# Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison

Dongxu Li , Cristian Rodriguez Opazo, Xin Yu, Hongdong Li
The Australian National University, Australian Centre for Robotic Vision (ACRV)
{dongxu.li, cristian.rodriguez, xin.yu, hongdong.li}@anu.edu.au

## Abstract

*Vision-based sign language recognition aims at helping the hearing-impaired people to communicate with others. However, most existing sign language datasets are limited to a small number of words. Due to the limited vocabulary size, models learned from those datasets cannot be applied in practice. In this paper, we introduce a new large-scale Word-Level American Sign Language (WLASL) video dataset, containing more than 2000 words performed by over 100 signers. This dataset will be made publicly available to the research community. To our knowledge,it is by far the largest public ASL dataset to facilitate word-level sign recognition research.*

*Based on this new large-scale dataset, we are able to experiment several deep learning methods for word-level sign recognition and evaluate their performances in large scale scenarios. Specifically we implement and compare two different models,i.e., (i) holistic visual appearance based approach, and (ii) 2D human pose based approach. Both models are valuable baselines that will benefit the community for method benchmarking. Moreover, we also propose a novel pose-based temporal graph convolution networks (Pose-TGCN) that models spatial and temporal dependencies in human pose trajectories simultaneously, which has further boosted the performance of the pose-based method. Our results show that pose-based and appearance-based models achieve comparable performances up to* $62.63\%$ *at top-10 accuracy on 2,000 words/glosses, demonstrating the validity and challenges of our dataset. We will make the large-scale dataset, as well as our baseline deep models, freely available on github.*

## 1. Introduction

Sign languages, as a primary communication tool for the deaf community, have their unique linguistic structures. Sign language interpretation methods aim at automatically translating sign languages using, for example, vision techniques. Such a process involves mainly two tasks, namely, word-level sign language recognition (or "isolated sign language recognition") and sentence-level sign language recognition (or "continuous sign language recognition"). In this paper, we target at word-level recognition task for American Sign Language (ASL) considering that it is widely adopted by deaf communities over 20 countries around the world [46].

Figure 1: ASL signs "read" (top row) and "dance" (bottom row) [14] differ only in the orientations of the hands.

Serving as a fundamental building block for understanding sign language sentences, the word-level sign recognition task itself is also very challenging:

- The meaning of signs mainly depends on the combination of body motions, manual movements and head poses, and subtle differences may translate into different meanings. As shown in Fig. 1, the signs for "dance" and "read" only differ in the orientations of hands.
- The vocabulary of signs in daily use is large and usually in the magnitude of thousands. In contrast, related tasks such as gesture recognition [5, 1] and action recognition [31, 59, 12] only contains at most a few hundred categories. This greatly challenges the scalability of recognition methods.
- A word in sign language may have multiple counterparts in natural languages. For instance, the sign

shown in Fig. 2 (a), can be interpreted as "wish" or "hungry" depending on the context. In addition, nouns and verbs that are from the same lemma usually correspond to the same sign. These subtleties are not well captured in the existing small-scale datasets.

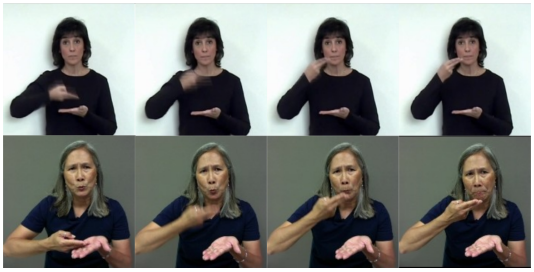
In order to learn a practical ASL recognition model, the training data needs to contain a sufficient number of classes and training examples. Considering that existing word-level datasets do not provide a large-scale vocabulary of signs, we firstly collect large-scale word-level signs in ASL as well as their corresponding annotations. Furthermore, since we want to leverage the minimal hardware requirement for the sign recognition, only monocular RGB-based videos are collected from the Internet. By doing so, the trained sign recognition models do not rely on special equipment, such as depth cameras [33] and colored gloves [52], and can be deployed in general cases. Moreover, when people communicate with each other, they usually sign in frontal views. Thus, we only collect videos with signers in near-frontal views to achieve a high-quality large-scale dataset. In addition, our dataset contains annotations for dialects that are commonly-used in ASL. In total, our proposed WLASL dataset consists 21,083 videos performed by 119 signers, and each video only contains one sign in ASL. Each sign is performed by at least 3 different signers. Thus, inter-signer variations in our dataset facilitates the generalization ability of the trained sign recognition models.

Based on WLASL, we are able to experiment with several deep learning methods for word-level sign recognition, based on (i) holistic visual appearance, and (ii) 2D human-pose. For *appearance-based methods*, we provide a baseline by re-training VGG backbone [58] and GRU [17] as a representative for convolutional recurrent networks. We also provide a 3D convolution networks baseline using fine-tuned I3D [12], which performs better than the VGG-GRU baseline. For *pose-based methods*, we firstly extract human poses from original videos and use them as input features. We provide a baseline using GRU to model the temporal movements of the poses. Giving that GRU captures explicitly only the temporal information in pose trajectories, it may not fully utilizes the spatial relationship between body keypoints. Motivated by this, we propose a novel pose-based model *temporal graph convolutional network (TGCN)* that captures the temporal and spatial dependencies in the pose trajectories simultaneously. Our results show that both pose-based approach and appearance-based approach achieve comparable classification performance on 2,000 words, reaching up to 62.63%.

