

MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language

Hamid Reza Vaezi Joze
hava@microsoft.com

Oscar Koller
Oscar.Koller@microsoft.com

Microsoft
Redmond, WA, USA

Microsoft
Munich, Germany

Abstract

Sign language recognition is a challenging and often underestimated problem comprising multi-modal articulators (handshape, orientation, movement, upper body and face) that integrate asynchronously on multiple streams. Learning powerful statistical models in such a scenario requires much data, particularly to apply recent advances of the field. However, labeled data is a scarce resource for sign language due to the enormous cost of transcribing these unwritten languages.

We propose the first real-life large-scale sign language data set comprising over 25,000 annotated videos, which we thoroughly evaluate with state-of-the-art methods from sign and related action recognition. Unlike the current state-of-the-art, the data set allows to investigate the generalization to unseen individuals (signer-independent test) in a realistic setting with over 200 signers. Previous work mostly deals with limited vocabulary tasks, while here, we cover a large class count of 1000 signs in challenging and unconstrained real-life recording conditions. We further propose I3D, known from video classifications, as a powerful and suitable architecture for sign language recognition, outperforming the current state-of-the-art by a large margin. The data set is publicly available to the community.

1 Introduction

In the US, approximately around 500,000 people use American sign language (ASL) as primary means of communication [49]. ASL is also used in Canada, Mexico and 20 other countries. Just like any other natural language, it covers its unique vocabulary as well as a grammar which is different from spoken English. We are intrigued by sign language and accessibility for the Deaf and believe sign recognition is an exciting field that offers many challenges for computer vision research.

For decades, researcher from different fields have tried to solve the challenging problem of sign language recognition. Most of the proposed approaches rely on external devices such as additional RGB [4] or depth cameras [57, 73], sensor [44, 43] or colored gloves [72]. However, such requirements limit the applicability to specific settings where such resources are available. Opposed to that, non-intrusive and purely vision based sign recognition will allow for general usage. With the appearance of deep learning based methods and their powerful performance on computer vision tasks, the requirements on training data have changed dramatically from few hundred to thousands of samples being needed to train strong models.

Unfortunately, public large scale sign language resources suitable for machine learning are very limited and there is currently no public ASL data set big enough to evaluate recent deep learning approaches. This prevents recent computer vision trends to be applied to this field. As such, our goal is to advance the sign language recognition community and the related state-of-the-art by releasing a new large-scale data set, establishing thorough baselines and carrying over recent computer vision trends. With this work, we make the following contributions: (1) We release the first large-scale ASL data set called MS-ASL that covers over 200 signers, signer independent sets, challenging and unconstrained recording conditions and a large class count of 1000 signs ¹. (2) We evaluate current state-of-the-art approaches: 2D-CNN-LSTM, body key-point, CNN-LSTM-HMM and 3D-CNN as baselines. (3) We propose I3D (known from action recognition) as a powerful and suitable architecture for sign language recognition that outperforms previous state-of-the-art by a large margin and provide new pre-trained model for it. (4) We estimate the effect of number of classes and number of samples on the performance.

2 Previous Works

Recognition methods: Researchers have tried to solve the challenges of sign language recognition in different ways. In 1983, the first work was a glove based device that allowed to recognize ASL fingerspelling based on a hardwired circuit [60]. In the meantime, there have been a lot of related approaches which rely on tracked hand movements based on sensor gloves for sign recognition [13, 26, 42, 43, 53]. Some works extended this by adding a camera as a new source of information [49] and they showed that adding video information improves the accuracy of detection but the method mainly relies on the glove sensors.

In 1988, Tamura *et al.* [62] were the first to follow vision-based sign language recognition. They built a system to recognize 10 isolated signs of Japanese sign language using simple color thresholding. Because the sign is performed in 3-dimensions, many vision based approaches use depth information [43, 62, 75] or multiple cameras [9]. Some rely on colored gloves to ease hand and finger tracking [46, 72]. In this paper we focus on non-intrusive sign language recognition using only a single RGB camera as we believe this will allow to design tools for general usage to empower everybody to communicate with a deaf person using ASL. The sole use of RGB for sign detection is not new, traditional computer vision techniques particularly with Hidden Markov Models [23, 61, 62, 71], mainly inspired by improvements in speech recognition, have been in use in the past two decades. With the advances of deep learning and convolutional networks for image processing the field has evolved tremendously. Koller *et al.* showed large improvements embedding 2D-CNNs in HMMs [53, 40], related works with 3D-CNNs exist [6, 57] and weakly supervised multi-stream modeling is presented in [40]. However, sign language recognition still lags behind related fields in the adoption of trending deep learning architectures. To the best of our knowledge no prior work exists that leverages latest findings from action recognition with I3D networks or complete body key-points which we will address with this paper.

