Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT 7), pages 13–20
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Isolated Sign Recognition using ASL Datasets with Consistent Text-based Gloss Labeling and Curriculum Learning

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

Abstract

We present a new approach for isolated sign recognition, which combines a spatial-temporal Graph Convolution Network (GCN) architecture for modeling human skeleton keypoints with late fusion of both the forward and backward video streams, and we explore the use of curriculum learning. We employ a type of curriculum learning that dynamically estimates, during training, the order of difficulty of each input video for sign recognition; this involves learning a new family of *data parameters* that are dynamically updated during training. The research makes use of a large combined video dataset for American Sign Language (ASL), including data from both the American Sign Language Lexicon Video Dataset (ASLLVD) and the Word-Level American Sign Language (WLASL) dataset, with modified gloss labeling of the latter—to ensure 1-1 correspondence between gloss labels and distinct sign productions, as well as consistency in gloss labeling across the two datasets. This is the first time that these two datasets have been used in combination for isolated sign recognition research. We also compare the sign recognition performance on several different subsets of the combined dataset, varying in, e.g., the minimum number of samples per sign (and therefore also in the total number of sign classes and video examples).

Keywords: ASL, Isolated Sign Recognition, Curriculum Learning, ASLLVD, WLASL

1. Introduction

There are >70 million deaf people worldwide, and >200 signed languages (World Federation of the Deaf, 2022). In the US, there are 28 million Deaf or Hard-of-Hearing people (Lin et al., 2011), and ASL is the primary language for an estimated 500,000 (or more) (Mitchell et al., 2006). Signed languages like ASL are full-fledged natural languages, but they are structurally distinct from spoken languages. Language in the visual modality involves movements of the hands and arms, as well as facial expressions and movements of the head and upper body. ASL has no standard written form.

Computer-based research on sign recognition from video will pave the way for technologies to benefit the Deaf community and to improve communication between deaf and hearing individuals, such as ASL-to-English translation, for which sign recognition is a precursor; or educational applications to support ASL learners. It will also enable development of a variety of computational tools for signers, such as Google-like sign search by example over videos on the Web.

However, this is a difficult problem, and research in this area is badly needed. Here we focus on recognition of isolated, citation-form signs. Sign recognition from continuous signing is a related but more complex problem. As with any other natural language, there is considerable variability in the production of signs in ASL, which poses a challenge for sign recognition. Progress in this area requires the availability of large, linguistically annotated, video datasets with consistent gloss labeling of signs, and with representation of many and diverse signers and a sufficient number of samples per sign, to serve as a basis for computer learning.

1.1. Issues related to Data

As observed in Dafnis et al. (2022) Neidle et al. (2022a), and Neidle and Ballard (2022), the Word-Level ASL (WLASL) video dataset (Li et al., 2020)—which is potentially valuable for sign recognition research in that it brings together multiple publicly shared ASL video datasets—is problematic in one critical respect: there is no enforced 1-1 correspondence between gloss labels and sign productions. Figure 1 illustrates the problem with using the WLASL gloss labels as "ground truth" for sign recognition research. Each of the ASL signs shown in this figure—one glossed as "A-LOT," the other as "MANY" in our ASLLRP Sign Bank (Neidle et al., 2022b), `https://dai.cs.rutgers.edu/dai/s/signbank`)—has several different gloss labels within the WLASL dataset, whereas particular gloss labels, such as "a lot" or "numerous," are used for totally different ASL signs.

For this reason, we have created, and shared publicly `http://dev.dai.cs.rutgers.edu/dai/s/aboutwlasl`, a spreadsheet that provides, for a large subset of the WLASL videos, gloss labels consistent with those used for the ASLLRP Sign Bank, where such 1-1 correspondences are enforced. This makes it possible to take advantage of the large and varied set of WLASL video files while ensuring internally consistent gloss labeling; this is precisely what was done in Dafnis et al. (2022).

Moreover, this also makes it possible to combine the WLASL and ASLLRP isolated sign datasets (of which

*Equal contribution.

