Hand Pose Guided 3D Pooling for Word-level Sign Language Recognition

Al Amin Hosain, Panneer Selvam Santhalingam, Parth Pathak, Huzefa Rangwala and Jana Košecká
George Mason University, Fairfax, USA
{ahosain, psanthal, phpathak, rangwala, kosecka}@gmu.edu

Abstract

Gestures in American Sign Language (ASL) are characterized by fast, highly articulate motion of upper body, including arm movements with complex hand shapes and facial expressions. In this work, we propose a new method for word-level sign recognition from American Sign Language (ASL) using video. Our method uses both motion and hand shape cues while being robust to variations of execution. We exploit the knowledge of the body pose, estimated from an off-the-shelf pose estimator. Using the pose as a guide, we pool spatio-temporal feature maps from different layers of a 3D convolutional neural network. We train separate classifiers using pose guided pooled features from different resolutions and fuse their prediction scores during test time. This leads to a significant improvement in performance on the WLASL benchmark dataset [25]. The proposed approach achieves 10%, 12%, 9.5% *and* 6.5% *performance gain on WLASL100, WLASL300, WLASL1000, WLASL2000 subsets respectively. To demonstrate the robustness of the pose guided pooling and proposed fusion mechanism, we also evaluate our method by fine tuning the model on another dataset. This yields* 10% *performance improvement for the proposed method using only* 0.4% *training data during fine tuning stage.*

1. Introduction

Sign language is the primary form of communication among the Deaf and Hard-of-Hearing (DHH) persons. There are 70 million DHH people around the world and there exists more than 300 sign languages [17]. This large section of population suffers from communication barrier in many ways. Sign language is a complete independent language from its counterpart of spoken language. The most commonly considered sign language recognition tasks tackled by computer vision techniques are finger spelling recognition, sentence parsing and world-level sign gesture recognition. In this paper we focus on the word-level sign language recognition problem. The basic components of a sign gesture are complex arm movements with articulated hand shapes, and facial expression. Both the motion and the shape of the hands are the most discriminative components of individual gestures. From a computer vision perspective, the word-level gesture recognition requires learning strong *spatio-temporal* representations from videos, capturing both the appearance of the hand as determined by its shape and pose, as well as motion of arms and hands. Several deep learning based approaches were found effective in capturing *spatio-temporal* representations of the data for action recognition or action localization tasks on commonly used action recognition benchmarks [40, 5]. The common building blocks of these models use a combination of Deep Convolutional Neural Networks (ConvNet) for extracting the spatial feature and Recurrent Neural Network (RNN) for modelling the temporal aspect [8, 57], or use combination of 2D and 3D convolutional networks for fusing *spatio-temporal* cues [20, 57, 50, 46]. These methods take action videos as input and compute class prediction probabilities over the available classes or occasionally action localization depending on the labels available in the training set. Alternative approaches for action recognition exploit the existing techniques for human pose estimation [4]

Figure 1: Demonstration of pose in feature map space. Top images show the variation in hand position in three different samples from a sign gesture of `city` class. Bottom images shows how the hand poses are mapped (for the middle sample) to corresponding activation maps for four randomly selected channels from a certain layer of a 3D ConvNet.

in individual frames and learn models on the top of these representations [26, 6]. Large motions, motion blur and limited resolution make the body pose estimation brittle. The same factors affect the estimation of hand finger joints, making these methods ineffective for capturing hand shape and pose. For both action recognition and word-level sign language recognition 3D ConvNets are currently one of the best performing models [5, 25] enabling end-to-end training of *spatio-temporal* component. For word-level gesture recognition, previous methods that shown the effectiveness of hand shape for gesture recognition required training separate models for hand shape classification, in a supervised or weakly-supervised way [15, 23].

Motivated by the effectiveness of hand shape for gesture recognition, we propose to augment a 3D ConvNet model by pose guided pooling from 3D convolutional features maps at different layers and levels of resolution. We propose to use the spatial locations of hands in the feature maps to guide the pooling. Location of hands on RGB video frames can be reliably estimated by any state-of-the-art of pose estimation method [4]. Figure 1 shows an example of body poses mapped to corresponding locations in different feature layers of a 3D ConvNet. We learn to predict the word-level gestures by training additional classifiers using pose guided pooling of 3D ConvNet features maps with different spatial support. During test time, we fuse the class probability scores from classifiers learned using these pooled multi scale features. In summary, our contributions can be listed as follow:

- We propose a novel pose guided pooling mechanism for word-level ASL recognition;
- We investigate the idea of pooling localized features from multiple feature map levels of 3D convolutional network;
- We evaluate the proposed architectures improving the state of the art results on word-level action recognition;
- We validate the feature transferability of our proposed method by evaluating on a different dataset, using only 0.4% training samples.

2. Related Work

In the following section we will review some related works for action recognition in video, gesture recognition as well as more specific sign language recognition. The approaches differ in the representation of appearance and motion and the datasets used to benchmark the models.