Sign language data sets: Some outdated reviews of sign language corpora exist [46]. Below, we have reviewed sign language data sets with explicit setups intended for reproducible pattern recognition research. The **Purdue RVL-SLLL ASL database** [35, 74] contains 10 short stories with a vocabulary of 104 signs and a total sign count of 1834 produced by 14 native signers in a lab environment under controlled lighting. The RWTH-BOSTON corpora were originally created for linguistic research [2] and packaged for pattern recognition purposes later. The **RWTH-BOSTON-50** [73] and the **RWTH-BOSTON-104**

¹Instructions and download links: <https://www.microsoft.com/en-us/research/project/ms-asl/>

corpus contain isolated sign language with a vocabulary of 50 and 104 signs. The **RWTH-BOSTON-400** corpus contains a vocabulary of 483 signs and also constitutes of continuous signing by 5 signers. The **SIGNUM corpus** [71] provides two evaluation sets: first a multi-signer set with 25 signers, each producing 603 predefined sentences with 3703 running gloss annotation and a vocabulary of 455 different signs. Second, it has a single signer setup where the signer produces three repetitions of the given sentences. In the scope of the DictaSign project, multi-lingual sign language resources have been created [8, 23, 37]. However, the produced corpora are not well curated and made available for reproducible research. The **Greek Sign Language (GSL) Lemmas Corpus** [2] constitutes such a data collection. It provides a subset with isolated sign language (single signs) that contains 5 repetitions of the signs produced by two native signers. However, different versions of this have been used in the literature disallowing fair comparisons and the use as benchmark corpus. The corpus has been referred to with 1046 signs [64, 65], with 984 signs [15] and with 981 signs [62]. Additionally, a continuous 100 sign version of the data set has been used in [65]. The reason for all these circulating subsets is that the data has not been made publicly available. **DEVISIGN** is a Chinese sign language data set featuring isolated single signs performed by 8 non-natives [12] in a laboratory environment (controlled background). The data set is organized in 3 subsets, covers a vocabulary of up to 2000 isolated signs and provides RGB with depth information in 24000 recordings. A more recent data set [53] covers 100 continuous chinese sign language sentences produced five times by 50 signers. It has a vocabulary of 178 signs. It can be considered as staged recording. The **Finish S-pot** sign spotting task [59] is based on the controlled recordings from the Finish sign language lexicon [27]. It covers 1211 isolated citation form signs that need to be spotted in 4328 continuous sign language videos. However, the task has not been widely adopted by the field. The **RWTH-PHOENIX-Weather 2014** [29, 32] and **RWTH-PHOENIX-Weather 2014 T** [8] are large scale real-life sign language corpora that feature professional interpreters recorded from broadcast news. They cover continuous German sign language with a vocabulary of over 1000 signs, about 9 hours of training data. It only features 9 signers and limited computer vision challenges. There are several groups which experimented with their own private data collection resulting in corpora with quite limited size in terms of total number of annotations and vocabulary such as **UWB-07-SLR-P corpus of Czech sign language** [8], data set by Barabara Loedings Group [46, 51] and small scale corpora [0, 24].

Data set	classes	signer	independ.	videos	real-life
Purdue ASL [74]	104	14	no	1834	no
Video-Based CSL [63]	178	50	yes	25000	no
Signum [71]	465	25	yes	15075	no
RWTH-Boston [46]	483	5	no	2207	no
RWTH-Phoenix [29]	1080	9	no	6841	yes
Devisign [12]	2000	8	no	24000	no
This work	1000	222	yes	25513	yes

Table 1: Comparison of public sign language data sets.

Table 1 summarizes the mentioned data sets. To the best of our knowledge RWTH-PHOENIX-Weather 2014 and DEVISIGN are currently the only publicly available data sets that are large enough to cover recent deep learning approaches. However, both data sets are lacking the variety and number of signers to advance the state-of-the-art with respect to the important issue of signer independence and computer vision challenges from natural unconstrained recordings. In the scope of this work, we propose the first ASL data set that covers over 200 signers, signer independent sets, challenging and unconstrained recording conditions and a large class count of 1000 gloss level signs.

3 Proposed ASL Data Set

Since there is no public ASL data set suitable for large-scale sign language recognition, we looked for realistic data sources. The deaf community actively uses public video sharing platforms for communication and study of ASL. Many of those videos are captured and uploaded by ASL students and teachers. They constitute challenging material with large variation in view, background, lighting and positioning. Also from a language point of view, we encounter regional, dialectal and inter-signer variation. This seems very appealing from a machine learning point of view as it may further close the gap in learning signer independent recognition systems that can perform well in realistic circumstances. Besides having access to well suited data, the main issue remains labeling which requires skilled ASL natives.

We noticed that a lot of the public videos have manual subtitles, captions, descriptions or a video title that indicates which signs are being performed in it. We therefore decided to access the public ASL videos and obtain the text from all those sources. We process these video clips automatically in three distinct ways: (1) For longer videos, we used Optical Character Recognition (OCR) to find printed labels and their time of occurrence. (2) Longer videos may contain video captions that provide the sign descriptor and the temporal segmentation. (3) In short videos we obtained the label directly from the title.

In the next step, we detected bounding boxes and used face recognition to find and track the signer. This allowed identification of descriptions that refer to a static image rather than an actual signer. If we identified multiple signers performing one after the other, we split the video up into smaller samples.

