patterns. This information trains and evaluates machine learning models that aim to recognize and transcribe sign language. Large-scale CSL datasets are crucial in advancing sign language processing technology and making it more accessible to the deaf and hard-of-hearing community. These datasets provide a foundation for building more accurate and effective sign language recognition and translation systems, which can help bridge the communication gap between the hearing and non-hearing worlds. Figure 2 shows the sample sign of the CSL dataset.

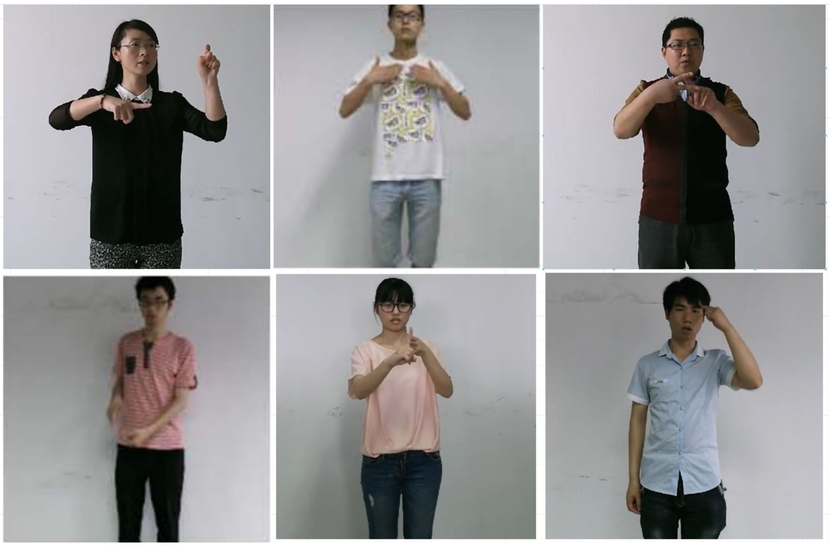

Figure 2. Sample of CSL Dataset.

3.3. WLASL Dataset

WLASL is one of the SLR's large-scale video datasets from the Word Level American Sign Langauge dataset (WLASL). This dataset was made by the University of Central Florida sign language re-searchers team, collecting 21089 individual videos with 200 unique signs of ASL [53]. There were 200 people who participated in the recording process as a subject. During the collecting dataset, only one sign was collected from one signer with some repetition and in most cases, it was a frontal view but included diverse, complex backgrounds. After collecting data, it was labelled using the English glosses, where each gloss annotation contained only one word. If an individual gloss contains more than one word, it will select only one word by discarding the others using unique rules. The aim of selecting only one word for a gloss is to ensure sufficient samples in the training and test sets. The video was collected with a camera known as Microsoft Kinect, and the signer gave the dataset from in front of the camera view where signers performed multiple times for unique signs for capturing the divers signing style and camera angles. After that, the sort gloss in descending order for the prepared sample number of gloss yields a better understanding formate for the word level scalability of the sign recognition method and sign recognition tasks. The dataset also contained a different number of glosses like 100, 300, 1000, and 2000, which is the defined as four subsets, namely WLASL100, WLASL300, WLASL1000 and WLASL2000, respectively. The WLASL dataset is an important resource for developing and evaluating computer vision and machine learning algorithms for sign language recognition. It has been used in several research papers and is freely available for non-commercial research purposes.