2.1. Gesture and Action Recognition

Commonly used approaches for action recognition from video extract various features from RGB data or use body poses (skeletal) data or a combination of both. For certain actions, the appearance information in a single frame can unambiguously determine the actions, for others the motion is the discriminant cues. While the history of the modeling approaches is rich, we focus the review on more recent methods. To capture both the spatial and temporal signals in the video, Simonyan *et al*. [41] explore different ways to fuse spatial and temporal information from a two stream appearance and flow convolutional network. Similar multi-stream architectures were introduced in [33] for gesture recognition. The architecture is based on 2D convolution and sparse fusion of scores from different channels of input stream where some of the channels are focused on hands. Later approaches explored the idea of 3D convolution for joint learning of the spatial and temporal features [48, 5]. Inflated 3D ConvNet (I3D) [5] network extended pre-trained 2D convolutional kernels to temporal dimension to bootstrap learning 3D convolutional filters. Some approaches focused on temporal modeling by either learning sparse frame sampling [53], learning hierarchical features [29] or generating temporal candidate proposals [56]. Using unsupervised techniques are also well studied in gesture or activity recognition. These methods typically try to capture the temporal order similarities of full or sub-activities of similar kinds [10, 32], bypassing the need for more detailed labeling.

Activity recognition using body pose (or skeletal) data is also a well studied problem [39, 27, 9]. Shahroudy *et al*. released a large scale dataset for human activity recognition [39] and proposed an extension of long short term memory (LSTM) model which leverages group motion of several body joints to recognize human activity from skeletal data. A different adaptation of the LSTM model was proposed by Liu *et al*. where spatial interactions among joints was considered in addition to the temporal dynamics [27]. Veeriah *et al*. [51] proposed to capture derivative of motion states among different body joints, meanwhile Du *et al*. [9] exploited the hierarchical arrangement of different body parts. Several attention based models were proposed for human activity analysis [42, 28]. Some approaches use skeletal sequences of body joints to develop new representations which captures the spatio-temporal cues in the videos [22, 26]. Often, the goal of these method is to generate an image like representation of a video pose sequence, to facilitate the use of pre-trained image models.

The aforementioned methods used either RGB video input (possibly with optical flow) or pose data separately or together to model sign or activity video samples. However, using pose guided feature extraction from a ConvNet is also well studied in video action recognition [52, 3]. These approaches deviate from the traditional use of poses of modeling sign videos. Instead, these methods use pose to localize a position in a feature map to extract from. The feature

map is usually generated from an RGB input video. Our approach follows a similar mechanism. However, to the best of our knowledge, we are the first to try this in a sign language recognition setting.

2.2. Sign Language Recognition

From recognition perspective, sign languages can be categorized into world level signs (WLSLR), continuous sign sentences (CSLR) and finger spelling. Early approaches for WLSLR used hand crafted features from either RGB or skeletal sequences with Hidden Markov Model (HMM) for parsing simple multi-word phrases [60, 58, 43]. More recent efforts used deep learning techniques to bypass the feature engineering [18, 21, 44]. These models are either based on 3D ConvNets [18] or Support Vector Machine classifiers using extracted features from ConvNets [21, 44]. Lionel *et al*. [35] used ConvNet to model Italian sign gestures in a similar manner.

For sentence level parsing CSLR tasks, most of the state-of-the-art methods are built upon RWTH-PHOENIX-Weather corpus [11]. The corpus contains weather forecasts simultaneously interpreted into sign language which were recorded from German public TV and manually annotated using glosses on the sentence level. Koller *et al*. [23] trained a 22 layer deep convolutional neural network (CNN) with more than 1 million images from videos of Danish and New Zealand sign language. The data is weakly labeled with only video level annotation. The ConvNet model for estimating the likelihood of hand shapes, is trained using EM algorithm, jointly with Hidden Markov Model (HMM) for parsing sign gestures. Several later approaches focused on temporal modeling of sign sentences with the help of Connectionist Temporal Classification (CTC) loss [7, 37, 14, 13]. For example, Cui *et al*. [7] proposed a method of staged optimization where first alignment proposal is learned for sign sequences and those used as stronger supervision in final task. Pu *et al*. [37] proposed dilated convolutional kernels, also followed a pseudo label based training, for capturing temporal dynamics. Guo *et al*. [13] also followed as similar pardigm with 2d convolutional pyramid features. Recently, Pu *et al*. [38] proposed a combination of 3D ConvNet and RNN encoder-decoder, with alternative iterative training technique, to model continuous sign sentences.

ASL Recognition There are several works that directly relate to world-level recognition problem in American Sign Language (ASL) that we study here. Early works proposed for the WLSLR for ASL relied on linguistic properties [2, 31, 30, 47]. Although, these works focused on the inherent part of the language, computer vision techniques were not properly exploited. Later approaches feature small datasets or restricted laboratory environment settings [18, 24, 59, 60, 58]. More recently Hosain *et al*. [16] introduced a medium scale dataset with 11k word sign samples and proposed skeletal based RNN models for recognition. Although, not small in size, the dataset was collected in a laboratory environment. The size of the dataset is an important contributing factor when it comes to training large deep ConvNets. Most recently, two large scale word-level datasets were introduced to the community [49, 25]. Both datasets used online wild video resources for retrieving sign samples. These datasets feature adequate variation to validate evolving deep learning based models for sign language recognition. For example MS-ASL [49] contains more than 20k sign samples of almost 1k sign classes, while WLASL [25] features almost same number of sign samples with 2k sign classes in total. Both works proposed several human pose based and RGB based baselines demonstrating the challenges and usefulness of the models. In both cases, 3D ConvNet based model was the best performing model. In our work, we show how to enhance this model by body pose guided pooling of *spatio-temporal* features maps, where the location of the hand is estimated by the state-of-the-art human body pose estimator.