In total we accessed more than 45,000 video samples that include words or phrases in their descriptions. We sorted the words based on frequency to find the most frequently used ones while removing misspellings and OCR mistakes. Since many of the ASL vocabulary publicly accessible videos belong to teachers performing a lesson vocabulary or students doing their homework, all top hundred words belong to ASL tutorial books [68, 73] vocabulary units. ASL-Lex has also been identified as source of quality recordings [10].

3.1 Manual Touch-up

Although many of the sample videos are good for training purposes, some of them include the instruction to the sign or several repeated performances with long pause in between. Therefore, we decided to manually trim all video samples with a duration of more than 8 seconds. For higher accuracy on the test set, we chose the threshold to be 6 seconds there. Although our annotators were not native in ASL, they could easily trim these video samples while considering other samples of the same label. We also decided to review video samples shorter than 20 frames. In this way, around 25% of the data set was manually reviewed. Figure 1 illustrates a histogram of the duration of the 25,513 video samples of signs after the manual touch-up. There are unusual peaks for multiples of 10 frames which seems to be caused by video editing software cutting and adding captions, which favors such duration. Despite that, the histogram looks like a Poisson distribution with the average of 60. Combined, the duration of the video samples is just over 24 hours long.

3.2 ASL synonyms

Sign languages all over the world are independent, fully fledged languages with their own grammar and word inventory, distinct from the related spoken language. Sign languages can be characterized as unwritten languages and have no commonly used written form. Therefore, a written word will usually represent a semantic description and refer to the meaning of

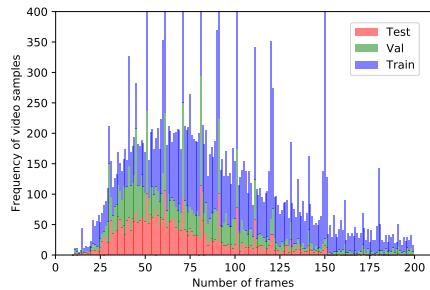


Figure 1: Histogram of frame numbers for ASL1000 video samples.

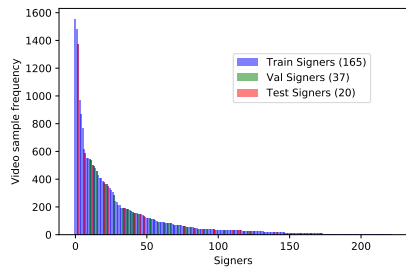


Figure 2: Showing the number of video samples for each of the 222 signers and the train/test/validation split of proposed data set.

a sign, not to the way it is executed. This is fundamentally different to most writing schemes of spoken languages. As an example, look at the two English words *Clean* and *Nice*. While they are clearly distinct in English, they have similar signs in ASL which share the same hand gesture and movement. On the other hand, the English word *Run* has a distinct sign for each of its meaning such as "walk and run", "run for office", "run away" and "run a business" [68]. With respect to the ASL videos we accessed from the internet and their descriptions, we needed to make sure that similar ASL signs merged to one class for training even if they have distinct English descriptors. This process was implemented based on a reference ASL Tutorial books [78]. This mapping of sign classes will be released as part of the MS-ASL data set. ASL-Lex has also been identified as source of quality recordings [80].

3.3 Signer Identification

Signer dependency is one of the most blocking challenges with current non-intrusive sign recognition approaches. To address this issue, our goal is to create a recognition corpus which covers signer independent sets. We want to ensure that the signers occurring in train, validation and test are distinct. Therefore, we aimed at identifying the signer in each sample video. To achieve this, we computed 5 face embeddings [66] for each video sample. Based on this, the video samples were then clustered into 457 clusters. Some of these clusters were merged later by using the prior knowledge that two consecutive samples from a video tend to have the same signer. Additionally, we manually labeled the low confidence clusters. Finally, we ended up having 222 distinct signers. The found individuals occurred in the corpus with very diverse frequency. We have 3 signers with more than one thousand video samples and 10 signers with a single video sample each. We then solve an optimization problem to distribute signers into train, validation and test set signers aiming to divide data set partitions to 80%, 10% and 10% for train, validation and test, respectively. However, due to the signer independency constraint and unbalances samples, an exact division into these sizes was impossible. We relaxed this condition, maintaining at least one sample in each set for each class. The final amount of signers in each of the sets was 165, 37 and 20 for train, validation and test, respectively. Figure 2 shows the frequency of samples by all 222 signers and the train/validation/test split.

3.4 MS-ASL Data Set with 4 Subsets

In order to have a good understanding of the ASL vocabulary and being a comprehensive benchmark for classifying signs with diverse training samples, We release 4 subsets including 100, 200, 500 and 1000 most frequent words. Each includes their own train, test and

Data set	Class	Subjects	Number of Videos				Duration [hours:min]	Videos per class	
			Train	Validation	Test	Total		Min	Mean
ASL100	100	189	3789	1190	757	5736	5:33	47	57.4
ASL200	200	196	6319	2041	1359	9719	9:31	34	48.6
ASL500	500	222	11401	3702	2720	17823	17:19	20	35.6
ASL1000	1000	222	16054	5287	4172	25513	24:39	11	25.5

Table 2: Showing statistics of the 4 proposed subsets of the MS-ASL data set.

validation sets. All these sets are signer independent and the signers for train (165), test (20) and validation (37) are the same as shown in Figure 2, therefore smaller sets are subset of the larger. We call these subsets *ASL100*, *ASL200*, *ASL500* and *ASL1000* for the rest of this paper. Table 2 shows the characteristics of each of these sets. In *ASL100*, there are at least 45 samples for each class while in *ASL1000* there are at least 11 samples for each class.