4. Proposed Methodology

In the study, we developed Multi-Stream Graph-based Deep Neural Networks (SL-GDN) to recognize sign language using potential skeleton points of whole body information. This idea we generated from the concept of Jiang [8]. The key idea is applying a Neural network(NN) with a fully connected layer to construct a fully connected graph from the selected whole-body key points. The objective is to edge and node features dynamically learned via a sequential graph and general convolutional network, which is performed in both spatial and temporal information. Our graph is mainly constructed with a spatial-temporal model for recognizing hand gestures based on human body skeleton information dynamics. We also adopt a multi-stream approach for various information to improve performance further. Although, many researchers developed an SLR system with the 21 key points extracted using a media pipe [35]. Researchers found that only hand information cannot express the sign's exact meaning and emotion. After that, researchers think a full-body skeleton is more efficient for the SLR systems [36]. To do this, some researchers think about the body with hand because this aims to localize the key points or joints of human bodies from a single image or video. Besides the traditional approach, such as pictorial structures [25] and probabilistic model [38] for estimating the single-person poses, many researchers develop their system using the ground truth skeleton, which is come from the motion capture device such as Kinect version 2 [55]. Nowadays, many deep learning-based techniques extract the whole body's key points. Although these deep learning-based pose estimations generated body key points, this is insufficient because it needs the spatial dependencies on the extracted key points. One of the researchers [8] extracted 133 key points for the whole body, including the body, face and two using oppose [12], where 42 key points come from the left and right hand and the rest of the key points come from the upper body [8,9]. Among the 133 we worked with, we selected the 27 most potential key points using the graph reduction approach, considering 20 key points from both the left and right hand and seven from the upper body. Figure 3 shows the detailed working flow architecture of the proposed model where we first took the 27 whole body joint keypoint then extracted joint motion, bone and bone motion key points stream from this based on the formula [8]. In each of the skeleton streams, we applied a NN with a fully connected layer which helps to make a fully connected graph where node and edges feature learn with the graph convolution and deep neural network. We extracted spatial features with the Graph convolutional network. We fed them to the convolutional neural network layer through bach normalization, relu and dropout layer and produced a feature vector. In the same way, we extracted features from the four streams and concatenated them to produce the final feature vector. We fed the final feature vector into the classification module to refine the final feature, and after converting the matrix feature into a vector, we used a classification layer. Figure 3 shows the working flowgraph of the proposed study, and 4 shows the NN, SL-GDN and classification modules separately.

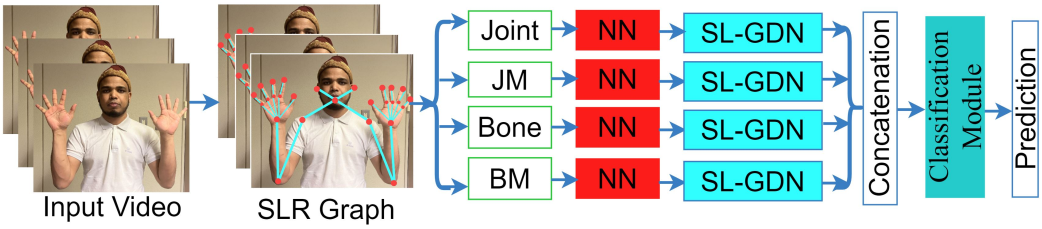

Figure 3. Working Flow Architecture

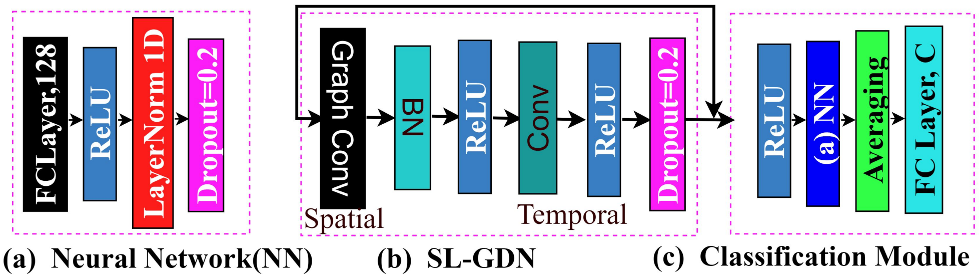

Figure 4. (a) Neural Network (NN) (b) SL-GDN (c) Classification Module

4.1. Key points Selection and Graph Construction

A sequence of vectors comes from the single frame, considered a raw skeleton where individual vectors represent 2D coordinates of the human joints. Moreover, a full hand gesture sign consists of multiple frames based on the number of frames and samples. We constructed a spatial-temporal graph by considering the natural connection among the adjacent skeleton points. We assumed a set of the node set $V = v_(i,t)|i = 1, \ldots, N, t = 1, \ldots, T$, which is mainly body, face and hand pose skeleton points. To construct the adjacent matrix for the graph, we used the following formulas Equation 1.