3. Our Approach

Given, a dataset of N training examples $\{V_i, W_i\}$, where V_i is an RGB video $\in \Re^{T\times H\times W\times 3}$ and W_i is a word level label; $H, W, 3$ is the dimension of a single frame of a video and T is its length. We seek to train a machine learning model to predict the word-level sign from videos. Figure 1 shows example gestures for the sign `city`, performed by three different signers. Specifically, our proposed method leverages motion and hand shape cues by being robust to variations while signing. We use the estimated body pose as a guide for pooling *spatio-temporal* feature maps at different layers of a 3D ConvNet. We train independent classifiers using pose guided pooled feature maps from different resolutions and fuse their prediction scores during test time.

In this section, we first describe the baseline 3D ConvNet model, followed by our pose guided pooling and fusion approach.

3.1. Inflated 3D ConvNet

Deep convolutional neural networks (ConvNet) for image classification proceed by learning layers of shared filter parameters that are used obtain spatial features maps by means of convolution. To model sequential nature of video data, convolution can be extended to 3D where weights of 3D filters can learned directly from *spatio-temporal* data [20]. However, training these models form scratch is quite challenging, due to large number parameters added by the temporal dimension of convolutional filters. To mitigate this issue, I3D network [5] was proposed to use already trained 2D convolutional filters and inflate them into 3D to

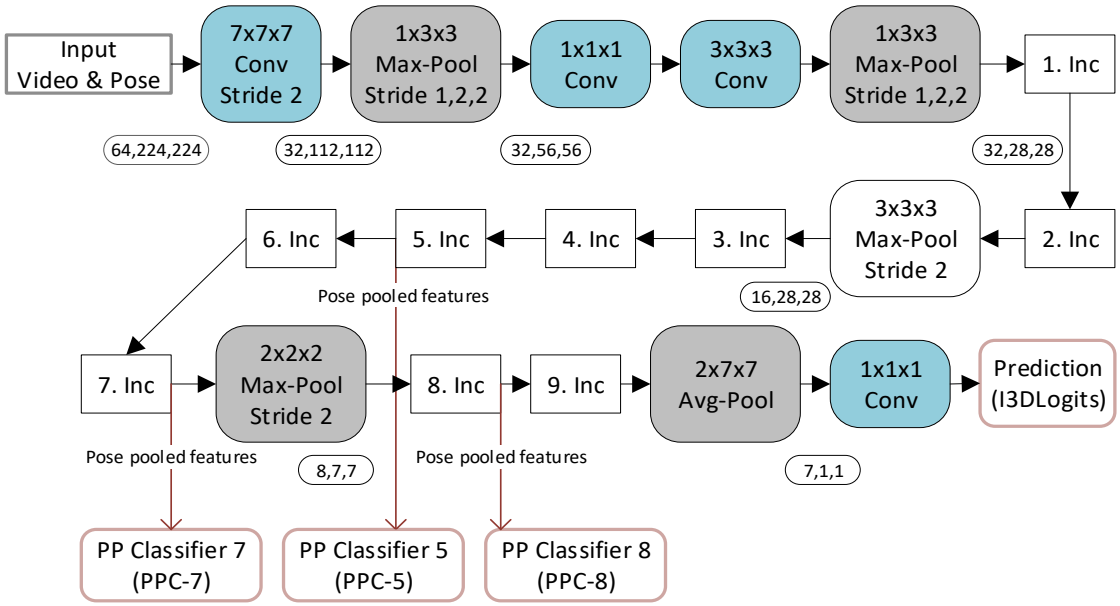


Figure 2: I3D Inception-v1 based sign video recognition pipeline. All inception blocks (Inc) are numbered for the convenience of description. Volume of output is labeled as "temporal,height,width" after any layer where it is being changed by the previous layer's sampling and convolution filters. Number of feature maps are not shown for simplicity in any output volume. Pose pooled classifier is shown for three output locations (PPC-5, PPC-7 & PPC-8), while it can be done from any output points in the network.

initialize the training of 3D convolutional network. Inflating pre-trained 2D filters into 3D allowed the network to learn seamless spatio-temporal features and has become the state-of-the-art method in action recognition and world-level sign gesture recognition [5, 25]. GoogleNet was proposed to find the optimal sparse local structure in data using readily available dense convolutional fiters [45]. The idea was motivated by the fact that, not every neuron, at each layer is equally responsible for the learning process of a ConvNet and optimal network structure can be approximated using the correlation statistics of highly activated neurons layer by layer [1]. The authors of I3D [5] inflated a version of GoogleNet and initialized 3D filters using 2D versions trained on image recognition task. This architecture is titled as Inflated Inception-V1 and we use at as one of our baselines as well as starting point in our proposed model. Figure 2 shows the inception module as `Inc`. Details can be found in the original paper [5].