## 2. Related Work

In this section, we briefly review some existing publicly sign language datasets, and state-of-the-art sign language recognition algorithms are also discuss to demonstrate the necessity of a large-scale ASL dataset.

(a) The verb "**Wish**" (top) and the adjective "**hungry**" (bottom) correspond to the same sign.



(b) The same sign represents different words "**Rice**" (top) and "**soup**" (bottom).

(c) Signers perform "**Scream**" with different hand positions and amplitude of hand movements.

Figure 2: Ambiguity and variations of Signing. (a, b) shows linguistic ambiguity in ASL. (c) shows signing variations of different signers.

### 2.1. Sign Language Datasets

There are three publicly released *word-level ASL* datasets[1], i.e. Purdue RVL-SLLL ASL Database [70], Boston ASLLVD [6] and RWTH-BOSTON-50 [79].

**Purdue RVL-SLLL ASL Database** [70] contains 39 motion primitives with different hand-shapes that are commonly encountered in ASL. Each primitive is produced by 14 native signers. Note that, the primitives in [70] are the elements constituting ASL signs but may not necessarily correspond to an English word. **Boston ASLLVD** [6] has

[1]We notice that an unpublished paper [32] aims at providing an ASL dataset containing 1,000 glosses. Since the dataset is not released, we cannot evaluate the quality and the usefulness of the dataset.

Figure 3: Illustrations of the diversity of our dataset, which contains different backgrounds, illumination conditions and signers with different appearances.

Table 1: Overview of word-level datasets in other languages.

| Datasets | #Gloss | #Videos | #Signers | Type | Sign Language |
|---|---|---|---|---|---|
| LSA64 [52] | 64 | 3,200 | 10 | RGB | Argentinian |
| PSL Kinect 30 [34] | 30 | 300 | - | RGB, depth | Polish |
| PSL ToF [34] | 84 | 1,680 | - | RGB, depth | Polish |
| DEVISIGN [15] | 2,000 | 24,000 | 8 | RGB, depth | Chinese |
| GSL [24] | 20 | 840 | 6 | RGB | Greek |
| DGS Kinect [3] | 40 | 3,000 | 15 | RGB, depth | German |
| LSE-sign [27] | 2,400 | 2,400 | 2 | RGB | Spanish |

2,742 words (*i.e.*, glosses) with 9,794 examples (3.6 examples per gloss on average). Although the dataset has large coverage of the vocabulary, more than 2,000 glosses have at most three examples, which is unsuitable to train thousand-way classifiers. **RWTH-BOSTON-50** [79] contains 483 samples of 50 different glosses performed by 2 signers. Moreover, **RWTH-BOSTON-104** provides 200 continuous sentences signed by 3 signers which in total cover 104 signs/words. **RWTH-BOSTON-400**, as a sentence-level corpus, consists of 843 sentences including around 400 signs, and those sentences are performed by 5 signers. **DEVISIGN** is a large-scale word-level Chinese Sign Language dataset, consists of 2,000 words and 24,000 examples performed by 8 non-native signers in controlled lab environment. Word-level sign language datasets exist for other regions, as summarized word-level sign language datasets in other languages in Table 1.

All the previously mentioned datasets have their own properties and provide different attempts to tackle the word-level sign recognition task. However, they fail to capture the difficulties of the task due to insufficient amount of instance and signer.,

## 2.2. Sign Language Recognition Approaches

Existing word-level sign recognition models are mainly trained and evaluated on either private [26, 38, 78, 28, 49] or small-scale datasets with less than one hundred words [26, 38, 78, 28, 49, 42, 47, 71]. These sign recognition approaches mainly consists of three steps: the feature extraction, temporal-dependency modeling and classification. Previous works first employ different hand-crafted features to represent static hand poses, such as SIFT-based features [72, 75, 64], HOG-based features [43, 8, 20] and features in the frequency domain [4, 7]. Hidden Markov Models (HMM) [61, 60] are then employed to model the temporal relationships in video sequences. Dynamic Time Warping (DTW) [41] is also exploited to handle differences of sequence lengths and frame rates. Classification algorithms, such as Support Vector Machine (SVM) [48], are used to label the signs with the corresponding words.

Similar to action recognition, some recent works [56, 35] use CNNs to extract the holistic features from image frames and then use the extracted features for classification. Several approaches [37, 36] first extract body keypoints and then concatenate their locations as a feature vector. The extracted features are then fed into a stacked GRU for recognizing signs. These methods demonstrate the effectiveness of using human poses in the word-level sign recognition task. Instead of encoding the spatial and temporal information separately, recent works also employ 3D CNNs [28, 76] to capture spatial-temporal features together. However, these methods are only tested on small-scale datasets. Thus, the generalization ability of those methods remains unknown. Moreover, due to the lack of a standard word-level large-scale sign language dataset, the results of different methods evaluated on different small-scale datasets are not comparable and might not reflect the practical usefulness of models.