Data Set Challenges: There are challenges in this data set which make it unique compared to other sign language data sets and more challenging compared to video classification data sets: (1) One sample video may include repetitive act of a distinct signs. (2) One word can sign differently in different dialects based on geographical regions. As an example, there are 5 common signs for the word *Computer*. (3) It includes large number of signers and is a signer independent data set. (4) They are large visual variabilities in the videos such as background, lighting, clothing and camera view point.

3.5 Evaluation Scheme

We suggest two metrics for evaluating the algorithms ran in these data sets: 1) average per class accuracy, 2) average per class top-five accuracy. We prefer per class accuracy compared to plain accuracy to better reflect performance given the unbalance test set inherited from the unbalance nature of the data set. To be more precise, we compute the accuracy of each class and reported the average value. In the top-5 accuracy, we call it correct if the ground-truth label appears in the top five guesses of the method being evaluated. We compute top-five accuracy for each class and report the average value. ASL, just like any other language can have ambiguity which can be resolved in context. Therefore, we picked top-five accuracy.

4 Baseline Methods

Although it is much more challenging, but we can consider isolated sign language recognition similar to action recognition or gesture detection as it is a video classification task for a human being. We can categorize current action recognition or gesture detection into three major categories or combination of them 1) Using 2D convolution on image and do a recurrent network on top of that [20, 25] 2) Extracting subject’s body joints in the form of skeleton and using skeleton data for recognition [21, 77] 3) Using 3D convolution [10, 50, 56]. In order to have baselines from each categories of human action recognition, we implement at least one method for each of these categories.

For all of the methods, we use the bounding box covering the signer as input image. We extract the person bounding box by the SSD network [45] and release it for each video sample as part of MS-ASL data set. For spatial augmentations, body bounding boxes are randomly scaled or translated by 10%, fit into a square and re-sized to fixed 224×224 pixels. We picked 64 as our temporal window which is the average number of frames across

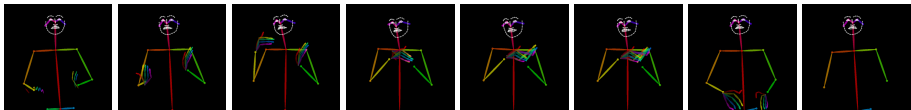


Figure 3: Extracted 137 body key-points for a video sample from MS-ASL by [9, 57].

all sample videos. In addition, the resulted video is randomly but consistently (per video) flipped horizontally because ASL is symmetrical and can be performed by either hands. We used fixed sized frame number as well as fixed size resolution for 2D and 3D convolution methods. For temporal augmentations: 64 consecutive frames are picked randomly from the videos and shorter videos are randomly elongated by repeating their first or last frame. We train for 40 epochs. In this paper, we focused on RGB only algorithms and did not use optical flow for any of the implementations. It is a proven fact that using optical flow as second stream in train and test stage [17, 25, 58] or just train stage [10] boosts the performance of prediction. Herein, we describe the methods used for determining baselines.

2D-CNN : The high performance of 2D convolutional networks on image classification makes them the first candidate for video processing. This is achieved by extracting features from each frame of the video independently. The first approach was to combine these features by simply pooling the prediction, but it ignored the frame ordering or timing. The next approach which proved more successful, was using recurrent layers on the top of 2D convolution networks. Motivated by [20], we picked LSTM [61] as our recurrent layer which records the temporal ordering and long range dependencies by encoding the states. We used VGG16 [59] network followed by an average pooling and LSTM layer of size 256 with batch normalization. The final layers are a 512 hidden units followed by a fully connected layer for classification. We considered the output on final frame for testing. We also have implemented [39] as the state-of-the-art on PHOENIX2014 data set [29, 57]. This method use GoogleNets [63] as 2D-CNN with 2 bi-directional LSTM layers and 3-state HMM. We report it as Re-Sign in the experimental result.

Body Key-Points : With the introduction of robust body key-points (so-called skeleton) detection [13], some studies try to solve human action recognition by body joints only [21, 27] or use body joints along with the image stream [14]. Since most body key-point techniques did not cover hand details, it was not rational to use it for sign language recognition task as it relies heavily on the movement of fingers. But a recent work has covered hand and face key-points along with classical skeleton [57]. We leveraged this technique which extracted 137 key-points in total, to do a baseline on our data set by body key-points. We extracted all the key-points for all samples using [9, 57]. Using 64 frames for time window, our input to the network would be $64 \times 137 \times 3$ representing x, y coordinates and confidence values for the 137 body key-points for all consecutive 64 frames. Figure 3 illustrates the extracted 137 body key-points for a video sample from MS-ASL. The hand key-points are not as robust as body and face.