3.2. Pose Estimation

Pose estimation is the process of estimating 2D or 3D body joint locations (e.g. wrist, elbow) in single image. Typically it is done by first detecting human subjects in an image frame and then parsing body joint location [34, 12] or directly inferring the body parts without first detecting the person [55, 54, 4, 19, 36]. For this work, to extract 2D body poses, we have chosen the state-of-the-art human body pose estimation approach OpenPose [4].

3.3. Proposed Method

In our approach we propose to augment sign video recognition using 3D ConvNet along with body pose. We assume that the body pose estimates are available for each frame using an off-the-shelve pose estimation approach [4]. The video is passed through the layers of 3D ConvNet generating spatio-temporal features maps with multiple channels at different levels of resolution. We then use the estimates of body joint locations to guide the pooling of the *spatio-temporal* feature maps to generate additional predictions. Details of our architecture is presented in this section.

We denote a sign gesture video input by $V^{T\times H\times W}$ where T is the video length in number of frames and H, W are the spatial dimensions. For a person performing a sign gesture in the video, the estimated body pose tensor of the person can be represented as $P^{T\times J\times 2}$ where T is the number of frames, J is the count of body locations and 2 for the (x, y) image coordinates for the pose of each body location. Even though the consecutive levels of 3D ConvNet

change the spatial and temporal dimensions of the feature maps due to stride and pooling operation, we can compute corresponding spatial coordinates of joint locations for respective feature maps. Suppose we have a feature map at level k with dimensions $F^{T' \times H' \times W'}$. The spatial joint coordinates can be scaled down based on the ratios of heights and widths between the input video and the 3D feature map. For the temporal dimension we uniformly sample T' frames from initial T, to match the temporal dimension of the feature maps. More specifically, we convert an input tensor index (t_v, x_v, y_v) to feature map index (t_f, x_f, y_f) using following equation:

$$\begin{gathered} s_x = \frac{H'}{H} \;\; s_y = \frac{W'}{W} \;\; s_t = \frac{T'}{T} \\ x_f = s_x \times x_v, \; y_f = s_y \times y_v \\ t_f = TemporalSample(T, T')[s_t \times t_v] \end{gathered} \tag{1}$$

where s_x and s_y are the scaling factors for height and width change of resolution. The function $TemporalSample$ divides the T temporal dimension into T' equal length windows, picks the middle index from each window and returns a list of temporal indices. To get the corresponding temporal index of pose data to the feature map, $(s_t \times t_v)^{th}$ number from the sampled indices is selected. Given the computed joint locations in the feature maps we use these to guide the pooling to generate the new feature vector at that layer.

3.3.1 Pose Guided Pooling Classifier (PPC)

We pool features around each joint location and stack all joints' feature. Pooling from a feature map of $[f \times t \times h \times w]$ using j joint locations leads to a feature representation of size $[f \times t \times j]$. Using these features we train a separate linear softmax classifier described in more detail in experiments section. The architecture learns several classifiers separately and during test time we fuse their prediction scores by summing. Since different classifiers use features from different scales of 3D ConvNet, they carry complementary information. This fact is reflected in overall performance improvement using our pose pooling and score fusion mechanism. Figure 2 shows the overall architecture, comprised of I3D backbone network with labelled inception modules. This figure shows, `PP Classifer 7 (PPC-7)` gets pose pooled features from the inception layer labeled 7, and `PPC-8 & PPC-5` from levels 8 and 5 respectively. We can extract features from any 3D feature map level and train a separate classifier. For the rest of the discussion, we refer to classifier from n^{th} inception module as `PPC-n`. We also consider the final prediction of baseline I3D network, termed as `I3DLogits` and experiment with different fusion strategies described in experiments section.

Datasets	#Gloss	#Videos	#Mean	#Signers
WLASL100	100	2,038	20.4	97
WLASL300	300	5,117	17.1	109
WLASL1000	1,000	13,168	13.2	116
WLASL2000	2,000	21,083	10.5	119

Table 1: Summary of the different subsets of WLASL dataset where mean is the average number of video samples per gloss. More details can be found in the paper [25].

4. Experiments

4.1. Dataset

We evaluate our method using recently introduced WLASL dataset [25]. The dataset is curated from online ASL videos, primarily created for teaching purposes. Being collected from different sources, the dataset contains unrestricted varieties in signing styles and background. The authors went through several manual and automated pre-processing steps to create four subsets of data, named WLASL100, WLASL300, WLASL1000 and WLASL2000, respectively. Table 1 shows the statistics of the dataset.