$$f(x) = \begin{cases} 1 & 1 \text{ if there is adjacent} \\ 0 & \text{if there is no adjacent.} \end{cases} \tag{1}$$

Where if there is adjacent denoted the calculation minimum distance or shortest path between two nodes. As mentioned, 133 key points in our pose, including body, face, and hand, are many nodes. Because of the large number of nodes and edges, there may be unnecessary noise. In addition, in any case, if their two nodes are far from each other, then it isn't easy to extract the relation between the two nodes. Because of the complexity, all key points produce noise, and hard to improve the performance accuracy [8,56]. After that, based on the visualization of the spatial-temporal graph, we selected 27 nodes based on the graph reduction algorithms. The selected 27 nodes included ten nodes for individual hands and seven key points from the upper body parts demonstrated in Figure 1 as an SLR graph. In the graph, we visualize the essential key points of the body to construct an SLR graph. Because of the reduction, the performance and efficiency of the system are increased.

4.2. Neural Network (NN)

In our study, we constructed a graph from the whole body skeleton and then extracted features from a skeleton-based graph using graph convolution and a general neural network. To increase the ability to modify the unified graph dynamically based on the different actions, we employed NN on the skeleton. The main purpose of the NN is to achieve generalizability property for making skeleton graphs, which is not depending on the number of skeleton points. We employed the NN to produce the initial feature from the skeleton points, where we first employed a fully connected layer along with the relu function, then normalised with layer normalization and dropout layer to reduce the overfitting and produced the initial feature F1 [1].

4.3. Graph Convolution

We are considering the spatial-temporal graph based on the spatial partitioning strategy to the dynamic skeleton models to extract the potential pattern embedded with the whole body skeleton graph [8,33]. To construct the spatial graph for the whole body point can be used following Equation 2.

$$G_{out} = D^{-(1/2)}(A + I)D^{-(1/2)}xW \tag{2}$$

Where A, I, D, W have denoted the intr-body connection, self-connection or an identity matrix, diagonal degrees of $(A + I)$ and trainable weight matrix of the convolution, respectively. In the implementation of the graph convolutional, we performed 2d convolutional and multiplied it with the $DD^{-(1/2)}(A + I)D^{-(1/2)}$, which is assumed as a spatial graph convolution. We also used a 2D convolution with the $k_t \times 1$ kernel size to implement the temporal graph convolution. We adopted a Neural network (NN) architecture consisting of a fully connected network to boost the network's capacity. The fully connected results of the NN are fed to the SL-GDN network to produce the final features.

4.4. SL-GDN Architecture Block

The proposed SL-GDN took the output of NN as an input which performed the $k_t \times 1$ convolution on the $C \times T \times N$ initial feature map where NTC denoted the number of vertexes, temporal length and the number of channels, respectively. Here SL-GDN architecture is mainly constructed with a graph convolution layer and batch normalization (BN) layer and relies on a convolutional layer, relu layer and dropout layer. After that, we concatenated this temporal feature with the initial feature and produced the final feature. Finally, we employed a classification module that included the NN concerning refining the final feature. After averaging the refined feature, we average them to convert the matrix into a vector and employ a fully connected layer based on the number of classes denoted by C in Figure 4(c).

4.5. Four-stream Approach

To overcome the inefficient feature problems of the skeleton-based SLR system, we employed the first and 2nd order representation of the skeleton points, namely joints coordinates and bone coordinates with their motion vector [11,48,49]. Figure 3 shows the multi-four stream SL-GDN uses joint, joint, bone, and bone motion. Joint data in a vector form indicated from source to target joints produced the Bone data based on the natural connection of the human body. Here we consider the nose a zero-number joint known as a root joint of the human body, and bone data for the nose is 0. Assume the source and target joint can be expressed as $v^J_{p,t} = (x_{p,t}, y_{p,t}, S_{p,t})$ and $v^J_{q,t} = (x_{q,t}, y_{q,t}, S_{q,t})$ where the $x - y$ score and confidence score are represented by x,y and S. The bone vectors can be calculated by subtracting the source joint and target joint as $v^B_{p,t} = (x_{p,t} - x_{q,t}, y_{p,t} - y_{q,t}, S_{p,t})$ here (p, q) is the set of keypoint joint for face, body, and hand pose. The difference between adjacent frames produces the motion data for both joint and bone motion [8]. Based on the mentioned formula, we can calculate the joint motion as follows $v^{J,M}_{p,t} = (x_{p,t} - x_{p,t+1}, y_{p,t} - y_{p,t+1}, S_{p,t})$. In the same bone motion can be calculated as $v^{B,M}_{p,t} = v^B_{p,t} - v^B_{p,t+1}$. We trained each stream with individual data and produced the feature, and finally, we concatenated all four features to produce the final feature vector [29,30]