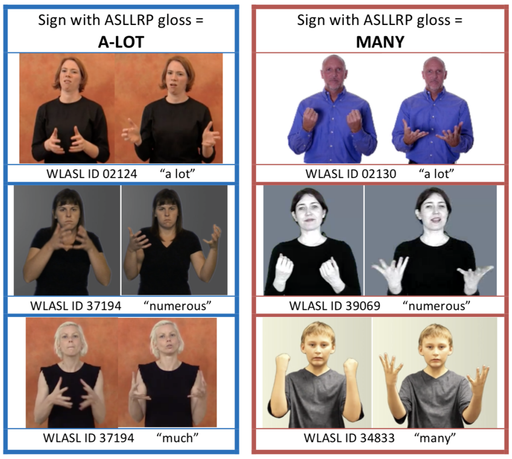

Figure 1: Inconsistent WLASL gloss labels: examples

the American Sign Language Lexicon Video Dataset (ASLLVD) (Athitsos et al., 2010; Neidle et al., 2012) is a part), with consistent gloss labeling across both, giving rise to a combined dataset larger and richer than either of the two. That is what we have done here.

The ASLLRP datasets include, for each sign: gloss labels (main entry plus variant labels); annotations of the linguistic start and end frames; start and end handshapes for each hand (in 1- and 2-handed signs); and sign type categorization (e.g., lexical, fingerspelled, loan sign, classifier, compound, etc.).

The current research relies on using both the ASLLVD and WLASL datasets in combination. In experiments to be reported on below, we used (1) lexical signs merged from both collections for which we had at least 6 or 12 examples per sign; and (2) these same datasets expanded to include not only lexical signs, but also loan signs and compounds, for which we had at least 6 or 12 examples per sign from the merged datasets. Complete details of the datasets used for each of these experiments are available from our website: http://www.bu.edu/asllrp/signrec.html.

1.2. Overview of our Approach

Our isolated sign recognition approach uses a spatial-temporal Graph Convolution Network (GCN) architecture for modeling human skeleton keypoints, with late fusion of forward and backward video streams, as in Dafnis et al. (2022). We also explore curriculum learning: dynamic estimation, during training, of the order of the difficulty of input videos for sign recognition; this involves learning a new family of parameters using a differentiable curriculum.

2. Related Work

Early research on isolated sign recognition from video, as well as more recent work (Cooper et al., 2012; Badhe and Kulkarni, 2015; Tamura and Kawasaki, 1988; Xiaohan Nie et al., 2015; Tornay et al., 2020), uses either color thresholding for feature extraction or hand-crafted features, such as hand positions, movement, location, and distances between the hands and specific body parts, in conjunction with classifiers, such as SVMs, KNNs, CRFs and HMMs (Memiş and Albayrak, 2013; Dardas and Georganas, 2011; Yang, 2010; Metaxas et al., 2018; Tornay et al., 2020). However, these features and the distribution assumptions inherent to these approaches result in systems with limited capability for generalization.

2.1. RGB-based Approaches

Over the past decade, most of this research shifted toward end-to-end deep learning methods, spurred by the success for computer vision problems of Convolutional Neural Networks (CNNs) in extracting spatial features and of Recurrent Neural Networks (RNNs) in capturing temporal information. Promising initial results were achieved in the domain of sign language recognition using CNN-based end-to-end deep learning methods, e.g., Pigou et al. (2016), which uses a 2D CNN for sign recognition of Flemish Sign Language (VGT) and Dutch Sign Language (NGT).

Later, many researchers leveraged modified CNNs (3D-CNN) in the context of sign and action recognition. For example, Li et al. (2020), who introduced the WLASL for isolated sign recognition, compare 4 different deep-learning architectures: 2 RGB-based and 2 pose-based approaches. The pose-based networks use body keypoints extracted using OpenPose (Cao et al., 2019; Simon et al., 2017) as input. These methods include a 2D-CNN in conjunction with an RNN, a pose-based RNN, a 3D-CNN, and a pose-based Temporal GCN. The authors show that the 3D-CNN outperforms the other approaches. While the 3D-CNN model performs better than previous approaches in learning short-term memory dependencies, a major drawback is that it restricts the learning of long-term dependencies at the final temporal global average pooling stage.

Recent architectures exploit the self-attention mechanism of Transformers for video understanding (Bertasius et al., 2021). De Coster et al. (2020) use a 2D-CNN and a Video Transformer Network for isolated sign recognition; they use the self-attention encoder layers without masking, while they remove the cross-attention decoder, and their results are promising.

2.2. Skeleton-based Approaches

Instead of using RGB frames as input, some methods, such as those mentioned in Li et al. (2020), use body keypoints to focus the learning procedure on the relevant information. When the off-the-shelf pretrained human pose estimation systems are robust, these methods show good performance in both learning and recognition, as the recognition models are not affected by irrelevant information from the background.

