Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7328–7338
Marseille, 20-25 June 2022

Bidirectional Skeleton-Based Isolated Sign Recognition using Graph Convolutional Networks

Konstantinos M. Dafnis[1], **Evgenia Chroni**[1], **Carol Neidle**[2], **Dimitris N. Metaxas**[1]
[1] Rutgers University, [2] Boston University
[1] 110 Frelinghuysen Road, Piscataway, NJ 08854,
[2] Boston University Linguistics, 621 Commonwealth Ave., Boston, MA 02215
kd703@cs.rutgers.edu, etc44@cs.rutgers.edu, carol@bu.edu, dnm@cs.rutgers.edu

Abstract

To improve computer-based recognition from video of isolated signs from American Sign Language (ASL), we propose a new skeleton-based method that involves explicit detection of the start and end frames of signs, trained on the ASLLVD dataset; it uses linguistically relevant parameters based on the skeleton input. Our method employs a bidirectional learning approach within a Graph Convolutional Network (GCN) framework. We apply this method to the WLASL dataset, but with corrections to the gloss labeling to ensure consistency in the labels assigned to different signs; it is important to have a 1-1 correspondence between signs and text-based gloss labels. We achieve a success rate of 77.43% for top-1 and 94.54% for top-5 using this modified WLASL dataset. Our method, which does not require multi-modal data input, outperforms other state-of-the-art approaches on the same modified WLASL dataset, demonstrating the importance of both attention to the start and end frames of signs and the use of bidirectional data streams in the GCNs for isolated sign recognition.

Keywords: ASL, Isolated Sign Recognition, GCN, Linguistic Modeling

1. Introduction

There are 28 million Deaf or Hard of Hearing people (Lin et al., 2011) in the US, and American Sign Language (ASL) is the primary language for $\geq$500,000 people (Mitchell et al., 2006). It is also the 3rd most studied "foreign" language (Looney and Lusin, 2019). Signed languages involve articulations of the hands and arms and non-manual expressions: facial expressions and movements of the head and upper body. Computer-based sign language recognition from video would pave the way for technologies to improve communication between deaf and hearing individuals, such as ASL-to-English translation; educational applications to support ASL learners; or Google-like sign search by example over videos on the Web. These same technologies could also be applied to other signed languages.

The research reported here focuses on recognition of isolated, citation-form signs. The linguistically significant aspects of sign production include particular hand configurations, palm orientations, locations (places of articulation), and movements, as well as non-manual components in some cases. As would be expected in any language, the production of words (signs) shows considerable variability, and inter- and intra-signer variations pose a challenge for computer-based sign recognition. For this reason, it is important to use a large video corpus with sufficient numbers of examples for each sign, including multiple signers. In this research, we used the WLASL dataset (Li et al., 2020), a collection of videos taken from various sources that includes 119 signers in 21,083 videos of 2,000 distinct isolated signs. However, there are inconsistencies in the gloss labels associated with signs. We thus modified the gloss labeling to enforce consistency. We also restricted the set of signs to lexical signs (the largest class), for which there is a fixed vocabulary. For example, we did not include fingerspelled signs and other sign types in this research. The resulting dataset, with corrected annotations, that we used for these experiments consists of 18,141 videos for 1,449 lexical signs, and we evaluated the performance of our deep learning approach on this modified dataset.

Before the advent of deep learning, there were several approaches to isolated sign recognition using traditional machine learning methods (e.g., Hidden Markov Models (HMMs), Conditional Random Fields (CRFs) (Lafferty et al., 2001)) to analyze the spatiotemporal information in a video sign sequence (Grobel and Assan, 1997; Dilsizian et al., 2014; Fatmi et al., 2017; Metaxas et al., 2018; Tornay et al., 2020). Some methods also incorporated some degree of linguistic modeling (Dilsizian et al., 2014; Metaxas et al., 2018). Recent advances in deep machine learning have given rise to new methodologies for sign recognition (Lim et al., 2019; Sincan et al., 2019; Sincan and Keles, 2021; Masood et al., 2018), which include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long short-term Memory (LSTMs) (Hochreiter and Schmidhuber, 1997), and often multi-modal input learning fusion approaches. However, such approaches still achieve limited success in sign recognition from large vocabularies.

To improve isolated sign recognition accuracy, we propose a new skeleton-based method that involves explicit detection of start and end frames of signs and uses linguistically relevant parameters based on the skeleton input, employing a bidirectional learning approach within a GCN framework. We achieve 77.43% top-1 and 94.54% top-5 accuracy using the modified WLASL dataset. Furthermore, we compare our method with that of Jiang et al. (2021) on the same modified dataset, demonstrating the importance of linguistically motivated parameters and attention to the start and end frames of signs; our method does not require multimodal data input.

2. Related Work

Before the rise of deep learning methods, sign recognition frameworks often used hand-crafted features, such as relative hand positions and distances between the hands and specific body parts (Tornay et al., 2020; Cooper et al., 2012; Badhe and Kulkarni, 2015; Xiaohan Nie et al., 2015), in conjunction with standard classifiers, such as Support Vector Machines (SVMs), k-Nearest Neighbors (kNNs), CRFs and HMMs (Memiş and Albayrak, 2013; Dardas and Georganas, 2011; Yang, 2010; Metaxas et al., 2018; Tornay et al., 2020). However, the hand-crafted features and underlying Gaussian distribution assumptions inherent to these approaches have resulted in systems with very limited capability for generalization and scalability. Recent deep neural network based learning methods address these limitations and have produced state-of-the-art results in computer vision tasks such as action and gesture recognition; these methods have also been applied to sign language recognition, which is a related but considerably more complex problem because of the importance of linguistic structure. For sign language recognition, several data inputs have been used, such as RGB video, depth video, or both (Rastgoo et al., 2021). Some recent methods fuse information from those inputs, using raw RGB video frames, 2D extracted skeletons from RGB video, and/or depth data to improve sign language recognition (Jiang et al., 2021).