4.6. Classification Module

After concatenating the four stream feature, we made a final feature vector and applied a classification module for the prediction. In the classification module, there are two parts; where first part includes a fully connected layer with relue and dropout layer to refine the final feature vector, and finally, we average the matrix into a vector and employ a fully connected layer with several classes and ready for classification.

5. Experimental result

Conducting experiments on sign language classification with three large-scale datasets, we investigate the proposed model's superiority and effectiveness in this section. We extracted the proposed model's performance accuracy first, then reported the state-of-the-art comparison.

5.1. Environmental Setting

To evaluate the model, we used three large-scale datasets, namely WLASL, CSL and AUTSL, with predefined training sets and tests of their dataset. Each dataset has four types of the stream: joint, bone, joint motion, and bone motion. Each stream produced individual features using the SL-GDN model, and we concatenated the 4 four features. Finally, we refined and classified the final feature with the classification module. To implement the proposed system, we used here google colab environment and python programming language. For the framework, we used here Pytorch [57] python programming language framework in the Google Colab Pro edition environment, which provided us Tesla 100 machine with 25GB GPU processing power [58]. Pytorch is one kind of a boon to the attention, transformer and deep learning model. This is because an open source requires a minimum computational cost and high compatibility and adaptability properties with minimum resources. In addition, we used here OpenCV package, open pose, pickle and csv for the initial processing [59,60]. The main goal of the pickle package is to convert the dataset into a byte stream for portable storage. We used the Numpy and Pandas packages that increase flexibility in matrix multiplication and other operations to process the statistical and mathematical procedures. We use initial learning here to reduce the high fluctuation rate to fasten the convergence of the training and testing with the Adam optimizer [60]. We used the 1000 epochs for tuning the model with various parameter tuning operations for the learning rate and optimizer for the multi classes of the study.

5.2. Performance Accuracy with the Three dataset

Table 2 demonstrates the performance of the proposed model, which includes the performance of the AUTSL, CSL and WLASL datasets. We visualize the performance for the individual four streams and the multi-stream of the proposed model. Table 1 reported 96.00% accuracy for the AUTSL data with Joint information and 95.00%, 94.00%, 93.00% and 96.00% for the joint motion, bone, bone motion, and multi-stream keypoint, respectively. In the same way, CSL showed 26.00%,27.00%, 26.00%, and 25.00%26.12% accuracy for the joint, joint motion, bone, bone motion, and multi-stream keypoint, respectively. We also included the WLASL dataset, which showed 50.00%,49.00%,48.00%,48.00%, and 50.00% accuracy for the joint, joint motion, bone, bone motion, and multi-stream keypoint, respectively.