Early research on action and sign recognition used pose-based CNNs, followed by an RNN for the relevant temporal information (Soo Kim and Reiter, 2017; Liu et al., 2017). However, a disadvantage of these models is that they cannot encode information about keypoint interactions in both space and time. In order to overcome this disadvantage, Yan et al. (2018) proposed a Spatial-Temporal Graph Convolutional Network (ST-GCN) and showed the effectiveness of GCNs for learning spatiotemporal skeleton dynamics. Shi et al. (2019b) exploited a 2-stream approach using both keypoints and bone information, while Shi et al. (2020) proposed a 4-stream approach in which bones and the motion of keypoints are added. Their approach resulted in improved action recognition. de Amorim et al. (2019) used an extension of the ST-GCN model for isolated sign recognition and achieved close to 60% accuracy on a vocabulary of 20 signs. Jiang et al. (2021) used a pose-based GCN approach, as in Shi et al. (2020), in conjunction with other modalities, such as RGB frames, optical flow, and depth video. Their proposed GCN was the first successful attempt to tackle isolated sign recognition using body skeleton graphs.

In Dafnis et al. (2022), we follow a similar GCN approach, with the addition of forward and backward data streams and use of the acceleration of keypoints and bones. This improved isolated sign recognition on 1,449 lexical signs from the WLASL dataset, with glosses modified as discussed in Section 1.2.

2.3. Curriculum Learning Approaches

Curriculum learning is a ”strategy that trains a machine learning model from easier data to harder data, which imitates the meaningful learning order in human curricula” (Wang et al., 2021). Curriculum learning was introduced by Bengio et al. (2009), who proved that training a neural network starting with easy examples and gradually increasing the difficulty of the data provides significant improvement to the overall accuracy and convergence of the model. The inspiration was derived from the way humans learn best: starting with easier concepts and gradually increasing complexity, rather than randomly learning different concepts.

However, deciding which samples to categorize as easy or hard is not trivial. Much research has been conducted on how to define which data samples to consider easy or difficult (e.g., Hacohen and Weinshall (2019), Weinshall et al. (2018), Wu et al. (2020), Zhou et al. (2020)). In this work, the order of difficulty is defined before training; the most common techniques are 1) to use pretrained models on the examined dataset; and 2) to create annotations, which could be time-consuming. Those techniques are task-specific and non generalizable. As a result, curriculum learning research later focused on finding a way to estimate the importance (or weight) of each sample directly during training, based on the observation that easy and hard samples behave differently and can therefore be separated.

The first step in this direction was taken by Kumar et al. (2010), who proposed a dynamic way to apply curriculum learning using the idea of self-paced learning. Instead of using a predefined order of difficulty of the samples, this method dynamically determines this order by feedback from the learner itself. Inspired by this idea, many classification tasks were further improved, since curriculum learning provided a quicker and better convergence (Cascante-Bonilla et al., 2020; Pi et al., 2016; Zhao et al., 2015; Saxena et al., 2019).

3. Technical Approach

The key aspects of our approach include a spatial-temporal GCN architecture for modeling the skeleton keypoints; dynamic estimation during training of the order of difficulty of each input video for sign recognition by learning a new family of data parameters using a differentiable curriculum; and a late ensemble method that fuses both the forward and backward video streams, as in Dafnis et al. (2022).

Section 3.1 presents our deep-learning model for isolated sign recognition based on skeleton keypoints. Our ensemble data fusion method is explained in 3.2. Section 3.3 then introduces the data parameters that we use for learning a differentiable curriculum and the training strategy we follow based on curriculum learning.

3.1. Sign Recognition Model

As mentioned in Section 2, previous studies on isolated sign recognition have revealed that spatial-temporal graph architectures, in conjunction with a self-attention mechanism, can boost recognition accuracy. Hence, we use a spatial-temporal GCN model similar to Jiang et al. (2021) and Dafnis et al. (2022) for isolated sign recognition on the reported dataset, as presented below.

GCN for human skeleton keypoints. Our adopted spatial-temporal GCN learning approach consists of 10 basic GCN blocks; see Figure 2. Each basic block consists of a sequence of Decoupled Spatial Graph Convolutional layers (Decoupled SGCNs) (Cheng et al., 2020), a cascaded spatial-temporal-channel attention mechanism (Shi et al., 2020), and a Temporal Convolutional layer (TCN). The Decoupled SGCN helps our GCN model boost its capacity with no extra cost. In addition, a DropGraph layer as in Cheng et al. (2020) is added. This module helps to avoid overfitting. At the end, we apply a global average pooling on both the spatial dimensions (within a skeleton) and the temporal dimensions (across skeletons), along with a dropout before a fully-connected layer for recognition.