We implemented hierarchical co-occurrence network (HCN) [17] which originally used 15 joints. We extended this work by using 137 body key-points including hand and face key-points. The input to this network is original 137 body key-points as well as per frame difference of them. The network includes three layers of 2D convolution on top of each input as well as two extra 2D convolution layers after the concatenation of two paths.

3D-CNN : Recently, 3D convolutional networks have shown promising performance for video classification and action recognition including C3D network [66] and I3D net-

Method	ASL100	ASL200	ASL500	ASL1000
Naive Classifier	0.99	0.50	0.21	0.11
VGG+LSTM [14], [18]	13.33	7.56	1.47	-
HCN [27]	46.08	35.85	21.45	15.49
Re-Sign [69]	45.45	43.22	27.94	14.69
I3D [10]	81.76	81.97	72.50	57.69

Table 3: The average per class accuracy for baseline method on proposed ASL data sets.

work [10]. We applied C3D [66] released code from author as well as our own implemented version to our proposed data sets with and without pre-trained model, trained on Sport-1M [67]. The model did not converge for any of our experiments. We adopted the architecture of I3D networks proposed in [10] and employed its suggested implementation details. This network is an inflated version of Inception-V1 [24], which contains several 3D convolutional layers followed by 3D max-pooling layers and inflated Inception-V1 submodules. We started with pre-trained network trained on Imagenet [19] and Kinetics [10]. We optimized the objective functions with standard SGD with momentum set to 0.9. We began the base learning rate at 10^{-2} with a $10\times$ reduction at epoch 20 when validation loss saturated.

5 Experimental Result

We trained all of the methods mentioned in section 4 on four MS-ASL subsets (*ASL100*, *ASL200*, *ASL500* and *ASL1000*) and computed the accuracy for test set which includes subjects that are not included in training phase. As described in subsection 3.5, we report two evaluation metrics: average per class accuracy and average per class top-five percent accuracy. The results are reported in Table 3 and Table 4 respectively. We did not over optimize training parameters. Hence, these results constitute baseline for 2D-CNN, 3D-CNN, CNN-LSTM-HMM and body key-point based approaches. The experimental result suggests that this data set is very difficult for 2D-CNN or at least LSTM could not propagate the recurrent information well. In video classification data sets such as UCF101 [60] or HMDB51 [42], the image itself carries context information regarding the classification while in MS-ASL there is minimum context information in a single image. Re-Sign [69] which report as state-of-the-art in few sign language dataset could not achieve well for challenging MS-ASL (This method could not predict top-five). Body key-point based approach (HCN) is doing relatively better compared to 2D-CNN but there is huge room for improvement because of network simplicity as well as future improvements for hand key-point extraction. On the other hand our 3D-CNN baseline did achieve good results in this challenging, uncontrolled data set and we propose it as powerful network for sign language recognition.

Method	ASL100	ASL200	ASL500	ASL1000
Naive classifier	4.86	2.49	1.05	0.58
VGG+LSTM [14], [18]	33.42	21.21	5.86	-
HCN [27]	73.98	60.29	43.83	32.50
I3D [10]	95.16	93.79	89.80	81.08

Table 4: The average per class top-five accuracy for baseline methods on proposed data sets.

5.1 Qualitative Discussion

Figure 5 illustrates the confusion matrix obtained by comparison of the grand-truth labels and the predicted labels from models trained by I3D on *ALS200* data set. As we expected,

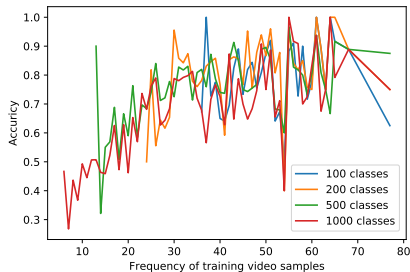


Figure 4: The accuracy of trained models based on frequency of training samples.

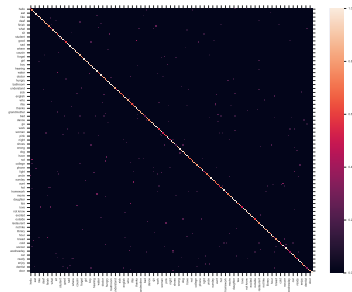


Figure 5: The confusion matrix for I3D on ALS200 data set.

most of the values lay on the diagonal element. Here is the list of brightest points off the diagonal with value of more than .25 which represents per class worst predictions:

- *Good* labeled as *Thanks* (.4): often the sign *Good* is done without the base hand, this sign can mean *Thanks* or *Good*
- *Water* labeled as *Mother* (.33): both by placing dominant hand around chin area while the detail is different.
- *Today* labeled as *Now* (.33): two versions for *Today* one of them is signing *Now* twice.
- *Not* labeled as *Nice* (.33), *Aunt* labeled as *Nephew* (.33), *Tea* labeled as *Nurse* (.33)
- *Start* labeled as *Finish* (.3)
- *My* labeled as *Please* (.28): both sign by place the dominant hand on the chest. A clockwise motion for *Please* and gentle slapping for *My*.