4.2. Preprocessing

We downloaded the data following the instructions come with the dataset release. We obtain estimates of poses for each frame the videos through OpenPose [4] and store the poses. Using the pose information on the image frame, we calculate a bounding box for each frame, making sure that both hands and whole body are visible over the video and crop the frames. After cropping the videos, we adjust the poses according to the cropped region. Finally each cropped video and corresponding adjusted pose form an input to our network.

4.3. Implementation Details

For our pose pooling methods, we used only 2 joints to extract features from the intermediate feature maps of I3D network. These 2 joints are calculated using mean joint location of all 21 finger and palm poses, produced by OpenPose, for each hand. Hence, a feature map of $[f \times t \times h \times w]$ is converted to $[f \times t \times 2]$. We empirically verified that, adding more joints, such as elbow or shoulder, in pooling does not improve results significantly. This is expected due to the fact that, most the variance explaining the data comes from the hand regions. We used maximum pooling using $3 \times 3 \times 3$ kernel around each hand pose location. We also noticed, adding fully connected layer afterwards does not improve or deteriorate the performance. Hence, we decided not to use it. Once we extract pose localized features, train-

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
I3D[25]	65.89	84.11	89.92	56.14	79.94	86.98	47.33	76.44	84.33	32.48	57.31	66.31
PPC-7	67.79	76.91	80.75	57.12	70.80	74.44	44.17	63.33	69.76	29.46	52.95	60.25
PPC-8	67.79	78.16	82.50	59.91	75.78	78.39	44.57	61.11	66.20	29.26	50.35	56.57
I3DLogits	68.70	86.66	89.58	57.62	78.63	82.46	48.29	68.25	73.17	33.18	60.04	68.87
Fusion-1	71.74	81.75	84.66	64.41	78.67	82.39	51.01	70.95	75.80	34.68	60.39	67.27
Fusion-2	74.16	**86.83**	**90.91**	67.79	**84.19**	**87.06**	55.71	77.90	83.77	38.57	**68.17**	75.71
Fusion-3	**75.67**	**86.00**	90.16	**68.30**	83.19	86.22	**56.68**	**79.85**	**84.71**	**38.84**	67.58	**75.71**

Table 2: Top-1, Top-5, top-10 accuracy (%) achieved by each model (by row) on the four WLASL subsets. First row shows the results reported in [25]. Next two rows (PPC 7 & PPC 8) shows performance from two classifiers of our pose localized pooling. The `I3D Logits` result is the basic I3D classifier without any pose pooling mechanism. Here, Fusion-1 = PPC-7 + PPC-8, Fusion-2 = PPC-7 + PPC-8 + I3DLogits and Fusion-3 = PPC-5 + PPC-7 + PPC-8 + I3DLogits.

ing mechanism follows the original I3D [5]. We initialize the I3D network using pre-trained weights on Charades [40] activity dataset. Each video is resized into 256×256 and the poses are adjusted accordingly. We used two video level data augmentation techniques : random cropping using 224×224 spatial support and random horizontal flipping. Input video length is set to 64, with possible temporal augmentation in case of longer videos. We padded the videos less than 64 frames either in the beginning or end. Poses are adjusted appropriately in case of any augmentation of video data. We used Adam optimizer with initial learning rate of 0.001, with 4 or 6 (for different subsets of data) mini-batch sizes. After fixing the hyper-parameters of a model, we train the model using the training and the validation split and report the results on the test split for each subset of the dataset. Micro average accuracy was used as metric due to the imbalance in number of test samples of different classes.

4.4. Evaluation

Table 2 shows the experimental results. The results indicate that our pose localized classifiers (PPCs) perform competitively with the base I3D network. It should be mentioned that, we have a performance gain (68.70% vs 65.89%) using base I3D, shown as `I3DLogits`, over the same baseline's reported in [25]. Although both of these are the same implementation, we believe, the performance increase is due to the cropping preprocessing step. The results indicate that any single `PPC` classifiers perform competitively with the I3D baseline. This shows improvement on memory footprint of the model and computation time, e.g. single `PPC-7` classifier uses smaller number of model parameters than the whole I3D. The results also shows that, fusing prediction score from different branches boost the performance significantly. The best accuracy is achieved by combining four sets of prediction scores from PPC-5, PPC-7, PPC-8 and I3D Logits. Overall, our best performing fusion model (Fusion-3) outperforms I3D implementation in [25] by 10%, 12%, 9.5% and 6.5% on WLASL100, WLASL300, WLASL1000, WLASL2000 subsets respectively.

Complementary Feature Learning We wanted to verify that, improvement from fusing the prediction scores is not just an effect of model ensemble. In this regard, we calculated the accuracy fusing scores from same branch but from separately trained model. For example if we fuse the scores from PPC-7 of two separately trained models, the top-1 accuracy the fusion achieves is 68.05% on WLASL100 subset. Fusing three models from PPC-7 gets 68.17%. Although, these are little better than result from a single PPC-7 (67.79%), they are far worse than the fusion of PPC-7 & PPC-8 (`Fusion-1`, 71.74%). Similarly, fusing scores from two PPC-8 layers from two separately trained models gives 68.15%. This suggests that performance improvement is not merely coming from using multiple models at test time. Rather, different pose localized branches pick on different class specific features and complement each other to obtain better performance.