Spatial-Temporal Graph Convolution. We first present the spatial convolution operations within a skeleton graph. To define the graph convolution in the spatial dimension for our human skeleton graph, we follow Yan et al. (2018). The implementation of the spatial part of the GCN is expressed as follows:

$$u_{out} = D^{-\frac{1}{2}}(I+A)D^{-\frac{1}{2}}u_{in}W, \tag{1}$$

where matrices A and I represent the intra-body and self-connections respectively. D is the diagonal matrix of (I+A), while W represents the weight matrix of the convolutions. In practice, the spatial graph convolution operation is implemented by performing standard 2D convolution and then multiplying the outcome by $D^{-\frac{1}{2}}(I+A)D^{-\frac{1}{2}}$.

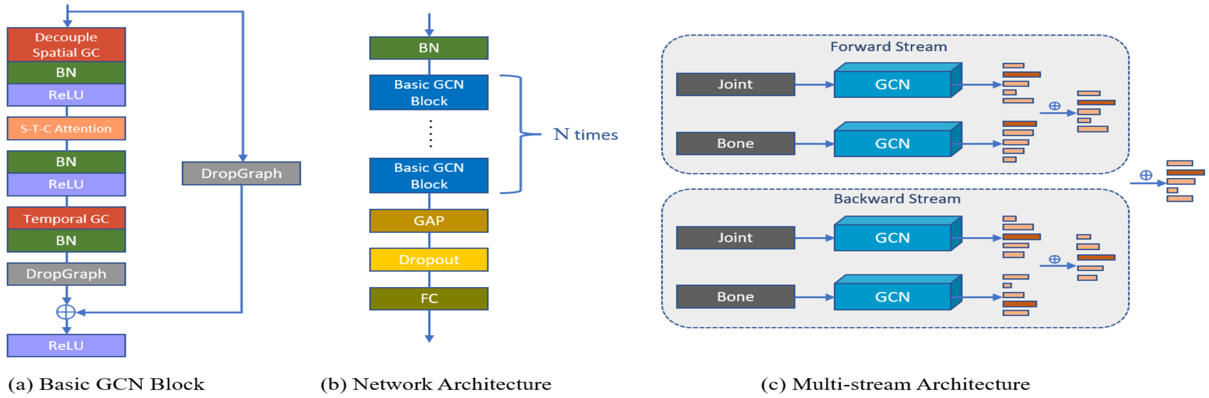

Figure 2: Illustration of the GCN pipeline: (a) Basic GCN block architecture; (b) GCN architecture. There are 10 basic GCN blocks in all. GAP represents the global average pooling layer and FC the fully connected layer. (c) The overall architecture of the Multi-stream GCN. The forward and backward scores are fused using weighted summation to obtain the final prediction.

To capture the temporal relationships among skeleton graphs in consecutive frames, we use temporal convolutions. These temporal graph convolution operations can be also expressed as a standard 2D convolution using a kernel size k_t×1, where k_t is the reception field. In practice, the human skeleton keypoints are connected to themselves in the temporal dimension. Thus, the traditional 2D convolution formulation is modified to a 1-dimensional convolution.

Spatial Graph Construction. To construct the skeleton graph, we extract 2D skeleton keypoints using Alphapose (Fang et al., 2017), a pretrained model that extracts 136 face and body keypoints from a given video frame. However, using all 136 keypoints for isolated sign recognition reduces the recognition rate. This is because the upper body keypoints are more informative than those of the lower body for sign recognition. In addition, because of blurriness during hand movements, it can be hard for the 2D skeleton extractor to detect the hand keypoints accurately. To overcome these issues, following Jiang et al. (2021) and Dafnis et al. (2022), we reduce the number of skeleton keypoints used for skeleton graph construction. Our graph consists of 27 nodes corresponding to 10 keypoints for each hand and 7 upper body keypoints: nose, eyes, shoulders, elbows. The 10 hand keypoints correspond to the base and tip of each finger. Given the variability in lexically related mouthing, and our current sample sizes, we did not include keypoints around the mouth on the graph. We found that including them did not increase accuracy, but we hope to incorporate this in the future. Each node on our graph has a (x, y, c) vector, where (x, y) are the 2D coordinates of the corresponding keypoints and c is the keypoint detection confidence score.