We did similar investigation for other data sets and find interesting evidence about language ambiguity that could solve within the context. Therefore, the error of the model is combination of language ambiguity and prediction error. Our observation shows when we have smaller training sets, model error mainly come from prediction errors but for classes with more samples the error could come from language ambiguity. This advise us to use five-top as our second metric since eventually these predication need to feed to language model with context.

5.2 The Effect of Pre-Trained Model

The fact that I3D training on *ASL200* outperformed I3D trained on *ASL100* was not convincing as it contains twice the classes as *ASL100*. We verified this result with further experiments. We evaluate the I3D model trained with the *ASL200* data on *ASL100* to study the effect of data. The average per class accuracy reaches 83.36% which made the results less convincing. The only proposed explanation is the lack of adequate training video samples which is less than four thousands. This prompted us to do a new experiment; We trained I3D on *ASL100* using the same setting as the last experiments except for using *ASL200* as pre-trained model instead of ImageNet+Kinetics pre-trained model. The result was 85.32% for average per class accuracy and 96.53% for average per class top-five accuracy which is more than 3.5% performance boost. This is a valid experimental approach as the test and train are still separated due to signer independency. This verifies our reasoning and suggests that the existing out-of-domain pre-trained models can be easily outperformed by in-domain pre-trained models specific for sign language recognition. We proposed the model trained on MS-ASL as a I3D pre-trained model for sign language recognition tasks.

5.3 The Effect of Number of Classes

In order to evaluate the effect of number of classes in model prediction, we tested the I3D model trained on *ASL1000* training sets on *ASL500*, *ASL200* and *ASL100* test sets. This allowed a comparison between the model trained on 100 classes with the one trained with 1000 classes on the same test set. We did similar experiments with all possible pairs and reported the average per class accuracy on Table 5. In this table we show subsets of the MS-ASL data set on the horizontal axis and the tested subsets on the vertical axis. Increasing the number of classes decreased the accuracy of either the train or the test phase. Doubling the size of test classes led to a small change from 83.36% to 81.97% and doubling the size of the train classes from 85.32% to 83.36%. This suggests that the observed effect is significantly less when we have more video samples per class.

I3D trained on	ASL100	ASL200	ASL500	ASL1000
ASL100	85.32%	-	-	-
ASL200	83.36%	81.97%	-	-
ASL500	80.61%	78.73%	72.50%	-
ASL1000	75.38%	74.78%	68.49%	57.69%

Table 5: Showing the average per class accuracy of the model trained on different subsets of the MS-ASL data set (horizontal), subsets tested on (vertical).

5.4 The Effect of Number of Video Samples

In order to determine the adequate number of video samples per word needed to a good model, we experimented with the number of samples illustrated figure 4. It shows the accuracy of the models based on frequency of training data for our experiments on test data. It shows a somewhat similar curve for all the four experiments suggesting that the accuracy correlates directly to the number of training video samples for classes with less than 40 video samples. However, for classes with more than 40 video samples, the difficulty of the signs may be more important. Although we have average accuracy of 80% for classes with more than 40 training video samples, it does not suggest that 40 is the sweet spot. Direct comparison cannot be made as this dataset lacks other classes which are significantly larger than 40 video samples. The curve deep at $x = 54$ for all networks belongs to the class *Good* which is the only class with 54 training samples. We have discussed this in subsection 5.1.

6 Conclusion

In this paper, we proposed the first large-scale ASL data set with 222 signers and signer independent sets. Our dataset contains a large class count of 1000 signs recorded in challenging and unconstrained conditions. We evaluated the state-of-the-art network architectures and approaches as the baselines on our data set and demonstrated that I3D outperforms current state-of-the-art methods by a large margin. We also estimated the effect of number of classes on the recognition accuracy.

For future works, we propose applying optical flow on the videos as it is a strong information extraction tool. We can also try leveraging body key-points and segmentation on the training phase only. We believe that the introduction of this large-scale data set will encourage and enable the sign language recognition community to catch up with latest computer vision trends.

