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

Hamid Reza Vaezi Joze
Microsoft
Redmond, WA
hava@microsoft.com

Oscar Koller
Microsoft
Munich, Germany
oscar.koller@microsoft.com

Abstract

Computer Vision has been improved significantly in the past few decades. It has enabled machine to do many human tasks. However, the real challenge is in enabling machine to carry out tasks that an average human does not have the skills for. One such challenge that we have tackled in this paper is providing accessibility for deaf individual by providing means of communication with others with the aid of computer vision. Unlike other frequent works focusing on multiple camera, depth camera, electrical glove or visual gloves, we focused on the sole use of RGB which allows everybody to communicate with a deaf individual through their personal devices. This is not a new approach but the lack of realistic large-scale data set prevented recent computer vision trends on video classification in this filed.

In this paper, we propose the first large scale American sign language (ASL) data set that covers over 200 signers, signer independent sets, challenging and unconstrained recording conditions and a large class count of 1000 signs. We evaluate baselines from action recognition techniques on the data set. We propose I3D, known from video classifications, as a powerful and suitable architecture for sign language recognition. We also propose new pre-trained model more appropriate for sign language recognition. Finally, We estimate the effect of number of classes and number of training samples on the recognition accuracy.

1. Introduction

Approximately 2 million people in the United States cannot understand normal speech, and of this number, around 500,000 use ASL to communicate [45], ASL is also used in Canada, Mexico and 20 other countries. Just like any other languages, ASL includes set of vocabulary as well as a grammar which is different from English language. With the improvement of machine learning and computer vision techniques in the past decade, many challenging problems in this domain have been solved. Some of the devices that we use daily benefit from these technical advances such as face detection, face recognition, body pose detection and others. We are intrigued by accessibility for the Deaf and believe sign language 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 [63, 71], sensor [40, 44] or colored [68] gloves. However, these requirements limit the applicability to specific settings where such resources are available. We want to support sign 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 language detection is not new but the lack of realistic large-scale data set prevent recent computer vision trends in 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 data set, establishing thorough baselines and carrying over recent computer vision trends. We make the following contributions with this work:

- 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.

- We evaluate approaches by 2D-CNN, body key-point and 3D-CNN as baselines on the data set.

- We propose I3D (known from action recognition) as a powerful and suitable architecture for sign language recognition and provide new pre-trained model for it.

- We estimate the effect of number of classes and and number of training samples on the performance.

The outline of the paper is as follows. In Section 2, we overview methods for sign language recognition as well as

current sign language data sets. Section 3 describes the technical details of our proposed ASL data set. Section 3 describes the baseline methods we applied for benchmarking as well as our proposed method. In Section 5, we experimentally show the performance of these methods on our proposed data set. Section 6 summarizes the paper with a conclusion.

2. Previous Works

2.1. Sign Language 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 [27]. In the meantime, there have been a lot of related approaches which rely on tracked hand movements based on sensor gloves for sign recognition [12, 23, 40, 44, 49]. Some works extended this by adding a camera as a new source of information [45] 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.* [60] were the first to follow vision-based sign language recognition. They built a system to recognize 14 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 [39, 63, 71] or multiple cameras [4]. Some rely on colored gloves to ease hand and finger tracking [68, 14].

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 [58, 57, 67, 25], 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 [36, 37], related works with 3D-CNNs exist [29, 6].

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.

2.2. Sign Language Data Sets

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 samples to thousands or even hundreds of 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.

Some reviews of sign language corpora exists. Though outdated, [42] provides a detailed table of sign language publications and their employed data sets until the year 2004. A less extensive, but more recent table can be found online [1]. Bellow, we have reviewed sign language data sets with explicit setups intended for reproducible pattern recognition research. The publicly available data set is presented first followed by the available private resources, all ordered by date.

- The **Purdue RVL-SLLL ASL database** [31, 70] 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 by RWTH Aachen University. The **RWTH-BOSTON-50** [72] corpus contains isolated sign language with a limited vocabulary of 50 signs. The **RWTH-BOSTON-104** constitutes of continuous sign language and covers a vocabulary of 104 signs performed by 3 signers. The **RWTH-BOSTON-400** corpus contains a vocabulary of 483 signs and also constitutes of continuous signing by 5 signers. It has a total of 7768 running gloss annotations.

- The **SIGNUM corpus** [66] provides two evaluation sets: first a multisigner 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. The corpus is available for purchase. With few variables, it is very controlled and [34] presented a method with less than 5% error rates for it.

- In the scope of the DictaSign project, multi-lingual sign language resources have been created [3, 20, 43]. However, the produced corpora are not well curated and made available for reproducible research. A lot of different versions exist which are based on the same video recordings and annotations, but represent different subsets. The **Greek Sign Language (GSL) Lemmas Corpus** [19] constitutes of 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 [50, 61], with 984 signs [15] and with 981 signs [48]. Additionally, a continuous 100 sign version of the data set has been used in [51]. 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 [11] 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.