2.1. RGB-based Approaches

Early sign language approaches used CNNs to extract spatial features in each RGB frame in combination with Recurrent Neural Networks, such as LSTM or Bidirectional LSTMs (Bi-LSTMs), to capture temporal information (Sincan et al., 2019; Koller et al., 2019; Papastratis et al., 2020; Cui et al., 2019). Some research has also used modified CNNs to capture short-term dependencies. For example, Tran et al. (2015) first proposed a 3D-CNN to improve action recognition. Many researchers subsequently leveraged modified CNNs in the context of action and sign language recognition (Liang et al., 2018; Li et al., 2020; Vaezi Joze and Koller, 2019). The most extensively used architecture based on 3D-CNN networks is the Inflated 3D-CNN (I3D) model (Carreira and Zisserman, 2017); variations include separable 3D-CNNs (S3D) (Xie et al., 2018). Although 3D-CNN models perform better than previous approaches in learning short-term memory dependencies, a major drawback is that they restrict the learning of long-term dependencies at the final temporal global average pooling stage. In order to overcome this disadvantage in the domain of action recognition, the authors in (Kalfaoglu et al., 2020), inspired by Natural Language Processing methods, used a Bidirectional Transformer (BERT) (Devlin et al., 2018). The attention mechanism of this Transformer worked quite well for dealing with most of the temporal dependencies.

Some newer architectures for video understanding have been based on the use of Transformers, exploiting their self-attention mechanism (Bertasius et al., 2021). This is a promising direction for action recognition and video classification, including sign language recognition. Transformers have the advantage of capturing space-time dependencies over the entire video. However, they require larger amounts of training data than are generally available for sign language recognition.

2.2. Skeleton-based Approaches

Instead of using the raw RGB frames, some methods have used frame-based extracted skeletons to focus the learning on the relevant information. When the skeleton extraction process is robust, these methods show improved learning and recognition performance, as they are not affected by irrelevant information, such as the background. Extracted skeletons can be in the form of sets of body joints (keypoints) or skeleton graphs that include the edges between the joints. The early approaches to action and sign language recognition used CNNs followed by RNNs to learn the relevant temporal information (Soo Kim and Reiter, 2017; Liu et al., 2017). A disadvantage of these models is their inability to encode keypoint interactions in both space and time. To overcome this limitation, Yan et al. (2018) proposed the first Spatial-Temporal Graph Convolutional Network (ST-GCN) and showed the effectiveness of GCNs for learning the spatiotemporal skeleton dynamics. However, ST-GCN extracts and processes spatiotemporal keypoint features using only the human body joint connections. Thus, interactions of keypoints that are not directly connected, such as the keypoints between the 2 hands, are largely ignored. This information is important for recognizing signs. There have recently been attempts to overcome this limitation. Li et al. (2019b) exploits the latent joint connections to improve human action recognition. Shi et al. (2019b) propose a 2-stream approach that uses keypoints and bone information (vectors between consecutive keypoints), while in Shi et al. (2020) the motion of keypoints and bones is added, resulting in improved action recognition. de Amorim et al. (2019) use an extension of the ST-GCN model for sign language recognition, achieving close to 60% accuracy on a vocabulary of 20 signs. This performance is significantly lower, however, than the recognition accuracy achieved with traditional sequence learning models that use skeleton and CNN-based RGB video frame features (Metaxas et al., 2018).

2.3. Multi-Feature Combination Approaches

Recent research has aimed to improve sign recognition accuracy by combining multiple features, such as raw RGB video frame features, 2D/3D extracted skeletons from RGB video frames, and depth data. Rastgoo et al. (2020) use spatial features extracted from pretrained CNNs and skeleton data, with an LSTM model to encode the temporal information of signs, achieving 86.32% sign recognition accuracy on a vocabulary of 249 signs. Jiang et al. (2021) use a GCN approach as in (Shi et al., 2020) and combine multiple features, such as skeleton-based data, RGB-based features, optical flow, and depth video features. This results in a 4-stream framework for isolated sign recognition. The GCN also employs a decoupled spatial convolutional layer to boost the GCN modeling capacity. Using this multi-feature combination learning approach, they achieve top-1 accuracy of 59.39% and top-5 accuracy of 91.48% on the WLASL dataset (2000 signs) (Li et al., 2020), and top-1 accuracy of 98.53% on the AUTSL dataset (226 signs) (Sincan and Keles, 2020).

3. Our Approach

The distinctive aspects of our approach include detection of start and end frames for ASL signs, as a first step; the use of a GCN model for keypoint graphs; and a late fusion strategy that utilizes both forward and backward video streams. We have also taken steps to ensure consistency of the gloss labeling of signs used for this research. Below we report on the data used for this project; we then describe the technical approach.

3.1. Data for This Project

We perform sign recognition on lexical signs in the WLASL dataset (Li et al., 2020), with enforced consistency of gloss labeling, as described below. We use the ASLLRP ASLLVD dataset (Neidle and Opoku, 2021; Neidle et al., 2018; Neidle et al., 2012), with almost 10,000 linguistically annotated citation-form sign examples corresponding to over 3,300 distinct signs from 6 signers, for training on detection of start and end points of signs.

3.1.1. Critical Importance of 1-1 Correspondence between Signs and Gloss Labels

Deficiencies in the quality and accuracy of annotations in sign language corpora are a key limitation for progress in sign recognition research (Bragg et al., 2019). Research based on gloss labels for signs faces a serious challenge, in light of the fact that: (1) there is no 1-1 correspondence between English words and ASL signs, and (2) there are also no established glossing conventions shared by the ASL/research community. The ASLLRP projects (Neidle and Opoku, 2021; Neidle et al., 2018; Neidle et al., 2012) have established conventions to ensure a 1-to-1 correspondence between gloss label and ASL sign production, which is critically important. See (Neidle et al., 2012) for discussion of the challenges posed in establishing glossing conventions. Serious problems arise, however, when researchers use datasets where 1-1 gloss label to sign correspondences have not been enforced; or when multiple datasets using inconsistent glossing conventions are combined. This is the situation for the WLASL dataset, which brings together multiple, publicly shared, ASL video corpora from different sources.