References

- [1] Mahdi Abavisani, Hamid Vaezi Joze, and VM Patel. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [2] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. The american sign language lexicon video dataset. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.
- [3] Annelies Braffort, Laurence Bolot, Emilie Chételat-Pelé, Annick Choisier, Maxime Delorme, Michael Filhol, Jérémie Segouat, Cyril Verrecchia, Flora Badin, and Nadège Devos. Sign Language Corpora for Analysis, Processing and Evaluation. In *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, pages 453–456, Valletta, Malta, May 2010. ELRA.
- [4] Helene Brashear, Thad Starner, Paul Lukowicz, and Holger Junker. Using multiple sensors for mobile sign language recognition. Georgia Institute of Technology, 2003.
- [5] Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793, Salt Lake City, UT, June 2018.
- [6] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Using Convolutional 3D Neural Networks for User-Independent Continuous Gesture Recognition. In *Proc. Int. Conf. on Pattern Recognition Workshops (ICPRW)*, pages 49–54, Cancun, Mexico, December 2016.
- [7] Necati Cihan Camgoz, Ahmet Alp Kindiroglu, Serpil Karabuklu, Meltem Kelepir, Ayse Sumru Ozsoy, and Lale Akarun. BosphorusSign: A Turkish Sign Language Recognition Corpus in Health and Finance Domains. In *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, pages 1383–1388, May 2016.
- [8] Pavel Campr, Marek Hruží, and Jana Trojanová. Collection and preprocessing of czech sign language corpus for sign language recognition. In *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, pages 3175–3178, May 2008.
- [9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.
- [10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017.
- [11] Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg, and Karen Emmorey. ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods*, 49(2):784–801, April 2017. ISSN 1554-3528. doi: 10.3758/s13428-016-0742-0.
- [12] Xiujuan Chai, Hanjie Wang, and Xilin Chen. The DEVISIGN Large Vocabulary of Chinese Sign Language Database and Baseline Evaluations. Technical report, Key Lab of Intelligent Information Processing of Chinese Academy of Sciences, 2014. 00000.
- [13] C. Charayaphan and A. E. Marble. Image processing system for interpreting motion in American Sign Language. *Journal of Biomedical Engineering*, 14(5):419–425, September 1992. ISSN 0141-5425. doi: 10.1016/0141-5425(92)90088-3.
- [14] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *CVPR 2018*, 2018.

- [15] H. Cooper, N. Pugeault, and R. Bowden. Reading the signs: A video based sign dictionary. In *Proc. IEEE Int. Conf. on Computer Vision Workshops (ICCVW)*, pages 914–919. IEEE, November 2011. doi: 10.1109/ICCVW.2011.6130349.
- [16] Helen Cooper and Richard Bowden. Sign Language Recognition using Linguistically Derived Sub-Units... In *LREC Workshop on the Representation and Processing of Sign Languages*, pages 57–61, Valletta, Malta, 2010.
- [17] R. Cui, H. Liu, and C. Zhang. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia*, 0:1–1, 2019. ISSN 1520-9210. doi: 10.1109/TMM.2018.2889563.
- [18] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7361–7369, Honolulu, HI, USA, July 2017.
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.
- [20] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.
- [21] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.
- [22] Eleni Efthimiou and Stavroula-Evita Fotinea. GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI. In *Universal Access in Human Computer Interaction. Coping with Diversity*, Lecture Notes in Computer Science, pages 657–666. Springer, Berlin, Heidelberg, July 2007. ISBN 978-3-540-73278-5 978-3-540-73279-2. doi: 10.1007/978-3-540-73279-2_73.
- [23] Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. Sign Language technologies and resources of the Dicta-Sign project. In *LREC Workshop on the Representation and Processing of Sign Languages*, pages 23–27, May 2012.
- [24] Marco Fagiani, Emanuele Principi, Stefano Squartini, and Francesco Piazza. A New Italian Sign Language Database. In Huaguang Zhang, Amir Hussain, Derong Liu, and Zhanshan Wang, editors, *Advances in Brain Inspired Cognitive Systems*, number 7366 in Lecture Notes in Computer Science, pages 164–173. Springer Berlin Heidelberg, January 2012. ISBN 978-3-642-31560-2 978-3-642-31561-9.
- [25] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.
- [26] S. S. Fels and G. E. Hinton. Glove-Talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE Transactions on Neural Networks*, 4(1):2–8, January 1993. ISSN 1045-9227. doi: 10.1109/72.182690.
- [27] Finish Association of the Deaf. Suvi Viittomat. <http://suvi.viittomat.net/>, 2015.
- [28] Jens Forster, Christian Oberdörfer, Oscar Koller, and Hermann Ney. Modality Combination Techniques for Continuous Sign Language Recognition. In *Iberian Conference on Pattern Recognition and Image Analysis*, Lecture Notes in Computer Science 7887, pages 89–99, Madeira, Portugal, June 2013. Springer.

- [29] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, pages 1911–1916, Reykjavik, Island, May 2014.
- [30] Gary J. Grimes. Digital data entry glove interface device, November 1983. US Patent.
- [31] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.
- [32] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Sign language recognition using 3d convolutional neural networks. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.
- [33] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *Proc. of the [AAAI] Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, February 2018.
- [34] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.
- [35] Avi C. Kak. Purdue RVL-SLLL ASL Database for Automatic Recognition of American Sign Language. In *Proc. IEEE Int. Conf. on Multimodal Interfaces (ICMI)*, pages 167–172, Washington, DC, USA, October 2002. IEEE Computer Society. ISBN 978-0-7695-1834-3. doi: 10.1109/ICMI.2002.1166987.
- [36] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
- [37] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding (CVIU)*, 141:108–125, December 2015. ISSN 1077-3142. doi: 10.1016/j.cviu.2015.09.013.
- [38] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *Proc. British Machine Vision Conference (BMVC)*, pages 1–12, York, UK, September 2016.
- [39] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-Sign: Re-Aligned End-To-End Sequence Modelling With Deep Recurrent CNN-HMMs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4297–4305, Honolulu, HI, USA, July 2017.
- [40] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *International Journal of Computer Vision (IJCV)*, 126(12):1311–1325, December 2018. ISSN 1573-1405. doi: 10.1007/s11263-018-1121-3.
- [41] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 0:15, 2019.
- [42] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *High Performance Computing in Science and Engineering '12*, pages 571–582. Springer, 2013.