4.5. Representation Transfer

We conducted this experiment to validate the effectiveness of pose pooled feature on unseen data samples from a different distribution than WLASL. We choose 12 overlapping classes from WLASL300 subset and another private dataset[1] we collected in a laboratory setting. The sign video samples from this dataset and WLASL dataset vary in appearance, lighting condition, distance from camera, hand motion, styles and in many other factors. Figure 3

[1]This dataset will be provided upon request for reproducing the results.

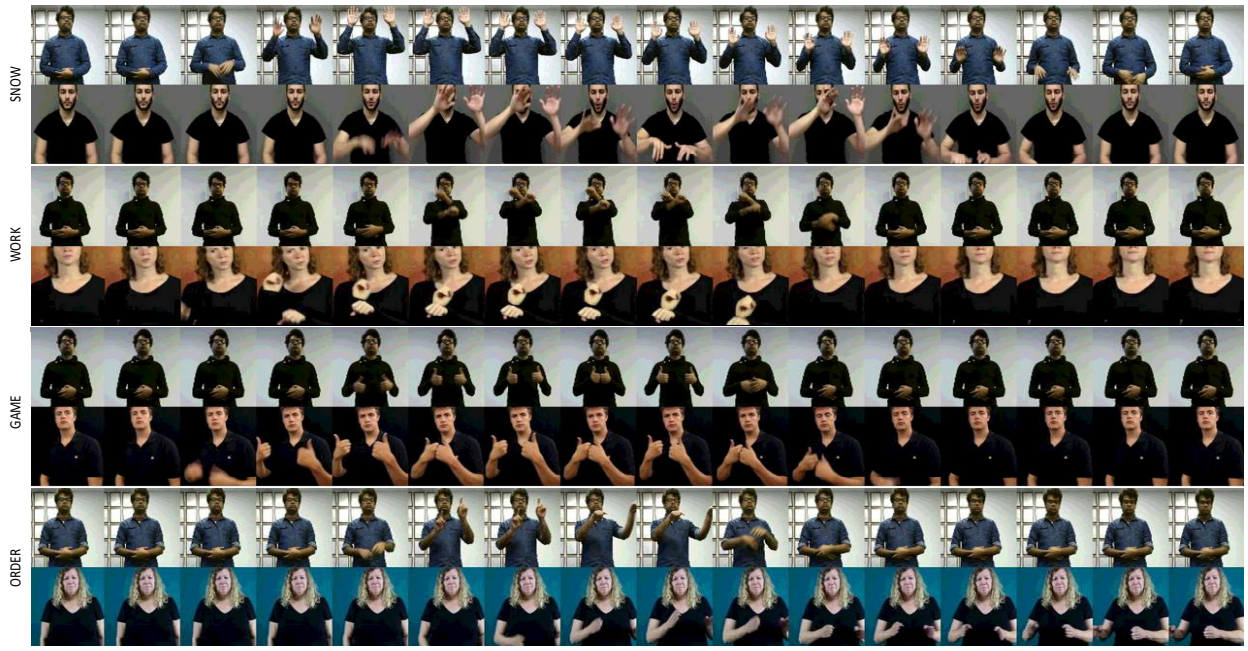


Figure 3: Examples of gesture samples from four classes. Each pair of sample shows one example (label is shown on the left) from WLASL dataset (bottom in a pair) and other different dataset (top in a pair) used in representation transfer experiments in the section 4.5.

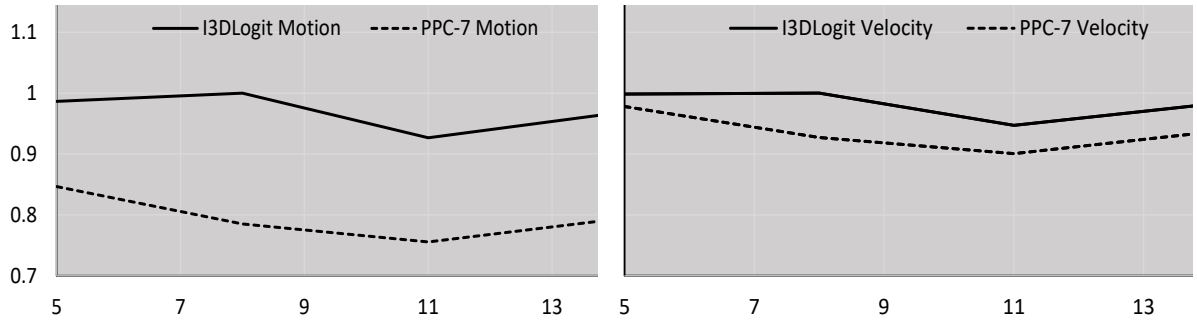
Method	Accuracy		
	top-1	top-5	top-10
I3DLogits	55.63	75.67	82.25
PPC-8	59.21	89.48	88.63
Fusion-3	**66.70**	**91.80**	**98.11**

Table 3: Fine tuned results using only 0.4% of data in training. Method names have same meaning as in Table 2.

shows some examples of from the both datasets. There are total 2976 samples from the other dataset for overlapping 12 classes. We fine tune our models using only a single sample from each of the 12 classes. Rest of the 2964 samples were used for evaluation. Table 3 shows the results of this experiment. The result demonstrates an approximate 10% increase in the top-1 recognition accuracy compared to baseline I3D network. Even without any fusion, our proposed pose pooling method, such as PPC-8, outperforms I3D by 3.58%. This demonstrates that the features learned by the proposed pose pooling methods are more invariant to domain changes than baseline I3D. This is not surprising because, pose guided pooling helps the network to extract features relevant to sign gesture dynamics and also helps to ignore nonessential appearance and motion cues.