3.1.2. Problems with the Gloss Labels Provided for the WLASL dataset

Although Li et al. (2020) claim that sign variations (often attributed to dialect variation rather than to variations in labeling) have been taken into account, serious labeling inconsistencies remain. For example, in the WLASL dataset the same sign is sometimes glossed as “reply” and sometimes as “answer”; compare, e.g., video ID 2718, 2713, and 4735, among the examples glossed as “answer” with 47343, 47345, and 47342, among those glossed as “reply,” as shown in Fig. 1.

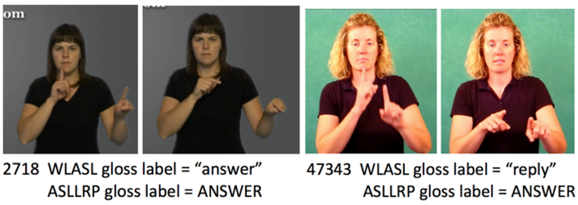
om
2718 WLASL gloss label = “answer”
ASLLRP gloss label = ANSWER
47343 WLASL gloss label = “reply”
ASLLRP gloss label = ANSWER

Figure 1: Example of 2 different WLASL gloss labels for a single sign.

Conversely, the gloss “right” is used sometimes to mean the opposite direction from “left,” as in 48107, 48109, and 48114; and sometimes for the sign that means “correct”, as in 48105, 48106, and 48115. This is shown in Fig. 2.

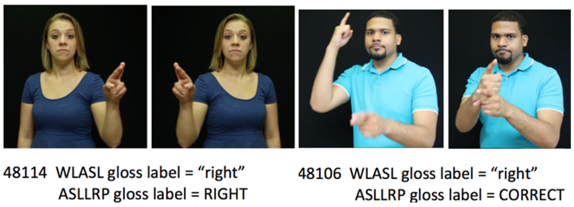
48114 WLASL gloss label = “right”
ASLLRP gloss label = RIGHT
48106 WLASL gloss label = “right”
ASLLRP gloss label = CORRECT

Figure 2: Example of 2 different ASL signs assigned the same WSLASL gloss label.

Note that the WLASL also has other occurrences of the sign shown in 48106 of Fig. 2 that are, in fact, glossed as “correct”, e.g. 13359, as shown in Fig. 3.

There is also an odd attribution of differences in sign productions to “dialect” variation. For example, the English word ‘correct’ can also be used as a verb, ‘to correct’. There is a different ASL sign used for that meaning of the word (which has a range of other possible meanings, as well), shown in Fig. 4. However, this is treated as a dialectal variant of “correct,” although it is clearly not a dialectal (or any other kind of) variant of the ASL sign shown in Fig. 3 (possible English translations of these signs notwithstanding).

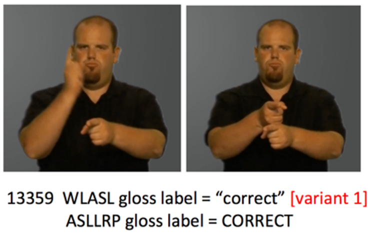
13359 WLASL gloss label = "correct" [variant 1]
ASLLRP gloss label = CORRECT

Figure 3: A different WLASL gloss label assigned to 1 of the signs shown in Fig. 2.

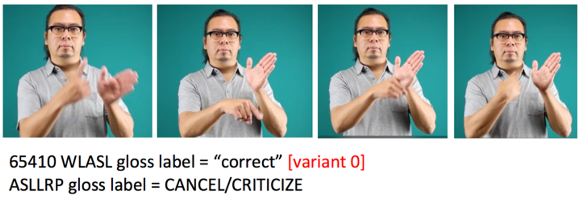
65410 WLASL gloss label = "correct" [variant 0]
ASLLRP gloss label = CANCEL/CRITICIZE

Figure 4: Supposed "dialectal variant" of the WLASL sign shown in Fig. 3.

These are not isolated examples. There are large numbers of signs in the WLASL data beset with the types of issues just mentioned.

3.1.3. Adjustments Made to Address This Issue

We made an effort to associate with the publicly shared WLASL data gloss labels consistent with those used for the ASLLRP (including the ASLLVD) datasets. This has 2 advantages: (1) the ASLLRP conventions are well established, with a Web-accessible Sign Bank (ASLLRP, 2017-2022), and thus such relabeling was feasible, thereby drastically improving the consistency of gloss labeling for the WLASL data; and (2) this will make it possible, in the future, to combine the WLASL and ASLLRP data to make an even larger set of data that can be used for sign recognition research.

3.2. Technical Approach

3.2.1. Detecting Sign Start and End Frames

In ASL datasets of citation-form signs (including the ASLLVD and WLASL datasets, among many others), the video clip typically includes frames both before and after the core region of the sign itself. Thus, processing the entirety of the clip, including frames external to the sign, for extraction of RGB-based features and/or skeleton-based keypoints, may result in introduction of additional noise to the model because of (1) movements of human body parts not directly related to the sign, and (2) the possibility of extensive blurriness in the transition frames. For this reason, we have introduced detection of the start and end points, so that we can restrict attention to the linguistically informative frames.

The start point of the sign is the frame in which the features of the sign are fully realized, i.e., when the initial hand configuration has been fully formed and the initial palm orientation and place of articulation (location) have been reached. Thus, there is a convergence of probable changes that may signal the start of the sign, including stabilization of the hand configuration and orientation, decrease in acceleration, and potentially an abrupt change in the direction of movement. The reverse occurs at the end of the sign, i.e., the hand configuration, palm orientation, and/or location and acceleration of the hands are likely to change.

Our methodology for detection of the start and end frames of isolated signs from RGB video consists of 2 steps. (1) We detect both hands in the video frames using a pre-trained YOLOv3 neural network (Redmon and Farhadi, 2018) that computes 2 bounding boxes around each of the visible hands. Each bounding box is modeled by computing the 2D coordinates of the upper left corner of the tightest bounding box, its width and height, as well as the confidence score of the bounding box detection. Based on a threshold, we use this confidence score to remove the frames at the beginning and end of the video where both hands are not visible. (2) We use a Bi-LSTM machine learning approach to estimate the start and end frames of the sign based on the detection of changes in the velocity and/or shape of the hands. We first calculate at each frame the location and velocity of the center of mass of the hands' bounding boxes. To estimate robustly the probability of a boundary point at time *t*, our Bi-LSTM model takes into account several frames before and after time *t*.