Forward and Backward Sign Recognition. Following Dafnis et al. (2022), we use both the forward and backward directions of the video data for isolated sign recognition. In each direction, we use two types of data streams as input: the human skeleton keypoint (joint) coordinates, and the bone vector (distance between keypoints). As demonstrated in Dafnis et al. (2022), these two streams are the most informative for isolated sign recognition since, because of noise in the estimation of joint locations, the joint velocities and acceleration vectors are not reliable.

We generate the bone vectors for our graph by setting the nose as the root keypoint on the skeleton graph. Let two ordered connected keypoints $v^K_{i,t}$, $v^K_{j,t}$ at frame t, with coordinates $v^K_{i,t} = (x_{i,t}, y_{i,t}, c_{i,t})$ and $v^K_{j,t} = (x_{j,t}, y_{j,t}, c_{j,t})$ respectively. Then the bone vector is computed as:

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

$$v^B_{j,t} = (x_{j,t} - x_{i,t}, y_{j,t} - y_{i,t}, c_{j,t} - c_{i,t})\ \forall (i, j) \in V, \quad (2)$$

where V contains all skeleton keypoint connections.

3.2. Score Fusion

In both the forward and backward directions, our framework uses multiple streams of information (i.e., joints and bones) to make aggregate predictions for each direction. We first fuse the prediction scores from all streams in each direction. We use the respective *softmax* scores in each stream (Shi et al., 2019b; Shi et al., 2019a; Shi et al., 2020; Cai et al., 2021; Dafnis et al., 2022) to compute an optimized weighted summation of the scores for each direction. We then fuse the prediction softmax scores for each direction by computing an optimized weighted summation that produces a prediction of the sign labels.

3.3. Curriculum Learning

To further enhance our recognition accuracy, we use a type of curriculum learning introduced in Saxena et al. (2019), which dynamically estimates during training the order of difficulty of each input video for sign recognition by using a new family of trainable parameters for deep neural networks called *data parameters*.

Each sign class and each sign instance are assigned *data parameters*, which are updated after every iteration during training. The respective learning process determines which sign samples and classes need more attention compared to the others to improve sign recognition automatically, as follows: We define

$$\{(x^i, y^i)\}^N_{n=1}, \quad (3)$$

where x^i is a data sample (a video of a sign) that is input to the neural network, y^i is the label of x^i, and N represents the number of input samples. The neural network is defined as f_θ, and the logits are z^i, i.e., $f_\theta(x^i) = z^i$. We also define the data parameter ϕ_i^* as the sum of the instance and class parameters as follows:

$$\phi_i^* = \phi_{y_i}^{class} + \phi_i^{instance} \tag{4}$$

We use the cross entropy loss as the loss function, where the logits are scaled using the data parameter ϕ_i^*:

$$L^i = -log(p_{y^i}^i), \tag{5}$$

where

$$p_{y^i}^i = \frac{exp(z_{y^i}^i/\phi_i^*)}{\Sigma_j exp(z_j^i/\phi_i^*}). \tag{6}$$

L^i is the cross entropy, ϕ_i^* is the data parameter, $z_{y^i}^i$ is the logit and $p_{y^i}^i$ is the probability of the target class y^i for sample x_i. In order to estimate the sign class given an instance we need to minimize L^i:

$$\min_{\theta,\phi^*} \frac{1}{N} \Sigma_{i=1}^N L^i \tag{7}$$

During training, the class parameters, $\phi_{y_i}^{class}$, take into account the average of the gradients from all the class samples in each mini-batch, while instance parameters, $\phi_i^{instance}$, aggregate the gradients from each individual sample. This process has the following advantages:

1) Some videos in our dataset are of low resolution and, as a consequence, those samples are blurry and noisy. This makes learning from those data difficult, and so they need to be ignored. Using the learnable instance parameters, the algorithm can learn which samples help the recognition part of the model and which samples should be ignored or paid less attention.

2) If, during training, the data samples of a class are correctly classified, the corresponding data parameter of this class is decreased, resulting in the acceleration of the learning process (the loss function is decreased). However, if they are misclassified, then the class parameter is increased, which results in the deceleration of the learning process (the loss function is increased).