- [43] Alina Kuznetsova, Laura Leal-Taixé, and Bodo Rosenhahn. Real-time sign language recognition using a consumer depth camera. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 83–90, 2013.
- [44] Rung-Huei Liang and Ming Ouhyoung. A real-time continuous gesture recognition system for sign language. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 558–567. IEEE, 1998.
- [45] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [46] B. Loeding, S. Sarkar, A. Parashar, and A. Karshmer. Progress in automated computer recognition of sign language. *Computers Helping People with Special Needs*, 1:624–624, 2004.
- [47] Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. Dicta-Sign—building a multilingual sign language corpus. In *LREC Workshop on the Representation and Processing of Sign Languages*, pages 117–122. Istanbul, Turkey, May 2012.
- [48] Syed Atif Mehdi and Yasir Niaz Khan. Sign language recognition using sensor gloves. In *Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on*, volume 5, pages 2204–2206. IEEE, 2002.
- [49] Ross E Mitchell, Travas A Young, BELLAMIE BACHELDA, and Michael A Karchmer. How many people use asl in the united states? why estimates need updating. *Sign Language Studies*, 6(3):306–335, 2006.
- [50] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4207–4215, 2016.
- [51] Sunita Nayak, Kester Duncan, Sudeep Sarkar, and Barbara Loeding. Finding Recurrent Patterns from Continuous Sign Language Sentences for Automated Extraction of Signs. *Journal of Machine Learning Research*, 13:2589–2615, September 2012.
- [52] Eng-Jon Ong, Oscar Koller, Nicolas Pugeault, and Richard Bowden. Sign Spotting using Hierarchical Sequential Patterns with Temporal Intervals. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1938, Columbus, OH, USA, June 2014.
- [53] Cemil Oz and Ming C Leu. American sign language word recognition with a sensory glove using artificial neural networks. *Engineering Applications of Artificial Intelligence*, 24(7):1204–1213, 2011.
- [54] V. Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos. Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–6, June 2011. doi: 10.1109/CVPRW.2011.5981681.
- [55] Anastasios Roussos, Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos. Dynamic affine-invariant shape-appearance handshape features and classification in sign language videos. *The Journal of Machine Learning Research*, 14(1):1627–1663, 2013.
- [56] Florian Schroff, Dmitry Kalenichenko, and James 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.

- [57] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.
- [58] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
- [59] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [60] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [61] Thad Starner and Alex Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion-Based Recognition*, pages 227–243. Springer, 1997.
- [62] Thad 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.
- [63] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [64] Shinichi Tamura and Shingo Kawasaki. Recognition of sign language motion images. *Pattern Recognition*, 21(4):343–353, 1988. ISSN 0031-3203. doi: 10.1016/0031-3203(88)90048-9.
- [65] Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos. Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. *Image and Vision Computing*, 32(8):533–549, 2014. 00003.
- [66] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
- [67] Dominique Uebersax, Juergen Gall, Michael Van den Bergh, and Luc Van Gool. Real-time sign language letter and word recognition from depth data. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 383–390. IEEE, 2011.
- [68] W Vicans. American sign language. *ASLU*, 2012.
- [69] Ville Viitaniemi, Tommi Jantunen, Leena Savolainen, Matti Karppa, and Jorma Laaksonen. Spot—a benchmark in spotting signs within continuous signing. In *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, pages 1892–1897, Reykjavik, Iceland, May 2014.
- [70] U. von Agris, J. Zieren, U. Canzler, B. Bauer, and K. F Kraiss. Recent developments in visual sign language recognition. *Universal Access in the Information Society*, 6(4):323–362, 2008. ISSN 1615-5289.
- [71] Ulrich von Agris, Moritz Knorr, and K.-F. Kraiss. The significance of facial features for automatic sign language recognition. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG)*, pages 1–6, Amsterdam, The Netherlands, September 2008. IEEE.
- [72] Robert Y Wang and Jovan Popović. Real-time hand-tracking with a color glove. *ACM transactions on graphics (TOG)*, 28(3):63, 2009.
- [73] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

- [74] Ronnie Wilbur and Avinash C. Kak. Purdue RVL-SLLL American Sign Language Database. Technical Report TR-06-12, School of Electrical and Computer Engineering, Purdue University, W. Lafayette, IN 47906, 2006.
- [75] 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*, pages 279–286. ACM, 2011.
- [76] Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Deutsche Arbeitsgemeinschaft für Mustererkennung Symposium*, volume 3663 of *Lecture Notes in Computer Science*, pages 401–408, Vienna, Austria, August 2005.
- [77] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, Xiaohui Xie, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *AAAI*, volume 2, page 6, 2016.
- [78] Jason E Zinza. Master asl. *Sign Media Inc*, 2006.