The input to the Bi-LSTM network is a sequence vector of fixed length T (T is set to the 95 percentile of the sign lengths) of x_t, $t = 1 : T$. It includes spatial hand features within the detected bounding boxes using a pre-trained CNN and the velocity and acceleration of the bounding boxes' center of mass. The output of the Bi-LSTM network is a probability sequence vector with the same length p_t, $t = 1 : T$, where p_t indicates the probability for time t to be a boundary point (start or end of a sign). Since start and end points have different statistics, we train separate Bi-LSTM networks for detection of the start and end points of the sign. To improve end point detection and avoid false positives due to repetitive movements, noise, blurring and other artifacts in the course of sign production, we reverse the input feature sequence. We train these networks on the ASLLVD dataset (Athitsos et al., 2008; Neidle et al., 2012; Neidle et al., 2018; Neidle and Opoku, 2021), which has start/end ground truth annotations. Furthermore, although the WLASL dataset lacks such annotations, we manually annotated a small set of signs for start and end points, and we used those to fine-tune the network parameters. Since we want to detect the beginning and end of a sign (in which there may be repetitive movements), we keep the first detected boundary point candidate in the original and reversed sign sequences.

3.2.2. GCN for Skeleton Keypoints

The GCN learning approach for isolated sign recognition consists of graph construction, skeleton graph keypoint selection, spatial-temporal graph convolutions, bidirectional stream learning, and score fusion. The innovations in our approach include bidirectional learning, keypoint selection, and late fusion of skeleton keypoint positions, velocities, and accelerations.

Graph Construction. To construct the spatial-temporal graphs, we first connect the spatially adjacent keypoints for the human anatomy. We also connect the keypoints to themselves in the temporal dimension. Thus, the corresponding adjacency matrix A in the spatial dimension is constructed using the keypoint set $V = \{u_i | i = 1, \dots, N\}$ as:

$$A = \begin{cases} 0, & \text{if } d(u_i, u_j) \neq 1 \\ 1, & \text{if } d(u_i, u_j) = 1 \end{cases} \tag{1}$$

where $d(u_i, u_j)$ is the minimum distance between any skeleton keypoints u_i, u_j, $i \neq j$, on the graph.

Spatial-Temporal Graph Convolution. To capture the structure embedded in connections of the body keypoints and model the spatiotemporal changes in the skeleton representation, we use spatial-temporal graph convolutions with a spatial partitioning strategy of the ST-GCN (Yan et al., 2018).The implementation of the spatial part of the GCN is expressed as follows:

$$x_{out} = \Lambda^{-\frac{1}{2}}(I + A)\Lambda^{-\frac{1}{2}} x_{in} W \tag{2}$$

where matrix I represents the self-connections and matrix A represents the intra-body connections. Λ is the diagonal matrix of (I+A) and W is the weight matrix of the convolutions, which is trainable. In practice, the spatial part of the GCN is implemented by performing standard 2D convolution and then multiplying the outcome by $\Lambda^{-\frac{1}{2}}(I + A)\Lambda^{-\frac{1}{2}}$.

In the temporal dimension, the skeleton keypoints are connected to themselves. In practice, 2 temporal neighbors (1 before and 1 after time t) are used. Thus, it is straightforward to perform the temporal graph convolutions by modifying the traditional 2D filter-based convolution formulation through use of 1-dimensional filters. Specifically, we use the output of the spatial feature map in equation (2), to perform a convolution in the temporal dimension using a kernel size k_t×1, where k_t is the reception field.

This spatial and temporal sequence of convolution operations constructs a spatial-temporal GCN block, the key component of the GCN. Inspired by the latest GCN-based action recognition models (Yan et al., 2018; Shi et al., 2019a; Wen et al., 2019; Shi et al., 2019b; Si et al., 2019; Li et al., 2019b; Li et al., 2019a), we create our own isolated sign recognition model by stacking several spatial-temporal GCN blocks (Fig. 6).

To improve learning efficiency and minimize overfitting with no extra computational cost, we employ a variation of the spatial GCN called a decoupling GCN, along with a DropGraph layer as in (Cheng et al., 2020). In the decoupling GCN layer, the input features are organized into g groups, where g is a hyperparameter. Features in each g group share 1 trainable adjacent matrix. To use the output features from the decoupling GCN layer, we concatenate the output features from all the g groups. To further improve GCN learning, we introduce a self-attention module that contains a sequence of 3 sub-modules: a spatial, temporal, and channel attention module, as in (Shi et al., 2020). Fig. 5 shows the Spatial-Temporal-Channel (STC) Attention mechanism. Fig. 6 presents the structure of our GCN block.

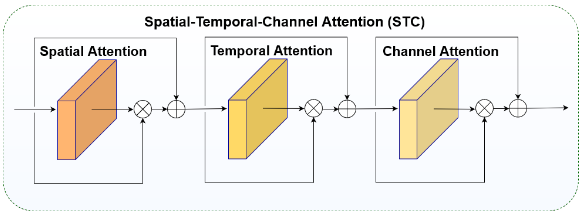

Figure 5: **The STC Attention mechanism**. The 3 sub-modules are arranged in the order: Spatial *before* Temporal *before* Channel attention. The generated attention maps are multiplied with the original feature maps. The residual connection in each attention sub-module is added to stabilize the training.