In the above curriculum learning method, we use 3 optimizers: 1 for training the model, 1 for training the class parameters, and 1 for training the instance parameters. The optimizers for the class and instance parameters are used only during training, since we do not have the data parameters ϕ^* for the test set.

This method is simple and effective, and it boosts the accuracy of sign recognition, as demonstrated in Section 4. Using those parameters, the algorithm can automatically learn to ignore noisy samples. In addition, it accelerates the learning of easier classes, while it decelerates and focuses on the learning of harder classes.

4. Experiments

The adopted GCN-based framework is tested for isolated sign recognition on the combined WLASL and ASLLVD isolated sign dataset (with consistent gloss labeling). Our training and testing protocol for both the forward and backward directions is described in Section 4.1. Section 4.2 explains the fusion of the forward and backward streams and the evaluation of the use of *data parameters* for curriculum learning.

Set ID	Sign Types	Min. # samples per sign	Total # class labels	Total # examples
LEX-6	Lexical	6	1,480	22,853
LEX-12	Lexical	12	983	18,362
ALL-6	All	6	1,502	23,016
ALL-12	All	12	990	18,482

Table 1: Dataset Statistics. *All* includes lexical signs, loan signs, and compounds

4.1. Training and Testing Protocol

4.1.1. Dataset Preprocessing

As described in (Dafnis et al., 2022), we modified the WLASL (Li et al., 2020) gloss labeling to make it consistent with the conventions of the ASLLRP datasets (which includes the ASLLVD), thereby also enforcing consistency of gloss labeling for the WLASL videos. As explained in Section 1.1, we merge the WLASL and ASLLVD isolated sign datasets (resulting in a set of 23,017 videos for 1,502 signs), and we use either lexical signs, or lexical plus loan signs and compounds; and we further restrict these sets to signs with at least either 6 or 12 examples. Increasing the minimum number of samples per sign also decreases the total number of available videos. Table 1 presents the numbers of sign classes and total videos for each set.

We split this dataset following (Li et al., 2020) into training, validation, and testing sets using a ratio of 4:1:1 for each sign. To evaluate the recognition performance, we use the mean scores of the *Top-K* recognition accuracy with $K = 1, 5$ over all sign instances.

4.1.2. Keypoint Extraction & Data Preprocessing

We use the pretrained Alphapose model of Fang et al. (2017), which estimates 136 keypoints of the whole body from single RGB images, and construct our skeleton graph of 27 nodes. To construct the graph, we first normalize the keypoint coordinates to [-1,1], and then apply random sampling, mirroring, rotation, scaling, and shifting as data augmentation techniques. Since the videos differ in total number of frames, the length of all videos is aligned to 200 frames. If a video has more than 200 frames, the first 200 are extracted from the video. However, given the length of the signs in our datasets, no information was lost as a result of this operation. If a video has fewer than 200, we repeat the frame sequence until the video length is 200 frames.

4.1.3. Training Details

To speed up and improve the training, we use a GCN model with pretrained weights from the AUTSL dataset (Sincan and Keles, 2020). The GCN models are implemented in PyTorch. All experiments were conducted using PyTorch 1.7.0 and an NVIDIA Quadro RTX8000s. To train the GCN model, the Stochastic Gradient Descent (SGD) with Nesterov Momentum

Streams	LEX-6				LEX-12				ALL-6				ALL-12			
	Forward		Backward		Forward		Backward		Forward		Backward		Forward		Backward	
	Top-1	*Top-5*	*Top-1*	*Top-5*	*Top-1*	*Top-5*	*Top-1*	*Top-5*	*Top-1*	*Top-5*	*Top-1*	*Top-5*	*Top-1*	*Top-5*	*Top-1*	*Top-5*
Joint	74.05	91.60	73.67	91.38	79.00	94.38	78.24	94.01	72.96	91.42	74.19	91.14	79.18	94.09	78.24	93.78
Bones	71.35	91.12	71.02	90.86	75.87	93.59	75.69	93.56	72.63	91.47	72.09	91.09	76.31	93.51	76.49	93.30
Multi-stream	77.35	94.08	77.54	93.70	82.95	96.08	82.22	95.87	77.58	94.21	77.65	94.24	83.07	95.87	82.26	95.87
Forward Multi-stream w/CL	**77.73**	**93.70**			**83.20**	**95.69**			**77.63**	**94.34**			**82.59**	**96.23**		
	Top-1		*Top-5*		*Top-1*		*Top-5*		*Top-1*		*Top-5*		*Top-1*		*Top-5*	
Fusion (no CL)	**78.54**		**94.72**		**84.23**		**96.69**		**78.70**		**94.79**		**84.70**		**96.56**	

Table 2: Recognition accuracy for all subsets.