To overcome the above issues in sign recognition, we propose a large-scale word-level ASL dataset, coined WLASL database. Since our dataset consists of RGB-only videos, the algorithms trained on our dataset can be easily applied to real world cases with minimal equipment requirements. Moreover, we provide a set of baselines using state-of-the-art methods for sign recognition to facilitate the evaluation of future works.

Table 2: Comparisons of our WLASL dataset with existing ASL datasets. Column "Mean" indicates the average number of video samples per gloss.

| Datasets | #Gloss | #Videos | Mean | #Signers | Year |
|---|---|---|---|---|---|
| Purdue RVL-SLLL [70] | 39 | 546 | 14 | 14 | 2006 |
| RWTH-BOSTON-50 [79] | 50 | 483 | 9.7 | 3 | 2005 |
| Boston ASLLVD [6] | 2,742 | 9,794 | 3.6 | 6 | 2008 |
| WLASL100 | 100 | 2,038 | 20.4 | 97 | 2019 |
| WLASL300 | 300 | 5,117 | 17.1 | 109 | 2019 |
| WLASL1000 | 1,000 | 13,168 | 13.2 | 116 | 2019 |
| WLASL2000 | 2,000 | 21,083 | 10.5 | 119 | 2019 |

# 3. Our Proposed WLASL Dataset

In this section, we introduce our proposed Word-Level American Sign Language dataset (WLASL). We first explain the data sources used and the data collection process. Following with the description of our annotation process which combines automatic detection procedures with manual annotations to ensure the correctness between signs and their annotations. Finally, we provide statistics of our WLASL.

## 3.1. Dataset Collection

In order to construct a large-scale signer-independent ASL dataset, we resort to two main resources from Internet. First, there are multiple educational sign language websites, such as ASLU [2] and ASL-LEX [14], and they also provide the lookup function for ASL signs. The mappings between the words/glosses and signs from those websites are accurate since those videos have been checked by experts before uploaded. Another main source is ASL tutorial videos on YouTube. We select videos with titles clearly describing the gloss of the sign. In total, we access 68,129 videos of 20,863 ASL glosses from 20 different websites. In each video, a signer performs only one sign (possibly multiple repetitions of the sign) in a nearly-frontal view but with different backgrounds.

After collecting all the resources for the dataset, if the gloss annotations are composed of more than two words in English, we will remove those videos to ensure that the dataset is word-level annotations. If the number of the videos for one gloss is less than seven, we also remove that gloss to guarantee that enough samples are split into the training and testing sets. Since most of the websites include daily used words, the small number of video samples for one gloss may imply those words are not frequently used. Therefore, removing those glosses with few video samples will not affect the usefulness of our dataset in practice. After this preliminary selection procedure, we have 34,404 video samples of 3,126 glosses for further annotations.

## 3.2. Annotations

In addition to providing a gloss label for each video, some meta information, including temporal boundary, body bounding box, signer annotation and sign dialect/variation annotations, is also given in our dataset.

*Temporal boundary:* A temporal boundary is used to indicate the start and end frames of a sign. When the videos do not contain repetitions of signs, the boundaries are labelled as the first and last frames of the signs. Otherwise, we manually label the boundaries between the repetitions. For the videos containing repetitions, we only keep one sample of the repeated sign to ensure samples in which the same signer performs the same sign will not appear in both training and testing sets. Thus, we prevent learned models from overfiting to the testing set.

*Body Bounding-box:* In order to reduce side-effects caused by backgrounds and let models focus on the signers, we use YOLOv3 [51] as a person detection tool to identify body bounding-boxes of signers in videos. Note that, the size of the bounding-box will change as a person signs, we use the largest bounding-box size to crop the person from the video.

*Signer Diversity:* A good sign recognition model should be robust to inter-signer variations in the input data, *e.g.* signer appearance and signing paces, in order to generalize well to real-world scenarios. For example, as shown in Fig. 2c, the same sign is performed with slightly different hand positioning by two signers. From this perspective, sign datasets should have a diversity of signers. Therefore, we identify signers in our collected dataset and then provide the IDs of the signers as the meta information of the videos. To this end, we first employ the face detector and the face embedding provided by FaceNet [54] to encode faces of the dataset, and then compare the Euclidean distances among the face embeddings. If the distance between two embeddings is lower than our pre-defined threshold (*i.e.*, 0.9), we consider those two videos signed by the same person. After automatic labeling, we also manually check the identification results and correct the mislabelled ones.

*Dialect Variation Annotation:* Similar to natural lan-

guages, ASL signs also have dialect variations [46] and those variations may contain different sign primitives, such as hand-shapes and motions. To avoid the situation where dialect variations only appear in testing dataset, we manually label the variations for each gloss. Our annotators receive training in advance to ensure that they understand the basic knowledge of ASL, in order to distinguish the differences from the signers variations and dialect variations. To speed up the annotation process and control the annotation quality, we design an interface which lets the annotators only compare signs from two videos displayed simultaneously. Then we count the number of dialects and assign labels for different dialects automatically. After the dialect annotation, we also give each video a dialect label. With the help of the dialect labels, we can guarantee the dialect signs in the testing set have corresponding training samples. We also discard the sign variations with less than five examples since there are not enough samples to be split into training, validation and testing sets. Furthermore, we notice that these variations are usually not commonly used in daily life.