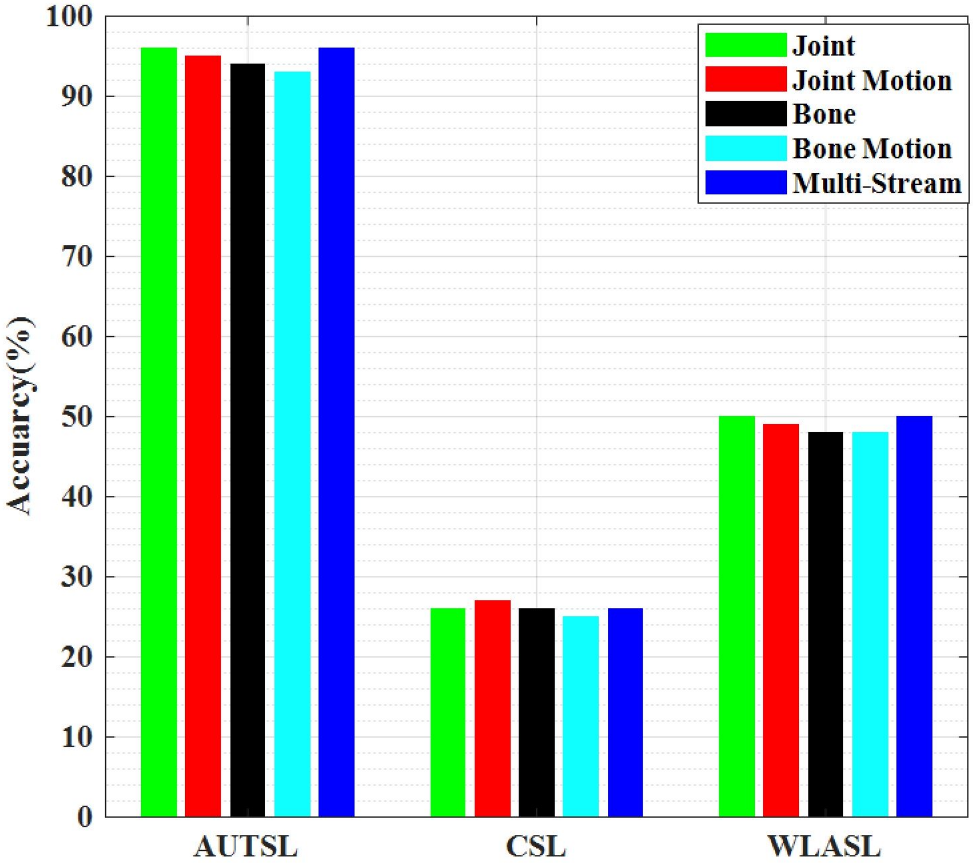

Figure 5. Proposed Model Performance for 3 Dataset.

Table 2. Proposed Model Performance for 3 Dataset.

Stream	AUTSL	CSL	WLASL
Joint	96.00	88.70	50.00
Joint Motion	95.00	87.00	49.00
Bone	94.00	86.00	48.00
Bone Motion	93.00	86.50	48.00
Multi-Stream	96.00	89.45	50.00

5.3. State of the art comparison of the proposed model with AUTSL Dataset

Table 3 shows the state-of-the-art comparison of the proposed model with Jiang [8]. The proposed model showed 96.00% accuracy for the Multi-Stream modal, whereas the existing model produced 95.54% accuracy. Jiang et al. proposed various models with various criteria and key points based on one method they used [8]. They also employed various models, and one of the models reported 95.02%,94.70%,93.10%,92.49% and 95.45% accuracy for the joint, joint motion, bone, bone motion, and multi-stream key point, respectively.

Table 3. State of the art comparison for the AUTSL datasets.

Dataset Types	Method Name	Performance
RGB+Depth	CNN+FPM+LSTM+Attention [54]	83.93
Skeleton Joint	Jiang[8]	95.02
Skeleton Joint Motion	Jiang [8]	94.70
Skeleton Bone	Jiang [8]	93.10
Skeleton Bone Motion	Jiang [8]	92.49
Skeleton Multi-Stream	Jiang [8]	95.45
Skeleton Joint	Proposed Model	96.00
Skeleton Joint Motion	Proposed Model	95.00
Skeleton Bone	Proposed Model	94.00
Skeleton Bone Motion	Proposed Model	93.00
Skeleton Multi-Stream	Proposed Model	96.45

5.4. State of the art comparison of the proposed model with CSL Dataset

Table 4 shows the state-of-the-art comparison for the CSL dataset, which reported 88.70% accuracy for the proposed model where the existing model 3D-CNN [27] achieved88.70% accuracy.

Table 4. State of the art comparison for the CSL datataset.