(0.9) is selected as the optimization algorithm. The Cross-Entropy loss function is used, and the weight decay is set to 10^{-4}. The batch size for both the training and testing processes is set to 64, while the total number of epochs used for training our models is 300. In addition, the learning rate is initially set to 0.1 and divided by 10 when 150 and 200 epochs are reached.

4.2. GCN Performance

Table 2 shows the *Top-1* and *Top-5* recognition performance of the forward and backward stream directions. Of the streams for which there is both the forward and backward direction, the keypoint stream provides the best accuracy. The score fusion approach for the forward and backward directions further improves overall recognition accuracy in all the test cases. Table 2 shows recognition accuracy for all signs with at least 6 and 12 samples. We observe that using more samples per sign with fewer total sign classes—resulting in a more balanced dataset—increases the recognition rate by 5%.

4.3. GCN Performance with CL

Table 2 also shows the contribution of using curriculum learning (CL) over just using fusion of the forward streams. The current results are inconclusive; we will explore varying the CL parameters in the future, in particular to adapt CL for imbalanced datasets. After optimizing the parameters, we will add CL to the backward as well as the forward stream prior to fusion, to assess the extent to which CL may improve overall results.

4.4. Overall Results

Figure 3 summarizes the recognition accuracy for the subsets of the combined dataset that included all sign types (lexical, loan signs, compounds) using signs for which we had a minimum number of samples per sign of either 6 or 12, showing our fusion results (without incorporation of improvements from curriculum learning) for top-1, top-2, top-3, top-4, and top-5.

Table 2 shows little difference in recognition accuracy for datasets restricted to lexical signs, in part because lexical signs still predominate in the larger datasets, but also because we have not yet incorporated into our approach methods tailored to the specificity of linguistic properties of lexical signs, as we have done in previous research (Thangali et al., 2011; Dilsizian et al., 2014).

5. Discussion and Conclusions

We presented a new GCN-based approach to isolated sign recognition. It is distinctive in these respects:

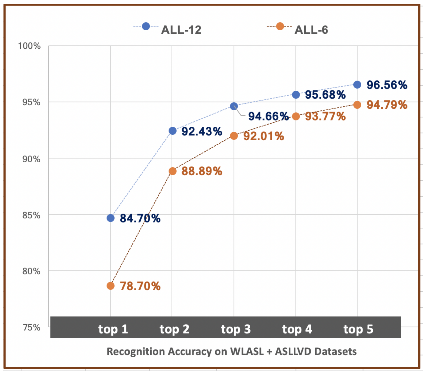

Figure 3: Summary of Sign Recognition Results: Based on Fusion (without curriculum learning)

1) Our method uses late fusion of forward and backward streams of joints and bones (following Dafnis et al. (2022)), not typically used in sign recognition.
2) This is the first time that ASL sign recognition research has been conducted by combining the ASLLVD and WLASL datasets, which gives rise to a large, rich, and diverse set of videos. This was made possible by our modifications to gloss labeling for WLASL videos, to enforce consistency of gloss labeling across these datasets, thereby also providing internally consistent gloss labels for the WLASL (not otherwise available).
3) This represents, to our knowledge, the first exploration of use of curriculum learning in sign recognition, by attending to the sign classes most difficult to learn, although our preliminary findings as to its promise for improving sign recognition accuracy are inconclusive.

To further improve recognition accuracy, in future research: 1) We will develop new curriculum learning methods to improve the estimation of difficult-to-recognize input signs, and integrate them with transformers. 2) We will further expand our dataset to include other data collections shared by the American Sign Language Linguistic Research Project (also with consistent gloss labeling). 3) We will conduct new machine learning research on extraction of 3D models from 2D video, with explicit integration of handshape recognition and incorporation of statistical information about the dataset that reflects linguistic constraints on the internal structure of signs.

6. Acknowledgments

We thank Matt Huenerfauth, Augustine Opoku, Carey Ballard, Indya-loreal Oliver, Lutece Dekker, Qilong Zhangli, Gregory Dimitriadis & Douglas Motto for assistance. We also gratefully acknowledge Stan

Sclaroff, Ashwin Thangali, Vassilis Athitsos, Joan Nash, Ben Bahan, Rachel Benedict, Naomi Caselli, Elizabeth Cassidy, Lana Cook, Braden Painter, Tyler Richard, Tory Sampson & Dana Schlang for collaboration in development of the ASLLVD. This research was supported in part by NSF grants 2040638, 1763486, 1763523 & 1763569. Any opinions, findings, or conclusions expressed here are those of the authors and do not necessarily reflect the views of the NSF.