### 3.3. Dataset Arrangement

After obtaining all the annotations for each video, we obtain videos with lengths ranging from 0.36 to 8.12 seconds, and the average length of all the videos is 2.41 seconds. The average intra-class standard deviation of the videos is 0.85 seconds.

We sort the glosses in a descending order in terms of the sample number of a gloss. To provide better understanding on the difficulties of the word-level sign recognition task and the scalability of sign recognition methods, we conduct experiments on the datasets with different vocabulary sizes. In particular, we select top-$K$ glosses with $K = \{100, 300, 1000, 2000\}$, and organize them to four subsets, named WLASL100, WLASL300, WLASL1000 and WLASL2000, respectively.

In Table 2, we present statistics of the four subsets of WLASL. As indicated by Table 2, we acquire 21,083 video samples with a duration of around 14 hours for WLASL2000 in total, and each gloss in WLASL2000 has 10.5 samples on average, which is almost three times larger than the existing large-scale dataset Boston ASLLVD. We show example frames of our dataset in Fig. 3.

## 4. Method Comparison on WLASL

Signing, as a part of human actions, shares similarities with human action recognition and pose estimation. In this section, we first introduce some relevant works on action recognition and human pose estimation. Inspired by network architectures of action recognition, we employ image-appearance based and pose based baseline models for word-level sign recognition. By doing so, we not only investigate the usability of our collected dataset but also exam the sign recognition performance of deep models based on different modalities.

### 4.1. Image-appearance based Baselines

Early approaches employ handcrafted features to represent the spatial-temporal information from image frames and then ensemble them as a high-dimensional code for classification [40, 69, 55, 39, 21, 66, 68].

Benefiting from the powerful feature extraction ability of deep neural networks, the works [57, 66] exploit deep neural networks to generate a holistic representation for each input frame and then use the representations for recognition. To better establish the temporal relationship among the extracted visual features, Donahue *et al*. [22] and Yue *et al*. [77] employ use recurrent neural networks (*e.g*., LSTM). Some works [23, 10] also employ the joint locations as a guidance to extract local deep features around the joint regions.

Sign language recognition, especially word-level recognition, needs to focus on detailed differences between signs, such as the orientation of hands and movement direction of the arms, while the background context does not provide any clue for recognition. Motivated by the action recognition methods, we employ two image-based baselines to model the temporal and spatial information of videos in different manners.

#### 4.1.1 2D Convolution with Recurrent Neural Networks

2D Convolutional Neural Networks (CNN) are widely used to extract spatial features of input images while Recurrent Neural Networks (RNN) are employed to capture the long-term temporal dependencies among inputs. Thus, our first baseline is constructed by a CNN and a RNN to capture spatio-temporal features from input video frames. In particular, we use VGG16 [58] pretrained on ImageNet to extract spatial features and then feed the extracted features to a stacked GRU [17]. This baseline is referred to as *2D Conv RNN*, and the network architecture is illustrated in Figure 4.

To avoid overfiting the training set, the hidden sizes of GRU for the four subsets are set to $64$, $96$, $128$ and $256$ respectively, and the number of the stacked recurrent layers in GRU is set to 2. In the training phase, we randomly select at most $50$ consecutive frames from each video. Cross-entropy losses is imposed on the output at all the time steps as well as the output feature from the average pooling of all the output features. In testing, we consider all the frames in the video and make predictions based on the average pooling of all the output features.

#### 4.1.2 3D Convolutional Networks

3D convolutional networks [13, 66, 63, 30] are able to establish not only the holistic representation of each frame

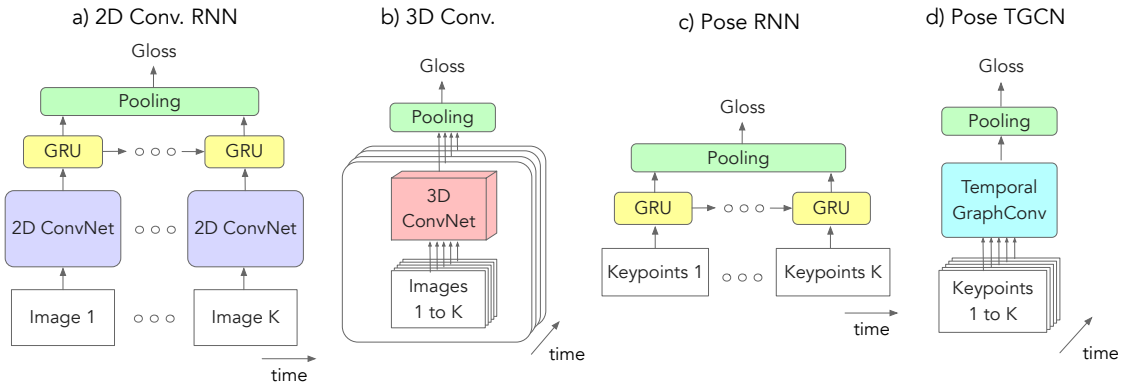



Figure 4: Illustrations of our baseline architectures.

but also the temporal relationship between frames in a hierarchical fashion. Carreira *et al*. [13] inflate 2D filters of the Inception network [62] trained on ImageNet [53], thus obtaining well-initialized 3D filters. The inflated 3D filters are also fine-tuned on the Kinetics dataset [13] to better capture the spatial-temporal information in a video.