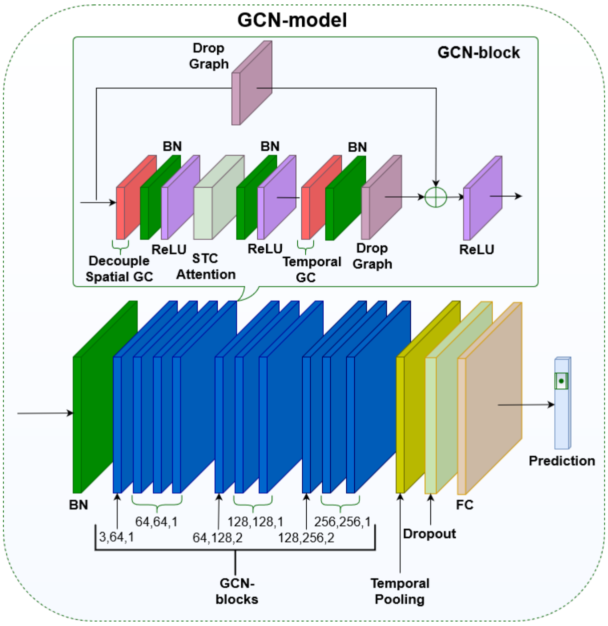

Figure 6: **The GCN model and the Graph Convolutional block in the network**. Each stream has 10 blocks. Each block has a Spatial Graph Convolutional layer, a Temporal Graph Convolutional layer, and a Spatial-Temporal-Channel Attention (STC) module.

Selecting keypoints for Graph Construction. To construct the skeleton graph, 2D keypoints are extracted from a given video frame using a pre-trained network (Fang et al., 2017) that detects 136 face and body keypoints. However, constructing a skeleton graph using all detected keypoints leads to several problems that reduce the recognition rate. This is because the upper body keypoints are more informative than those of the lower body for isolated sign recognition. Because of noise and image blurriness, it can be hard to detect the hand keypoints. The skeleton graph topology can result in keypoint pairs with large distances, leading to inaccurate learning of interactions between keypoints and the skeleton-adjacent matrix.

To overcome these potential problems, we use a subset of the upper body keypoints to construct the skeleton graph, which results in significantly higher recognition rates. More specifically, our skeleton graph consists of 27 nodes corresponding to the nose, eyes, shoulders, elbows, and hands. We use 10 nodes for each hand: the base and tip of each finger. Each node in our graph consists of a (x, y, c) vector, where (x, y) are the 2D coordinates of the corresponding keypoints and c is keypoint detection confidence score.

Bidirectional Stream GCN. Inspired by (Shi et al., 2020), which adopts a multi-stream approach, we use both the forward and backward direction of the video frame sequence to recognize each isolated sign. For each of the 2 streams, we use 6 types of data input: the coordinates of the skeleton keypoints (1^{st}-order information), the distance between consecutive keypoints, the bone vector (2^{nd}-order information), and their motion and acceleration vectors (Fig.7).

To generate the bone vectors for our graph, we set the nose as the root keypoint. Then, we calculate the bone vectors by following the connections of consecutive body skeleton keypoints from the root. Let $v^K_{i,t} = (x_{i,t}, y_{i,t}, c_{i,t})$ and $v^K_{j,t} = (x_{j,t}, y_{j,t}, c_{j,t})$ be 2 ordered, connected keypoints on the skeleton at frame t. Then, the bone vector is calculated as:

$$v^B_{j,t} = v^K_{j,t} - v^K_{i,t},$$

$$v^B_{j,t} = (x_{j,t} - x_{i,t}, y_{j,t} - y_{i,t}, c_{j,t} - c_{i,t}) \, \forall (i,j) \in V \quad (3)$$

where set V contains all the keypoint connections.

The motion streams for the keypoints as well as of the bone vectors are obtained by calculating the difference between their corresponding coordinates in 2 consecutive frames. For example, given a keypoint i at frame t, the keypoint velocity $v^{KV}_{i,t}$ is calculated as:

$$v^{KV}_{i,t} = v^K_{i,t} - v^K_{i,t-1} \, \forall \, i \geq 2 \quad (4)$$

Similarly, for a bone vector, the bone velocity $v^{BV}_{i,t}$ is calculated as:

$$v^{BV}_{i,t} = v^B_{i,t} - v^B_{i,t-1} \, \forall \, i \geq 2 \quad (5)$$

Then, it is straightforward to calculate the keypoints and bone accelerations, using keypoints or bone velocities, respectively:

$$v^{KA}_{i,t} = v^{KV}_{i,t} - v^{KV}_{i,t-1} \, \forall \, i \geq 3 \quad (6)$$

$$v^{BA}_{i,t} = v^{BV}_{i,t} - v^{BV}_{i,t-1} \, \forall \, i \geq 3 \quad (7)$$

Score Fusion. Since our method consists of multiple streams, we need to aggregate the prediction scores from these streams. There are several approaches, such as fusing multiple features obtained from middle layers in streams (middle fusion) (Li et al., 2019c; Hong et al., 2020; Tatulli and Hueber, 2017), or fusing multiple probabilities obtained from the last layers (late fusion) (Shi et al., 2019b; Shi et al., 2019a; Shi et al., 2020; Cai et al., 2021). For isolated sign recognition, the forward and backward sign video streams are adapted for a late fusion approach. First, the probability in both the forward and the backward stream is calculated as a weighted summation of the output scores of the 6 correlated streams. Next, we obtain the sign scores and predict the sign label by assigning weights, and we sum the results from the forward and backward streams.

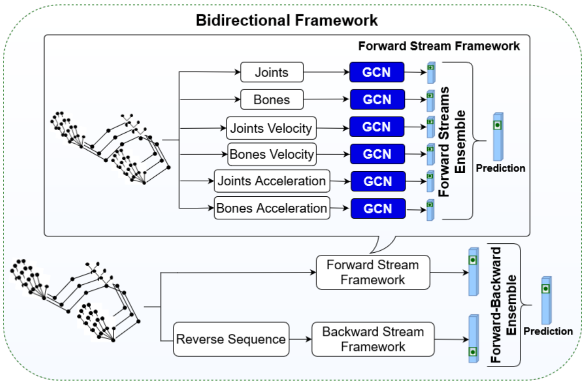

Figure 7: Illustration of our bidirectional framework.