4.6. Qualitative Findings

Our approach is built on the hypothesis that pooling features from different layers contributes differently to for overall performance. We also found this in our experiments. Experimental findings reveal that, the classes having less motion get best predictions using features from intermediate layer while the sign gestures with relatively heavy motion benefit from final layer representations of the I3D network. To corroborate this intuition, we first pick top performing classes that have higher fraction of samples correctly classified by `PPC-7` branch. We calculate the hand motion using pose locations of both hands for those selected classes. We repeat the same for top performing classes from the final level logits (`I3DLogits` in Table 2) of the network. Figure 4 (a) shows the plot of hand motion where horizontal axis is the number of top classes we pick. We observe that, calculated average motion of best performing classes for `PPC-7` is always lower. This suggests that, the classes having smaller motion get better features from `PPC-7` than `I3DLogits`. We believe, due to slower motion, hand shapes of these classes are more perceivable to the network and the pose localized pooling helps to extract those informative shapes. Figure 4 (b) shows similar phenomenon, except we calculate average motion of a certain sized window. We calculate distance of the first and the last frame of each window and the window slides over the video



(a) Average hand motion

(b) Average hand velocity

Figure 4: Hand motion of top performing classes from PPC-7 branch and I3D logits.

Method	WLASL300					
	Without Pose Pooling			With Pose Pooling		
	top-1	top-5	top-10	top-1	top-5	top-10
PPC-7	**57.37**	76.33	82.05	57.12	70.80	74.44
PPC-8	55.59	73.08	77.55	**59.91**	75.78	78.39
I3DLogits	**57.62**	78.63	82.46	**57.62**	78.63	82.46
Fusion-1	60.12	80.05	85.25	**64.41**	78.67	82.39
Fusion-2	63.83	83.39	87.69	**67.79**	84.19	87.06
Fusion-3	64.45	83.86	87.58	**68.30**	83.19	86.22

Table 4: Ablation studies of using pose as indexes while pooling activation from feature maps. Fusion methods have similar meaning as Table 2, i.e. Fusion-1 = PPC-7 + PPC-8, Fusion-2 = PPC-7 + PPC-8 + I3DLogits and Fusion-3 = PPC-5 + PPC-7 + PPC-8 + I3DLogits.

with stride 1. This is proportional to the average velocity of hands in a sign sample and we observe the similar results as the motion.

4.7. Ablation Studies

To validate the effectiveness of pose localized features, we implement similar fusion experiment as we described in the result section, but without pose localized pooling. Instead, we use basic maximum pooling sub-sampling. In detail, after extracting a feature map from any point of the network in Figure 2, we using maximum pooling over that feature map to produce a representation of the video. Table 4 shows the results from this experiment. The result indicates, when single layer features are used, pose pooling features perform equally. However, when scores from several layers are combined (fusion cases in Table 4), pose pooling features provide around 4% performance gain in `top-1` accuracy. This suggests predictions made using pose pooled representations exploit alternative information across different layers of the network.

5. Conclusion

In this work, we propose a novel pose guided pooling strategy for extraction of additional features from 3D ConvNet in the context of world level sign language recognition. Our experiments show that, combining features from different levels of the network can improve overall recognition accuracy. For future direction, our goal is to consider phrase level sign language modeling. We plan to use this work to localize sign words in phrase level sign language recognition tasks.

References

[1] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 584–592, Bejing, China, 22–24 Jun 2014. PMLR.

[2] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, , 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, June 2008.

[3] Congqi Cao, Yifan Zhang, Chunjie Zhang, and Hanqing Lu. Action recognition with joints-pooled 3d deep convolutional descriptors. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3324–3330. AAAI Press, 2016.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[5] João Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *CoRR*, abs/1705.07750, 2017.

[6] G. Chéron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3218–3226, 2015.

[7] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[9] Yong Du, W. Wang, and L. Wang. Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, June 2015.

[10] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal Cycle-Consistency Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[12] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. Using k-Poselets for Detecting People and Localizing Their Keypoints, booktitle = Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. CVPR '14, pages 3582–3589, Washington, DC, USA, 2014. IEEE Computer Society.