In this paper, we employ the network architecture of I3D [13] as our second image-appearance based baseline, and the network architecture is illustrated in Figure 4. As mentioned above, the original I3D network is trained on ImageNet [53] and fine-tuned on Kinetics-400 [13]. In order to model the temporal and spatial information of the sign language, such as focusing on the hand shapes and orientations as well as arm movements, we need to fine-tune the pre-trained I3D. In this way, the fine-tuned I3D can better capture the spatio-temporal information of signs. Since the class number varies in our WLASL subsets, only the last classification layer is modified in accordance with the class number.

## 4.2. Pose-based Baselines

Human pose estimation aims at localizing the keypoints or joints of human bodies from a single image or videos. Traditional approaches employ the probabilistic graphical model [74] or pictorial structures [50] to estimate single-person poses. Recently, deep learning techniques have boosted the performance of pose estimation significantly. There are two mainstream approaches: regressing the keypoint positions [65, 11], and estimating keypoint heatmaps followed by a non-maximal suppression technique [9, 19, 18, 73]. However, pose estimation only provides the locations of the body keypoints, while the spatial dependencies among the estimated keypoints are not explored.

Several works [29, 67] exploit human poses to recognize actions. The works [29, 67] represent the locations of body joints as a feature representation for recognition. These methods can obtain high recognition accuracy when the oracle annotations of the joint locations are provided. In order to exploit the pose information for SLR, the spatial and temporal relationships among all the keypoints require further investigation.

### 4.2.1 Pose based Recurrent Neural Networks

Pose based approaches mainly utilize RNNs [45] to model the pose sequences for analyzing human motions. Inspired by this idea, our first pose-based baseline employs RNN to model the temporal sequential information of the pose movements, and the representation output by RNN is used for the sign recognition.

In this work, we extract 55 body and hand 2D keypoints from a frame on WLASL using OpenPose [9]. These keypoints include 13 upper-body joints and 21 joints for both left and right hands as defined in [9]. Then, we concatenate all the 2D coordinates of each joint as the input feature and feed it to a stacked GRU of 2 layers. In the design of GRUs, we use the empirically optimized hidden sizes of $64$, $64$, $128$ and $128$ for the four subsets respectively. Similar to the training and testing protocols in Section 4.1.1, $50$ consecutive frames are randomly chosen from the input video. Cross-entropy losses is employed for training. In testing, all the frames in a video are used for classification.

### 4.2.2 Pose Based Temporal Graph Neural Networks

We introduce a novel pose-based approach to ISLR using *Temporal Graph Convolution Networks (TGCN)*. Consider the input pose sequence $\mathbf{X}_{1:N} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_N]$ in $N$ sequential frames, where $\mathbf{x_i} \in \mathbb{R}^K$ represents the concatenated 2D keypoint coordinates in dimension K. We propose a new graph network based architecture that models the spatial and temporal dependencies of the pose sequence. Different from existing works on human pose estimation, which usually model motions using 2D joint angles, we encode temporal motion information as a holistic representation of the trajectories of body keypoints.

Motivated by the recent work on human pose forecasting [16, 44], we view a human body as a fully-connected graph with $K$ vertices and represent the edges in the graph as a

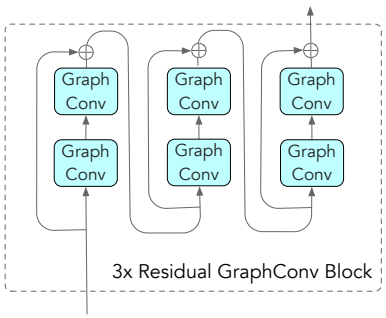



Figure 5: Residual Graph Convolution Block.

weighted adjacency matrix $\mathbf{A} \in \mathbb{R}^{K\times K}$. Although a human body is only partially connected, we construct the human body as fully-connected graph in order to learn the dependencies among joints via a graph network. In a deep graph convolutional network, the n-th graph layer is a function $\mathcal{G}_n$ that takes as input features a matrix $\mathbf{H}_n \in \mathbb{R}^{K\times F}$, where $F$ is the feature dimension output by its previous layer. In the first layer, the networks takes as input the $K \times 2N$ matrix coordinates of body keypoints. Given this formulation and a set of trainable weights $\mathbf{W}_n \in \mathbb{R}^{F\times F'}$, a graph convolutional layer is expressed as:

$$\mathbf{H}_{n+1} = \mathcal{G}_n(\mathbf{H}_n) = \sigma(\mathbf{A}_n\mathbf{H}_n\mathbf{W}_n), \quad (1)$$

where $\mathbf{A}_n$ is a trainable adjacency matrix for n-th layer and $\sigma(\cdot)$ denotes the activation function $tanh(\cdot)$. A residual graph convolutional block stacks two graph convolutional layers with a residual connection as shown in Fig. 5. Our proposed TGCN stacks multiple residual graph convolutional blocks and takes the average pooling result along the temporal dimension as the feature representation of pose trajectories. Then a softmax layer followed by the average pooling layer is employed for classification.