4. Experiments

The proposed bidirectional GCN-based framework was tested for isolated sign recognition on the modified WLASL dataset (with corrections to gloss labeling). We compare our method with an RGB-based approach as a baseline model (Zhou et al., 2018) and with a GCN-based approach that uses skeleton data for sign recognition (Jiang et al., 2021). First, we describe how we extract the skeleton data. Then we compare our forward and backward stream-based learning approach with the other methods on the modified WLASL dataset. We also compute sign start and end frames, and use transfer learning from the AUTSL (Sincan and Keles, 2020) and SLR500 (Zhang et al., 2016) datasets. In addition, we provide ablation studies to demonstrate the improvements achieved through the detection of start and end frames and use of transfer learning.

4.1. The WLASL Dataset

The WLASL dataset (Li et al., 2020) is a collection of isolated ASL sign videos taken from various sources that includes 119 signers and 2,000 distinct signs. It is an imbalanced dataset, consisting of 21,083 videos with unconstrained recording conditions. However, we have found that there are inconsistencies in the gloss labels associated with the signs. Thus, we modified the gloss labeling to enforce consistency (as discussed in Section 3.1). The modified dataset consists of 18,141 videos for 1449 lexical signs.

We split the modified dataset following (Li et al., 2020) into training, validation, and testing sets using a ratio of 4:1:1 for the samples of each sign. We evaluate the recognition performance and report the per-instance Top-1 and Top-5 sign recognition accuracy.

4.2. Skeleton Extraction-Data Preparation

We use the pre-trained model in (Fang et al., 2017) to estimate 136 keypoints of the whole body (torso, face, hands, and legs) from the RGB video frames and construct our skeleton graph of 27 nodes. First we normalize the keypoint coordinates to [-1,1], and then we apply the following data augmentation techniques: (a) random sampling, (b) mirroring, (c) rotation, (d) scaling, and (e) shifting. Since each video contains multiple frames with different lengths, the length of all the videos is aligned to 150 frames. If a video has more than 150 frames, the first 150 frames are extracted from the video. If a video has fewer than 150 frames, we repeat the frame sequence until the video length is 150 frames. Moreover, we use the information about the detected start and end frames of each sign video to redefine our input.

4.3. Performance of the Bidirectional GCN-based Framework

The Top-1 and Top-5 recognition performance of the proposed forward and backward GCN-based framework is reported in Table 1 and Table 2.

Forward Streams	WLASL	
	Top-1	Top-5
Keypoint	74.52%	92.41%
Bone	68.95%	90.02%
Keypoint Velocity	56.53%	78.87%
Bone Velocity	53.40%	76.56%
Keypoint Acceleration	45.73%	68.92%
Bone Acceleration	45.32%	68.27%
Multi-stream	**76.75%**	**94.18%**

Table 1: Forward stream sign recognition accuracy.

Backward Streams	WLASL	
	Top-1	Top-5
Keypoint	73.87%	92.44%
Bone	68.22%	88.66%
Keypoint Velocity	56.93%	79.58%
Bone Velocity	51.88%	76.59%
Keypoint Acceleration	46.08%	69.98%
Bone Acceleration	44.21%	68.03%
Multi-stream	**75.75%**	**93.88%**

Table 2: Backward stream sign recognition accuracy.

Of the streams for which there is both forward and backward information, the keypoint stream provides the best accuracy. The score fusion approach for the forward and backward models further improves the overall recognition rate. Table 3 shows the score fusion of the forward and backward models. The bidirectional approach results in higher performance compared to using only the forward or backward stream. This demonstrates the advantage of using the proposed fusion approach.

Streams	WLASL	
	Top-1	Top-5
Forward Stream	76.75%	94.18%
Backward Stream	75.75%	93.88%
Bidirectional stream	**77.43%**	**94.54%**

Table 3: Fused bidirectional framework performance.

In Table 4 we report the performance of our proposed framework compared to the RGB-based model (TRN) in (Zhou et al., 2018) and the GCN-based model in (Jiang et al., 2021) on the same modified WLASL dataset. We use pre-training on the AUTSL dataset (Sincan and Keles, 2020) to improve sign recognition. Our proposed late fusion model raises the recognition rate by 6.39%. Top-1 accuracy is improved to 77.43%. Compared with other state-of-the-art methods, our proposed method achieves the best recognition accuracy.

Method	WLASL	
	Top-1	Top-5
TRN (Zhou et al., 2018)	49.32%	77.91%
SL-GCN (SAM-SLR-v2) (Jiang et al., 2021)	71.04%	91.44%
Ours	**77.43%**	**94.54%**

Table 4: Comparison on the modified WLASL dataset.

In Table 5, we show the classification accuracy resulting from use of keypoints, velocities, accelerations, and their combinations in forward and backward data streams. Keypoints generally contribute more to the accuracy than the velocity streams.

Tables 6 and 7 present ablation studies. For transfer learning, pre-training on the Turkish sign language dataset AUTSL (Sincan and Keles, 2020) results in higher accuracy than pre-training on the Chinese sign language dataset SLR500 (Zhang et al., 2016). These 2 datasets have similar characteristics, although the AUTSL dataset consists of half the number of signs as are in the SLR500 dataset. This leads to the conclusion that, to achieve higher recognition rates, pre-training our framework using the AUTSL dataset is more efficient than pre-training on the SLR500 dataset. Furthermore, using the detected start and end frames results in higher accuracy than using raw data.

4.4. Training Details

The experiments were conducted using Pytorch 1.7.0 and 1 NVIDIA Quadro RTX8000. To train the GCN models in each stream (forward and backward), a Cross-Entropy loss function is used, and the weight decay is set to 0.0001. The Stochastic Gradient Descent (SGD) with Nesterov Momentum is selected as the optimization algorithm; the momentum is set to 0.9. The learning rate is initially set to 0.1 and divided by 10 when 150 and 200 epochs are reached. The total number of epochs used for training our models is 300. The batch size for both the training and testing processes is set to 64. We randomly select 64 videos during training as input in an iteration. Moreover, we ensure that all the videos are used in an epoch for training.