7. Bibliographical References

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Thangali, A., Wang, H., and Yuan, Q. (2010). Large Lexicon Project: American Sign Language Video Corpus and Sign Language Indexing/Retrieval Algorithms. In *Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*, volume 2.

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.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

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

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.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cascante-Bonilla, P., Tan, F., Qi, Y., and Ordonez, V. (2020). Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *arXiv preprint arXiv:2001.06001*.

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. *JMLR*, 13:2205–2231.

Dafnis, K. M., Chroni, E., Neidle, C., and Metaxas, D. N. (2022). Bidirectional Skeleton-Based Isolated Sign Recognition using Graph Convolution Networks. In *13th International Conference on Language Resources and Evaluation, LREC 2022*, Marseille, France, June 2022.

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.

De Coster, M., Van Herreweghe, M., and Dambre, J. (2020). Sign language recognition with transformer networks. In *12th International Conference on Language Resources and Evaluation*, pages 6018–6024.

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.

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.

Hacohen, G. and Weinshall, D. (2019). On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*, pages 2535–2544. PMLR.

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.

Kumar, M., Packer, B., and Koller, D. (2010). Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23.

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.

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*.

Memiş, A. and Albayrak, S. (2013). A kinect based sign language recognition system using spatio-temporal features. In *6th 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 *Pro-*

ceedings of the 11th 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 Ballard, C. (2022). Why Alternative Gloss Labels Will Increase the Value of the WLASL Dataset. ASLLRP Project Report No. 21. `http://www.bu.edu/asllrp/rpt21/asllrp21.pdf`, Boston, MA: Boston University, March 2022.

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*.

Neidle, C., Opoku, A., Ballard, C., Dafnis, K. M., Chroni, E., and Metaxas, D. (2022a). Resources for Computer-Based Sign Recognition from Video, and the Criticality of Consistency of Gloss Labeling across Multiple Large ASL Video Corpora. In *10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, Marseille, France, June 2022.

Neidle, C., Opoku, A., and Metaxas, D. (2022b). ASL Video Corpora & Sign Bank: Resources Available through the American Sign Language Linguistic Research Project (ASLLRP). https://arxiv.org/abs/2201.07899 .

Pi, T., Li, X., Zhang, Z., Meng, D., Wu, F., Xiao, J., and Zhuang, Y. (2016). Self-paced boost learning for classification. In *IJCAI*, pages 1932–1938.

Pigou, L., Van Herreweghe, M., and Dambre, J. (2016). Sign classification in sign language corpora with deep neural networks. In *International Conference on Language Resources and Evaluation (LREC), Workshop, Proceedings*, pages 175–178.

Saxena, S., Tuzel, O., and DeCoste, D. (2019). Data parameters: A new family of parameters for learning a differentiable curriculum. *Advances in Neural Information Processing Systems*, 32.

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.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

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.

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.

Tamura, S. and Kawasaki, S. (1988). Recognition of sign language motion images. *Pattern recognition*, 21(4):343–353.

Thangali, A., Nash, J. P., Sclaroff, S., and Neidle, C. (2011). Exploiting phonological constraints for handshape inference in ASL video. In *CVPR 2011*, pages 521–528. IEEE.

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.

Wang, X., Chen, Y., and Zhu, W. (2021). A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Weinshall, D., Cohen, G., and Amir, D. (2018). Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*, pages 5238–5246. PMLR.

World Federation of the Deaf. (2022). `https://wfdeaf.org/our-work/`. Accessed: 2022-04-12.

Wu, X., Dyer, E., and Neyshabur, B. (2020). When do curricula work? *arXiv preprint arXiv:2012.03107*.

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.

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.

Zhao, Q., Meng, D., Jiang, L., Xie, Q., Xu, Z., and Hauptmann, A. G. (2015). Self-paced learning for matrix factorization. In *Twenty-ninth AAAI conference on artificial intelligence*.

Zhou, T., Wang, S., and Bilmes, J. (2020). Curriculum learning by dynamic instance hardness. *Advances in Neural Information Processing Systems*, 33:8602–8613.