## 4.3. Training and Testing Protocol

### 4.3.1 Data Pre-processing and Augmentation

We resize the resolution of all original video frames such that the diagonal size of the person bounding-box is 256 pixels. For training VGG-GRU and I3D, we randomly crop a $224 \times 224$ patch from an input frame and apply a horizontal flipping with a probability of 0.5. Note that, the same crop and flipping operations are applied to the entire video frames instead of in a frame-wise manner. Similar to [12], when training VGG-GRU, Pose-GRU and Pose-TGCN, for each video consecutive 50 frames are randomly selected and the models are asked to predict labels based on only partial observations of the input video. In doing so, we increase the discriminativeness of the learned model. For I3D, we follow its original training configuration.

### 4.3.2 Implementation details

The models, *i.e*., VGG-GRU, Pose-GRU, Pose-TGCN and I3D are implemented in PyTorch. It is important to notice that we use the I3D pre-train weights provided by Carreira *et al*. [13]. We train all the models with Adam optimizer [34]. Note that, I3D was trained using stochastic gradient descent (SGD) in [13]. However, I3D does not converge when using SGD to fine-tune it in our experiments. Thus, Adam is employed to fine-tune I3D. All the models are trained with 200 epochs on each subset. We terminate the training process when the validation accuracy stop increasing.

We split the samples of a gloss into the training, validation and testing sets following a ratio of 4:1:1. We also ensure that each split has at least one sample per sign dialect. The split information will be released publicly as part of WLASL.

### 4.3.3 Evaluation Metric

We evaluate the models using the mean scores of top-$K$ classification accuracy with $K = \{1, 5, 10\}$ over all the sign instances. As seen in Figure 2, different meanings have very similar sign gestures, and those gestures may cause errors in the classification results. However, some of the erroneous classification can be rectified by contextual information. Therefore, it is more reasonable to use top-K predicted labels for the word-level sign language recognition.

## 4.4. Discussion

### 4.4.1 Performance Evaluation of Baseline Networks

Table 3 indicates that the performance of our baseline models based on poses and image-appearance. The results demonstrate that our pose-based TGCN further improves the classification accuracy in comparison to the pose-based sign recognition method Pose-GRU. This indicates that our proposed pose-TGCN captures both spatial and temporal relationships of the body keypoints since Pose-GRU mainly explores the temporal dependencies of the keypoints for classification. On the other hand, our fine-tuned I3D model achieves better performance compared to the other image-appearance based model VGG-GRU since I3D has larger network capacity and is pretrained on not only ImageNet but also Kinetics.

Although I3D is larger than our TGCN, Pose-TGCN can still achieve comparable results with I3D at top-5 and top-10 accuracy on the large-scale subset WLASL2000. This demonstrates that our TGCN effectively encodes human motion information. Since we use an off-the-shelf pose estimator [9], the erroneous estimation of poses may degrade the recognition performance. In contrast, image appearance-based baselines are trained in an end-to-end fashion for sign recognition and thus the errors residing in spatial features can be reduced during training. Therefore, training pose-based baselines in an end-to-end fashion could further improve the recognition performance.

Table 3: Top-1, top-5, top-10 accuracy (%) achieved by each model (by row) on the four WLASL subsets.

| Method | WLASL100 | | | WLASL300 | | | WLASL1000 | | | WLASL2000 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | top-1 | top-5 | top-10 | top-1 | top-5 | top-10 | top-1 | top-5 | top-10 | top-1 | top-5 | top-10 |
| Pose-GRU | 46.51 | 76.74 | 85.66 | 33.68 | 64.37 | 76.05 | 30.01 | 58.42 | 70.15 | 22.54 | 49.81 | 61.38 |
| Pose-TGCN | 55.43 | 78.68 | 87.60 | 38.32 | 67.51 | 79.64 | 34.86 | 61.73 | 71.91 | 23.65 | 51.75 | 62.24 |
| VGG-GRU | 25.97 | 55.04 | 63.95 | 19.31 | 46.56 | 61.08 | 14.66 | 37.31 | 49.36 | 8.44 | 23.58 | 32.58 |
| I3D | **65.89** | **84.11** | **89.92** | **56.14** | **79.94** | **86.98** | **47.33** | **76.44** | **84.33** | **32.48** | **57.31** | **66.31** |

Table 4: Top-10 accuracy (%) of I3D (and Pose-TGCN when trained (row) and tested (column) on different WLASL subsets.