Dataset Name	Dataset Types	Methodology	Performance [%]
CSL	RGB-D+Skeleton	3D-CNN [27]	88.70
Proposed Model	Skeleton	SL-GDN	89.45

6. Conclusions

In our work, we proposed a Multi-Stream Graph-based Deep Neural Network (MSL-GDN) for a skeleton-based SLR system where we consider the four streams of the skeleton-based sign language dataset. Specifically, we made a graph that is known as a skeleton graph for the whole body pose key points and then applied MSL-GDN to compute the spatial and temporal features. We extracted individual features from each stream, and finally, we concatenated and applied the classification module to refine the feature vector and classification. The performance table proved the superiority and efficiency of the proposed model because of the high accuracy with WLASL, AUTSL and CSL large-scale datasets. The reason for the high efficiency of the proposed model is that we selected 27 whole-body key points among the 133 body pose key points. We plan to combine the skeleton's final feature with the other modal dataset like the RGB and Depth. In addition, we will work to calculate the inverse dynamic from the video with different pose models and then apply the MSL-GDN model to the inverse dynamic features.

Author Contributions: Conceptualization, Abu Saleh Musa Miah; methodology, Abu Saleh Musa Miah, Si-Woong Jang,H youn-Sup Lee, Md Al Mehedi Hasan, and Jungpil Shin; investigation, Abu Saleh Musa Miah, Md Al Mehedi Hasan, and Jungpil Shin; data collection, Abu Saleh Musa Miah, data curation, Abu Saleh Musa Miah, and Md Al Mehedi Hasan, Si-Woong Jang, Jungpil Shin; writing—original draft preparation, Abu Saleh Musa Miah, Jungpil Shin; writing—review and editing, Abu Saleh Musa Miah, Jungpil Shin; visualization, Abu Saleh Musa Miah, and Md Al Mehedi Hasan; supervision, Jungpil Shin; funding acquisition, Si-Woong Jang, Hyoun-Sup Lee, Jungpil Shin.

Funding: This research work is supported by the Ministry of Science and ICT (MSIT), Korea, under the Grand Information Technology Research Center support program (IITP-2023-2020-0-01791) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This work was supported by the Competitive Research Fund of The University of Aizu, Japan

Conflicts of Interest: The authors declare no conflict of interest

References

1. Miah, A.S.M.; Hasan, M.A.M.; Shin, J. Dynamic Hand Gesture Recognition using Multi-Branch Attention Based Graph and General Deep Learning Model. *IEEE Access* **2023**.
2. Miah, A.S.M.; Hasan, M.A.M.; Shin, J.; Okuyama, Y.; Tomioka, Y. Multistage Spatial Attention-Based Neural Network for Hand Gesture Recognition. *Computers* **2023**, *12*, 13.
3. Miah, A.S.M.; Shin, J.; Hasan, M.A.M.; Rahim, M.A. BenSignNet: Bengali Sign Language Alphabet Recognition Using Concatenated Segmentation and Convolutional Neural Network. *Applied Sciences* **2022**, *12*, 3933.
4. Miah, Abu Saleh Musa, S.J.; Hasan, M.A.M.; Rahim, M.A.; Okuyama, Y. Rotation, Translation And Scale Invariant Sign Word Recognition Using Deep Learning. *Computer Systems Science and Engineering*, *44*.
5. Miah, A.S.M.; Shin, J.; Islam, M.M.; Molla, M.K.I.; others. Natural Human Emotion Recognition Based on Various Mixed Reality (MR) Games and Electroencephalography (EEG) Signals. 2022 IEEE 5th Eurasian Conference on Educational Innovation (ECEI). IEEE, 2022, pp. 408–411.
6. Miah, A.S.M.; Mouly, M.A.; Debnath, C.; Shin, J.; Sadakatul Bari, S. Event-Related Potential Classification Based on EEG Data Using xDWAN with MDM and KNN. International Conference on Computing Science, Communication and Security. Springer, 2021, pp. 112–126.
7. Emmorey, K. *Language, cognition, and the brain: Insights from sign language research*; Psychology Press, 2001.
8. Jiang, S.; Sun, B.; Wang, L.; Bai, Y.; Li, K.; Fu, Y. Skeleton aware multi-modal sign language recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3413–3423.
9. Yang, Q. Chinese sign language recognition based on video sequence appearance modeling. 2010 5th IEEE Conference on Industrial Electronics and Applications. IEEE, 2010, pp. 1537–1542.
10. Valli, C.; Lucas, C. *Linguistics of American sign language: An introduction*; Gallaudet University Press, 2000.
11. Mindess, A. *Reading between the signs: Intercultural communication for sign language interpreters*; Nicholas Brealey, 2014.
12. Shin, J.; Musa Miah, A.S.; Hasan, M.A.M.; Hirooka, K.; Suzuki, K.; Lee, H.S.; Jang, S.W. Korean Sign Language Recognition Using Transformer-Based Deep Neural Network. *Applied Sciences* **2023**, *13*, 3029.
13. Miah, A.S.M.; Shin, J.; Hasan, M.A.M.; Molla, M.K.I.; Okuyama, Y.; Tomioka, Y. Movie Oriented Positive Negative Emotion Classification from EEG Signal using Wavelet transformation and Machine learning Approaches. 2022 IEEE 15th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC). IEEE, 2022, pp. 26–31.
14. Miah, A.S.M.; Rahim, M.A.; Shin, J. Motor-imagery classification using Riemannian geometry with median absolute deviation. *Electronics* **2020**, *9*, 1584.
15. Miah, A.S.M.; Islam, M.R.; Molla, M.K.I. Motor imagery classification using subband tangent space mapping. 2017 20th International Conference of Computer and Information Technology (ICCIT). IEEE, 2017, pp. 1–5.
16. Lowe, D.G. Object recognition from local scale-invariant features. Proceedings of the seventh IEEE international conference on computer vision. Ieee, 1999, Vol. 2, pp. 1150–1157.
17. Zhu, Q.; Yeh, M.C.; Cheng, K.T.; Avidan, S. Fast human detection using a cascade of histograms of oriented gradients. 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). IEEE, 2006, Vol. 2, pp. 1491–1498.