Forward Data Stream						Backward Data Stream						(%)	
K	B	K-V	B-V	K-A	B-A	K	B	K-V	B-V	K-A	B-A	Top-1	Top-5
✓	✓											76.21	93.94
		✓	✓									60.39	82.68
				✓	✓							52.15	75.77
✓	✓	✓	✓	✓	✓							**76.75**	**94.18**
						✓	✓					75.07	93.45
								✓	✓			60.63	83.74
										✓	✓	51.58	75.61
						✓	✓	✓	✓	✓	✓	**75.75**	**93.88**
✓	✓					✓	✓					77.19	94.29
✓	✓	✓	✓			✓	✓	✓	✓			77.35	94.67
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**77.43**	**94.54**

Table 5: Comparison of the GCN Accuracy Resulting from Use of Combinations of Various Types of Data, using Forward and Backward Streams – for recognition of the 1,449 lexical signs in the modified WLASL dataset. Abbreviations: K=Keypoints; B=Bone; K-V=Keypoint Velocity; B-V=Bone Velocity; K-A=Keypoint Acceleration; B-A=Bone Acceleration.

Transfer Learning	WLASL	
	Top-1	Top-5
SLR500	73.65%	92.36%
AUTSL	74.52%	92.41%

Table 6: Performance of the forward data stream using transfer learning from SLR500 and AUTSL datasets.

Start/end frames	WLASL	
	Top-1	Top-5
Yes	74.52%	92.41%
No	68.95%	87.82%

Table 7: Performance of the forward data stream with and without detection of sign start and end frames.

5. Conclusion

We propose here a bidirectional GCN-based framework for accurate isolated sign recognition. Some key methodological aspects of this approach are: use of a dataset with enforced consistency of text-based gloss labeling of signs; pre-training, to leverage transfer learning; use of start and end frame information for selecting the input to our framework; and the cooperative use of forward and backward data streams (including various sub-streams), for recognition of isolated signs. The proposed framework outperforms the state-of-the-art methods in isolated sign recognition on the challenging WLASL dataset (with modifications of gloss labeling), as shown in Fig. 8. The ablation studies demonstrate the effectiveness of representing both forward and backward relations of intra-body keypoints over time. Future research will explore other fusion methods for the forward and backward streams, and exploitation of statistical information that reflects linguistic constraints governing the relationships between the 2 hands and between the start and end frames of lexical signs, which has been demonstrated to improve recognition accuracy (Thangali et al., 2011; Dilsizian et al., 2014).

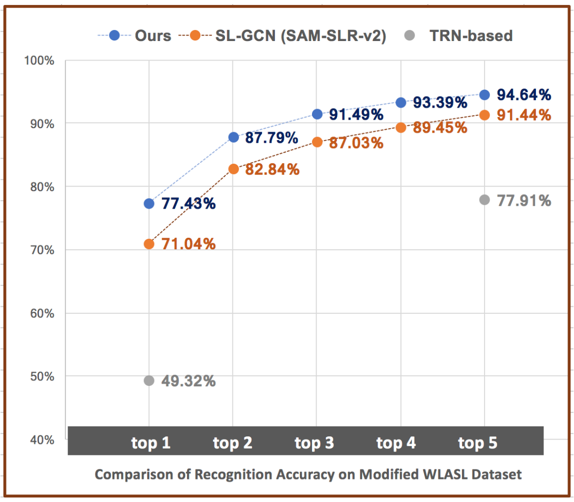

Figure 8: Comparison of recognition accuracy of our GCN-based method with another GCN-based method and with a CNN-based method.

6. Acknowledgments

For invaluable assistance with this project, we thank: Matt Huenerfauth, Augustine Opoku, Carey Ballard, Indya-loreal Oliver, Lutece Dekker, Qilong Zhangli, Gregory Dimitriadis, and Douglas Motto. We also gratefully acknowledge Stan Sclaroff, Vassilis Athitsos, Joan Nash, Ben Bahan, Rachel Benedict, Naomi Casselli, Elizabeth Cassidy, Lana Cook, Braden Painter, Tyler Richard, Tory Sampson, and Dana Schlang for collaboration on the development of the ASLLVD dataset. The work reported here has been supported in part by grants from NSF: no. 2040638, 1763486, 1763523, and 1763569. Any opinions, findings, or conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. Bibliographical References

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., and Thangali, A. (2008). The American Sign Language Lexicon Video Dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE.

Badhe, P. C. and Kulkarni, V. (2015). Indian sign language translator using gesture recognition algorithm. In *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, pages 195–200. IEEE.

Bertasius, G., Wang, H., and Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? *arXiv preprint arXiv:2102.05095*.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31.

Cai, J., Jiang, N., Han, X., Jia, K., and Lu, J. (2021). JOLO-GCN: mining joint-centered light-weight information for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2735–2744.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., and Lu, H. (2020). Decoupling GCN with DropGraph Module for Skeleton-Based Action Recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 536–553. Springer.

Cooper, H., Ong, E.-J., Pugeault, N., and Bowden, R. (2012). Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13:2205–2231.

Cui, R., Liu, H., and Zhang, C. (2019). A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891.

Dardas, N. H. and Georganas, N. D. (2011). Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Transactions on Instrumentation and measurement*, 60(11):3592–3607.

de Amorim, C. C., Macêdo, D., and Zanchettin, C. (2019). Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition. In *International Conference on Artificial Neural Networks*, pages 646–657. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Dilsizian, M., Yanovich, P., Wang, S., Neidle, C., and Metaxas, D. (2014). A new framework for sign language recognition based on 3D handshape identification and linguistic modeling. In *9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 1924–1929. European Language Resources Association (ELRA).

Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). RMPE: Regional Multi-person Pose Estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343.

Fatmi, R., Rashad, S., Integlia, R., and Hutchison, G. (2017). American Sign Language Recognition using Hidden Markov Models and Wearable Motion Sensors. *Trans. Mach. Learn. Data Min.*, 10(2):41–55.

Grobel, K. and Assan, M. (1997). Isolated sign language recognition using Hidden Markov Models. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 1, pages 162–167 vol.1.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., and Zhang, B. (2020). More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4340–4354.

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., and Fu, Y. (2021). Sign Language Recognition via Skeleton-Aware Multi-Model Ensemble. *arXiv e-prints*, pages arXiv–2110.

Kalfaoglu, M. E., Kalkan, S., and Alatan, A. A. (2020). Late Temporal Modeling in 3D CNN architectures with BERT for Action Recognition. In *European Conference on Computer Vision*, pages 731–747. Springer.