| | WLASL100 | | WLASL300 | | WLASL1000 | | WLASL2000 | |
|---|---|---|---|---|---|---|---|---|
| | I3D | TGCN | I3D | TGCN | I3D | TGCN | I3D | TGCN |
| WLASL100 | 89.92 | 87.60 | - | - | - | - | - | - |
| WLASL300 | 88.37 | 81.40 | 86.98 | 79.64 | - | - | - | - |
| WLASL1000 | 85.27 | 77.52 | 86.22 | 74.25 | 84.33 | 71.91 | - | - |
| WLASL2000 | 72.09 | 67.83 | 71.11 | 65.42 | 67.32 | 64.55 | 66.31 | 62.24 |

#### 4.4.2 Effect of Vocabulary Size

As seen in Table 3, our baseline methods can achieve relatively high classification accuracy on small-size subsets. *i.e.*, WLASL100 and WLASL300. However, the subset WLASL2000 is very close to the real-world word-level classification scenario due to its large vocabulary. Pose-GRU, pose-TGCN and I3D achieve similar performance on WLASL2000. This implies that the recognition performance on small vocabulary datasets does not reflect the model performance on large vocabulary datasets, and the large-scale sign language recognition is very challenging.

We also evaluate how the class number, *i.e.*, vocabulary size, impacts on the model performance. There are two factors mainly affecting the performance: (i) deep models themselves favor simple and easy tasks, and thus they perform better on smaller datasets. As indicated in Table 3, the models trained on smaller vocabulary size sets perform better than larger ones (comparing along columns); (ii) the dataset itself has ambiguity. Some signs, as shown in Figure 2, are hard to recognize by even humans, and thus deep models will be also misled by those classes. As the number of classes increases, there will be more ambiguous signs.

In order to explain the impacts of the second factor, we dissect the models, *i.e.*, I3D and Pose-TGCN, trained on WLASL2000. Here, we test our models on the WLASL100, WLASL300, WLASL1000 and WLASL2000. As seen in Table 4, when the test class number is smaller, the models achieve higher accuracy (comparing along rows). The experiments imply that as the number of classes decreases, the number of ambiguous signs becomes smaller, thus making classification easier.

#### 4.4.3 Effect of Sample Numbers

As the class number in the dataset increases, training a deep model requires more samples. However, as illustrated in Table 1, although in our dataset each gloss contains more samples than other datasets, the number of training examples per class is still relatively small compared to some large-scale generic activity recognition datasets [25]. This brings some difficulties for the network training. Note that, the average training samples for each gloss in WLASL100 are twice large as those in WLASL2000. Therefore, models obtain better classification performance on the glosses with more samples, as indicated in Table 3 and Table 4.

Crowdsourcing via Amazon Mechanism Tucker (AMT) is a popular way to collect data. However, annotating ASL requires specific domian knowledge and makes crowdsourcing infeasible.

## 5. Conclusion

In this paper, we proposed a large-scale Word-Level ASL (WLASL) dataset covering a wide range of daily words and evaluated the performance of deep learning based methods on it. To the best of our knowledge, our dataset is the largest public ASL dataset in terms of the vocabulary size and the number of samples for each class. Since understanding sign language requires very specific domain knowledge, labelling a large amount of samples per class is unaffordable. After comparisons among deep sign recognition models on WLASL, we conclude that developing word-level sign language recognition algorithms on such a large-scale dataset requires more advanced learning algorithms, such as few-shot learning. In our future work, we also aim at utilizing word-level annotations to facilitate sentence-level and story-level machine sign translations.

## References


[1] The 20bn-jester dataset-v1. https://20bn.com/datasets/jester. Accessed: 2019-07-16.

[2] Asl university. http://asluniversity.com/. Accessed: 2019-07-16.

[3] Kinect gesture dataset. https://www.microsoft.com/en-us/download/details.aspx?id=52283. Accessed: 2019-07-16.

[4] M. Al-Rousan, K. Assaleh, and A. Talaa. Video-based signer-independent arabic sign language recognition using hidden markov models. *Applied Soft Computing*, 9(3):990–999, 2009.

[5] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7243–7252, 2017.

[6] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali. The american sign language lexicon video dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

[7] P. C. Badhe and V. Kulkarni. 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, 2015.

[8] P. Buehler, A. Zisserman, and M. Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2968. IEEE, 2009.

[9] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*, 2017.

[11] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

[12] J. Carreira and A. Zisserman. 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, 2017.

[13] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[14] N. K. Caselli, Z. S. Sehyr, A. M. Cohen-Goldberg, and K. Emmorey. Asl-lex: A lexical database of american sign language. *Behavior research methods*, 49(2):784–801, 2017.

[15] X. Chai, H. Wanga, M. Zhoub, G. Wub, H. Lic, and X. Chena. Devisign: Dataset and evaluation for 3d sign language recognition. Technical report, Beijing, Tech. Rep, 2015.

[16] H.-k. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019.

[17] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[18] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4715–4723, 2016.

[19] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2017.

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

[21] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. 2005.

[22] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[23] W. Du, Y. Wang, and Y. Qiao. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3725–3734, 2017.

[24] E. Efthimiou and S.-E. Fotinea. Gslc: Creation and annotation of a greek sign language corpus for hci. In *HCI*, 2007.

[25] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[26] K. Grobel and M. Assan. 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. IEEE, 1997.

[27] E. Gutierrez-Sigut, B. Costello, C. Baus, and M. Carreiras. Lse-sign: A lexical database for spanish sign language. *Behavior Research Methods*, 48(1):123–137, 2016.