18. Dardas, N.H.; Georganas, N.D. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Transactions on Instrumentation and measurement* **2011**, *60*, 3592–3607.
19. Memiş, A.; Albayrak, S. A Kinect based sign language recognition system using spatio-temporal features. Sixth International Conference on Machine Vision (ICMV 2013). SPIE, 2013, Vol. 9067, pp. 179–183.
20. Li, Y.; Wang, X.; Liu, W.; Feng, B. Deep attention network for joint hand gesture localization and recognition using static RGB-D images. *Information Sciences* **2018**, *441*, 66–78.
21. Lim, K.M.; Tan, A.W.C.; Lee, C.P.; Tan, S.C. Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. *Multimedia Tools and Applications* **2019**, *78*, 19917–19944.
22. Shi, B.; Del Rio, A.M.; Keane, J.; Michaux, J.; Brentari, D.; Shakhnarovich, G.; Livescu, K. American sign language fingerspelling recognition in the wild. 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 145–152.
23. Li, D.; Rodriguez, C.; Yu, X.; Li, H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 1459–1469.
24. Pigou, L.; Van Den Oord, A.; Dieleman, S.; Van Herreweghe, M.; Dambre, J. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision* **2018**, *126*, 430–439.
25. Sincan, O.M.; Tur, A.O.; Keles, H.Y. Isolated sign language recognition with multi-scale features using LSTM. 2019 27th signal processing and communications applications conference (SIU). IEEE, 2019, pp. 1–4.
26. Tur, A.O.; Keles, H.Y. Isolated sign recognition with a siamese neural network of RGB and depth streams. IEEE EUROCON 2019-18th International Conference on Smart Technologies. IEEE, 2019, pp. 1–6.
27. Huang, J.; Zhou, W.; Li, H.; Li, W. Attention-based 3D-CNNs for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology* **2018**, *29*, 2822–2832.
28. Huang, J.; Zhou, W.; Zhang, Q.; Li, H.; Li, W. Video-based sign language recognition without temporal segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, Vol. 32.
29. Cheng, K.; Zhang, Y.; Cao, C.; Shi, L.; Cheng, J.; Lu, H. Decoupling gcn with dropgraph module for skeleton-based action recognition. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. Springer, 2020, pp. 536–553.
30. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing* **2020**, *29*, 9532–9545.
31. Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Stronger, faster and more explainable: A convolutional graph baseline for skeleton-based action recognition. proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1625–1633.
32. Wang, H.; Wang, L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 499–508.
33. Yan, S.; Xiong, Y.; Lin, D. Spatial, temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI conference on artificial intelligence, 2018, Vol. 32.
34. Oberweger, M.; Lepetit, V. Deepprior++: Improving fast and accurate 3d hand pose estimation. Proceedings of the IEEE international conference on computer vision Workshops, 2017, pp. 585–594.
35. Shin, J.; Matsuoka, A.; Hasan, M.A.M.; Srizon, A.Y. American sign language alphabet recognition by extracting feature from hand pose estimation. *Sensors* **2021**, *21*, 5856.
36. Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; Luo, P. Whole-body human pose estimation in the wild. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer, 2020, pp. 196–214.
37. Xiao, Q.; Qin, M.; Yin, Y. Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural networks* **2020**, *125*, 41–55.
38. Mejía-Peréz, K.; Córdova-Esparza, D.M.; Terven, J.; Herrera-Navarro, A.M.; García-Ramírez, T.; Ramírez-Pedraza, A. Automatic recognition of Mexican Sign Language using a depth camera and recurrent neural networks. *Applied Sciences* **2022**, *12*, 5523.
39. Rahim, M.A.; Miah, A.S.M.; Sayeed, A.; Shin, J. Hand gesture recognition based on optimal segmentation in human-computer interaction. 2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII). IEEE, 2020, pp. 163–166.