Koller, O., Camgoz, N. C., Ney, H., and Bowden, R. (2019). Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2306–2320.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data.

Li, B., Li, X., Zhang, Z., and Wu, F. (2019a). Spatio-Temporal Graph Routing for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8561–8568.

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., and Tian, Q. (2019b). Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603.

Li, Z., Huang, L., and He, J. (2019c). A multiscale deep middle-level feature fusion network for hyperspectral classification. *Remote Sensing*, 11(6):695.

Li, D., Rodriguez, C., Yu, X., and Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469.

Liang, Z.-j., Liao, S.-b., and Hu, B.-z. (2018). 3D Convolutional Neural Networks for Dynamic Sign Language Recognition. *The Computer Journal*, 61(11):1724–1736.

Lim, K. M., Tan, A. W. C., Lee, C. P., and Tan, S. C. (2019). Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. *Multimedia Tools and Applications*, 78(14):19917–19944.

Lin, F. R., Niparko, J. K., and Ferrucci, L. (2011). Hearing loss prevalence in the United States. *Archives of internal medicine*, 171(20):1851–1853.

Liu, H., Tu, J., and Liu, M. (2017). Two-Stream 3D Convolutional Neural Network for Skeleton-Based Action Recognition. *arXiv preprint arXiv:1705.08106*.

Looney, D. and Lusin, N. (2019). Enrollments in Languages Other Than English in United States Institutions of Higher Education, Summer 2016 and Fall 2016 Report. *MLA Web Publication. New York: MLA*.

Masood, S., Srivastava, A., Thuwal, H. C., and Ahmad, M. (2018). Real-time sign language gesture (word) recognition from video sequences using CNN and RNN. In *Intelligent Engineering Informatics*, pages 623–632. Springer.

Memiş, A. and Albayrak, S. (2013). A kinect based sign language recognition system using spatio-temporal features. In *Sixth International Conference on Machine Vision (ICMV 2013)*, volume 9067, page 90670X. International Society for Optics and Photonics.

Metaxas, D., Dilsizian, M., and Neidle, C. (2018). Linguistically-driven framework for computationally efficient and scalable sign recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mitchell, R. E., Young, T. A., Bachelda, B., and Karchmer, M. A. (2006). How many people use ASL in the United States? Why estimates need updating. *Sign Language Studies*, 6(3):306–335.

Neidle, C. and Opoku, A. (2021). Update on Linguistically Annotated ASL Video Data Available through the American Sign Language Linguistic Research Project (ASLLRP). June.

Neidle, C., Thangali, A., and Sclaroff, S. (2012). Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus. In *5th workshop on the representation and processing of sign languages: interactions between corpus and Lexicon, LREC*. Citeseer.

Neidle, C., Opoku, A., Dimitriadis, G., and Metaxas, D. (2018). NEW Shared Interconnected ASL Resources: SignStream® 3 Software; DAI 2 for Web Access to Linguistically Annotated Video Corpora; and a Sign Bank. *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, Miyazaki, Language Resources and Evaluation Conference 2018*.

Papastratis, I., Dimitropoulos, K., Konstantinidis, D., and Daras, P. (2020). Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space. *IEEE Access*, 8:91170–91180.

Rastgoo, R., Kiani, K., and Escalera, S. (2020). Video-based isolated hand sign language recognition using a deep cascaded model. *Multimedia Tools and Applications*, 79:22965–22987.

Rastgoo, R., Kiani, K., and Escalera, S. (2021). ZS-SLR: Zero-Shot Sign Language Recognition from RGB-D Videos. *arXiv preprint arXiv:2108.10059*.

Redmon, J. and Farhadi, A. (2018). YOLOv3: An Incremental Improvement. *CoRR*, abs/1804.02767.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019a). Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019b). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12026–12035.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2020). Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545.

Si, C., Chen, W., Wang, W., Wang, L., and Tan, T. (2019). An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1227–1236.

Sincan, O. M. and Keles, H. Y. (2020). AUTSL: A Large Scale Multi-modal Turkish Sign Language Dataset and Baseline Methods. *IEEE Access*, 8:181340–181355.

Sincan, O. M. and Keles, H. Y. (2021). Using Motion History Images with 3D Convolutional Networks in Isolated Sign Language Recognition. *arXiv preprint arXiv:2110.12396*.

Sincan, O. M., Tur, A. O., and Keles, H. Y. (2019). Isolated sign language recognition with multi-scale features using LSTM. In *2019 27th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Soo Kim, T. and Reiter, A. (2017). Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28.

Tatulli, E. and Hueber, T. (2017). Feature extraction using multimodal convolutional neural networks for visual speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2971–2975. IEEE.

Thangali, A., Nash, J. P., Sclaroff, S., and Neidle, C. (2011). Exploiting phonological constraints for handshape inference in ASL video. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 521–528, June.

Tornay, S., Aran, O., and Doss, M. M. (2020). An HMM Approach with Inherent Model Selection for Sign Language and Gesture Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6049–6056.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning Spatiotemporal Features with 3C Convolutional Networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Vaezi Joze, H. and Koller, O. (2019). MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. In *The British Machine Vision Conference (BMVC)*, September.

Wen, Y.-H., Gao, L., Fu, H., Zhang, F.-L., and Xia, S. (2019). Graph CNNs with motif and variable temporal block for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8989–8996.

Xiaohan Nie, B., Xiong, C., and Zhu, S.-C. (2015). Joint action recognition and pose estimation from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1293–1301.

Xie, S., Sun, C., Huang, J., Tu, Z., and Murphy, K. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Thirty-second AAAI conference on artificial intelligence*.

Yang, Q. (2010). Chinese sign language recognition based on video sequence appearance modeling. In *2010 5th IEEE Conference on Industrial Electronics and Applications*, pages 1537–1542. IEEE.

Zhang, J., Zhou, W., Xie, C., Pu, J., and Li, H. (2016). Chinese sign language recognition with adaptive HMM. In *2016 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE.

Zhou, B., Andonian, A., Oliva, A., and Torralba, A. (2018). Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818.