[28] J. Huang, W. Zhou, H. Li, and W. Li. Sign language recognition using 3d convolutional neural networks. In *2015 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2015.

[29] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013.

[30] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[31] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014.

[32] H. R. V. Joze and O. Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.

[33] T. Kapuscinski, M. Oszust, M. Wysocki, and D. Warchol. Recognition of hand gestures observed by depth cameras. *International Journal of Advanced Robotic Systems*, 12(4):36, 2015.

[34] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] P. Kishore, G. A. Rao, E. K. Kumar, M. T. K. Kumar, and D. A. Kumar. Selfie sign language recognition with convolutional neural networks. *International Journal of Intelligent Systems and Applications*, 10(10):63, 2018.

[36] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683, 2019.

[37] S.-K. Ko, J. G. Son, and H. Jung. Sign language recognition with recurrent neural network using human keypoint detection. In *Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems*, pages 326–328. ACM, 2018.

[38] V. S. Kulkarni and S. Lokhande. Appearance based recognition of american sign language using gesture segmentation. *International Journal on Computer Science and Engineering*, 2(03):560–565, 2010.

[39] I. Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3):107–123, 2005.

[40] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. 2008.

[41] J. F. Lichtenauer, E. A. Hendriks, and M. J. Reinders. Sign language recognition by combining statistical dtw and independent classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):2040–2046, 2008.

[42] K. M. Lim, A. W. Tan, and S. C. Tan. Block-based histogram of optical flow for isolated sign language recognition. *Journal of Visual Communication and Image Representation*, 40:538–545, 2016.

[43] S. Liwicki and M. Everingham. Automatic recognition of fingerspelled words in british sign language. In *2009 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 50–57. IEEE, 2009.

[44] W. Mao, M. Liu, M. Salzmann, and H. Li. Learning trajectory dependencies for human motion prediction. *arXiv preprint arXiv:1908.05436*, 2019.

[45] J. Martinez, M. J. Black, and J. Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

[46] C. McCaskill, C. Lucas, R. Bayley, and J. Hill. *The hidden treasure of Black ASL: Its history and structure*. Gallaudet University Press Washington, DC, 2011.

[47] D. Metaxas, M. Dilsizian, and C. Neidle. Scalable asl sign recognition using model-based machine learning and linguistically annotated corpora. In *8th Workshop on the Representation & Processing of Sign Languages: Involving the Language Community, Language Resources and Evaluation Conference 2018*, 2018.

[48] S. Nagarajan and T. Subashini. Static hand gesture recognition for sign language alphabets using edge oriented histogram and multi class svm. *International Journal of Computer Applications*, 82(4), 2013.

[49] L. Pigou, M. Van Herreweghe, and J. Dambre. Gesture and sign language recognition with temporal residual networks. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[50] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013.

[51] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[52] F. Ronchetti, F. Quiroga, C. A. Estrebou, L. C. Lanzarini, and A. Rosete. Lsa64: an argentinian sign language dataset. In *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016).*, 2016.

[53] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[54] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[55] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360. ACM, 2007.

[56] H. Shin, W. J. Kim, and K.-a. Jang. Korean sign language recognition based on image and convolution neural network. In *Proceedings of the 2nd International Conference on Image and Graphics Processing*, pages 52–55. ACM, 2019.

[57] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[58] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[59] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[60] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1371–1375, 1998.

[61] T. E. Starner. Visual recognition of american sign language using hidden markov models. Technical report, Massachusetts Inst Of Tech Cambridge Dept Of Brain And Cognitive Sciences, 1995.

[62] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich.

Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[63] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *European conference on computer vision*, pages 140–153. Springer, 2010.

[64] A. Tharwat, T. Gaber, A. E. Hassanien, M. K. Shahin, and B. Refaat. Sift-based arabic sign language recognition system. In *Afro-european conference for industrial advancement*, pages 359–370. Springer, 2015.

[65] A. Toshev, C. Szegedy, and G. DeepPose. Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA*, pages 24–27, 2014.

[66] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[67] C. Wang, Y. Wang, and A. L. Yuille. An approach to pose-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2013.

[68] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. 2011.

[69] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. 2009.

[70] R. Wilbur and A. C. Kak. Purdue rvl-slll american sign language database. 2006.

[71] Q. Xue, X. Li, D. Wang, and W. Zhang. Deep forest-based monocular visual sign language recognition. *Applied Sciences*, 9(9):1945, 2019.

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

[73] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3073–3082, 2016.

[74] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011*, pages 1385–1392. IEEE, 2011.

[75] F. Yasir, P. C. Prasad, A. Alsadoon, and A. Elchouemi. Sift based approach on bangla sign language recognition. In *2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA)*, pages 35–39. IEEE, 2015.

[76] Y. Ye, Y. Tian, M. Huenerfauth, and J. Liu. Recognizing american sign language gestures from within continuous videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2064–2073, 2018.

[77] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[78] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. American sign language recognition with the kinect. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 279–286. ACM, 2011.

[79] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Joint Pattern Recognition Symposium*, pages 401–408. Springer, 2005.