40. Cai, Z.; Wang, L.; Peng, X.; Qiao, Y. Multi-view super vector for action recognition. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 596–603.
41. Neverova, N.; Wolf, C.; Taylor, G.; Nebout, F. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2015**, *38*, 1692–1706.
42. Pu, J.; Zhou, W.; Li, H. Iterative alignment network for continuous sign language recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4165–4174.
43. Koller, O.; Zargaran, S.; Ney, H.; Bowden, R. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. *International Journal of Computer Vision* **2018**, *126*, 1311–1325.
44. Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; Saenko, K. Sequence to sequence-video to text. Proceedings of the IEEE international conference on computer vision, 2015, pp. 4534–4542.
45. Cui, R.; Liu, H.; Zhang, C. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia* **2019**, *21*, 1880–1891.
46. Guo, D.; Zhou, W.; Li, H.; Wang, M. Hierarchical LSTM for sign language translation. Proceedings of the AAAI conference on artificial intelligence, 2018, Vol. 32.
47. Parelli, M.; Papadimitriou, K.; Potamianos, G.; Pavlakos, G.; Maragos, P. Exploiting 3D Hand Pose Estimation in Deep Learning-Based Sign Language Recognition from RGB Videos. Computer Vision – ECCV 2020 Workshops; Bartoli, A.; Fusiello, A., Eds.; Springer International Publishing: Cham, 2020; pp. 249–263.
48. Cai, J.; Jiang, N.; Han, X.; Jia, K.; Lu, J. JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition. 2021, pp. 2734–2743. doi:10.1109/WACV48630.2021.00278.
49. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3595–3603.
50. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5457–5466.
51. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12026–12035.
52. de Amorim, C.C.; Macêdo, D.; Zanchettin, C. Spatial-temporal graph convolutional networks for sign language recognition. Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings 28. Springer, 2019, pp. 646–657.
53. Chen, Y.; Zuo, R.; Wei, F.; Wu, Y.; Liu, S.; Mak, B. Two-Stream Network for Sign Language Recognition and Translation. *arXiv preprint arXiv:2211.01367* **2022**.
54. 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.
55. Pagliari, D.; Pinto, L. Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors. *Sensors* **2015**, *15*, 27569–27589.
56. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer, 2016, pp. 816–833.
57. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; others. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **2019**, *32*.
58. Tock, K. Google CoLaboratory as a platform for Python coding with students. *RTSRE Proceedings* **2019**, *2*.
59. Gollapudi, S. *Learn computer vision using OpenCV*; Springer, 2019.
60. Dozat, T. Incorporating nesterov momentum into adam **2016**.

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.