[13] Dan Guo, Shengeng Tang, and Meng Wang. Connectionist Temporal Modeling of Video and Language: A Joint Model for Translation and Sign Labeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 751–757. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[14] Dan Guo, Shuo Wang, Qi Tian, and Meng Wang. Dense Temporal Convolution Network for Sign Language Translation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 744–750. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[15] A. Hosain, P. Santhalingam, P. Pathak, H. Rangwala, and J. Kosecka. Finehand: Learning hand shapes for american sign language recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 397–404, Los Alamitos, CA, USA, may 2020. IEEE Computer Society.

[16] Al Amin Hosain, Panneer Selvam Santhalingam, Parth Pathak, Jana Kosecka, and Huzefa Rangwala. Sign Language Recognition Analysis using Multimodal Data. In *International Conference on Data Science and Advanced Analytics (DSAA'19)*, 2019.

[17] http://wfdeaf.org/. World federation of deaf : Our work, 2020.

[18] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Sign Language Recognition using 3D convolutional neural networks. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, June 2015.

[19] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model, 2016.

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, Jan. 2013.

[21] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, Jan. 2013.

[22] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Ahmed Sohel, and Farid Boussaïd. A New Representation of Skeleton Sequences for 3D Action Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4570–4579, 2017.

[23] Oscar Koller, Hermann Ney, and Richard Bowden. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] S. Kulkarni. Appearance based recognition of american sign language using gesture segmentation.

[25] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020.

[26] Jian Liu, Naveed Akhtar, and Ajmal Mian. Skepxels: Spatio-temporal Image Representation of Human Skeleton Joints for Action Recognition. *CoRR*, abs/1711.05941, 2017.

[27] J. Liu, A. Shahroudy, D. Xu, A. Kot Chichung, and G. Wang. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2017.

[28] J. Liu, G. Wang, L. Y. Duan, K. Abdiyeva, and A. C. Kot. Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, April 2018.

[29] Kun Liu, Wu Liu, Chuang Gan, Mingkui Tan, and Huadong Ma. T-C3D: Temporal Convolutional 3D Network for Real-Time Action Recognition, 2018.

[30] Dimitris Metaxas, Mark Dilsizian, and Carol Neidle. Linguistically-driven Framework for Computationally Efficient and Scalable Sign Recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[31] Dimitris Metaxas, Mark Dilsizian, and Carol Neidle. Scalable ASL sign recognition using model-based machine learning and linguistically annotated corpora, 2018.

[32] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 527–544, Cham, 2016. Springer International Publishing.

[33] Pradyumna Narayana, Ross Beveridge, and Bruce A. Draper. Gesture Recognition: Focus on the Hands. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[34] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards Accurate Multi-person Pose Estimation in the Wild, 2017.

[35] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. Sign Language Recognition Using Convolutional Neural Networks. In *Computer Vision - ECCV 2014 Workshops*. Springer International Publishing, 2015.

[36] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[37] Junfu Pu, Wengang Zhou, and Houqiang Li. Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 885–891. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[38] Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative Alignment Network for Continuous Sign Language Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[39] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, June 2016.

[40] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *CoRR*, abs/1604.01753, 2016.

[41] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc., 2014.

[42] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. In *AAAI Conference on Artificial Intelligence*, pages 4263–4270, 2017.

[43] Thad Starner, Alex Pentland, and Joshua Weaver. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1371–1375, Dec. 1998.

[44] Chao Sun, Tianzhu Zhang, and Changsheng Xu. Latent Support Vector Machine Modeling for Sign Language Recognition with Kinect. *ACM Trans. Intell. Syst. Technol.*, 6(2):20:1–20:20, Mar. 2015.

[45] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[46] Graham W. Taylor, Rob Fergus, Yann LeCun, and Christopher Bregler. Convolutional Learning of Spatio-Temporal Features. In *Proceedings of the 11th European Conference on Computer Vision: Part VI*, ECCV'10, page 140–153, Berlin, Heidelberg, 2010. Springer-Verlag.

[47] A. Thangali, J. P. Nash, S. Sclaroff, and C. Neidle. Exploiting phonological constraints for hand shape inference in ASL video. In *CVPR 2011*, pages 521–528, June 2011.

[48] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features With 3D Convolutional Networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

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

[50] G. Varol, I. Laptev, and C. Schmid. Long-Term Temporal Convolutions for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1510–1517, 2018.

[51] V. Veeriah, N. Zhuang, and G. J. Qi. Differential Recurrent Neural Networks for Action Recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4041–4049, Dec 2015.

[52] Limin Wang, Yu Qiao, and Xiaoou Tang. Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors. *CoRR*, abs/1505.04868, 2015.

[53] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV*, 2016.

[54] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *CVPR*, 2016.

[55] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018.

[56] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[57] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond Short Snippets: Deep Networks for Video Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[58] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. American Sign Language Recognition with the Kinect. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pages 279–286, New York, NY, USA, 2011. ACM.

[59] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. American Sign Language Recognition with the Kinect. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, page 279–286, New York, NY, USA, 2011. Association for Computing Machinery.

[60] Zahoor Zafrulla, Helene Brashear, Pei Yin, Peter Presti, Thad Starner, and Harley Hamilton. American Sign Language Phrase Verification in an Educational Game for Deaf Children, 08 2010.