- The **Finish S-pot** sign spotting task [65] is based on the controlled recordings from the Finish sign language lexicon [24]. 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** [26, 33] and **RWTH-PHOENIX-Weather 2014 T** [5] 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 data for training with about 800k frames. The data set is very challenging from a segmentation and language point of view however, it only features 9 signers and limited computer vision challenges.

There are several groups which experimented with their own data collection resulting in corpora with quite limited size in terms of total number of annotations and vocabulary. Often these resources are not made publicly available.

One such example is the non-public **UWB-07-SLR-P corpus of Czech sign language** [8] which contains recordings of 4 signers with a 378 sign vocabulary and mostly 5 repetitions in 3 different camera views. It has a total length of 11.1 hours. Unfortunately, no recognition results could be found in the literature.

Another example is a data set by Barabara Loedings Group [42, 47]. In their publications they describe different data sets all containing at most 155 sequences with a limited vocabulary. Other small scale corpora exist, which are often not publicly available [7, 21].

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 re-

cent 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 signs.

3. Proposed ASL Data Set

Since there is no ASL public data set suitable for large scale video classification training, 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. Labeling video requires skilled ASL natives, this is difficult to crowdsource with platforms like Amazon Mechanical Turk or similar.

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:

- For longer videos, we used Optical Character Recognition (OCR) to find printed labels and their time of occurrence.

- Longer videos may contain video captions that provide the sign descriptor and the temporal segmentation.

- 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 splited 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
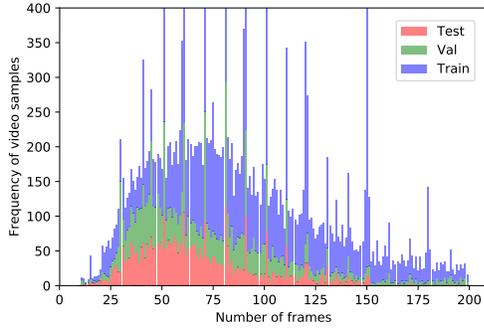
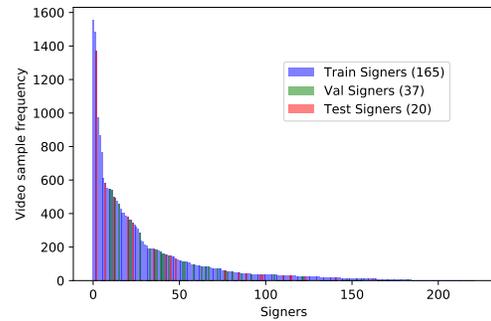Figure 1. Histogram of frame numbers for ASL1000 video samples.

Figure 2. Showing the number of video samples for each of the 222 signers and the train/test/validation split of proposed data set ASL1000. The signers are ordered by the number of their video samples.

mistakes. Since many of the ASL vocabulary publicly accessible videos are belong to teachers performing a lesson vocabulary or students doing their homework, all top hundred words belong to ASL tutorial books [74, 64] vocabulary units.

3.1. Manual Touch-up

Although many of the samples 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. There are few samples outside of the defined criteria which also have been reviewed by our annotators. In this way, around 25% of the data set was manually reviewed. After the touch-up, almost all the samples have less than 200 and more than 15 frames. 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. Typically, sign languages have no standardized written form. Therefore, a written word will usually just refer to the meaning of a sign, not to the way it is executed. This is fundamentally differ to most writing schemes of spoken languages which also applies to ASL. 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" [64]. 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[74]. This mapping of sign classes will be released as part of the MS-ASL data set.

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 [52] 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 randomly distributed signers to 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
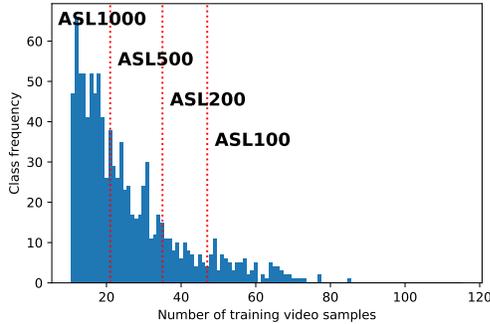
Figure 3. Histogram of number of samples per class for sums of Train, Test and Validation set for ASL1000 data set.

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 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[2]. Table 1 shows the characteristics of each of this sets. In *ASL100*, there are at least 45 samples for each class while in *ASL1000* there are at least 11 samples for each class.

Figure 3 illustrates the histogram of the number of video samples per class for *ASL1000*. The bars at the right of $x = 47$ form *ASL100*, the bars at the right of $x = 35$ form *ASL200* and the bars at the right of $x = 21$ form *ASL500*.

3.5. Data Set Challenges

We used automatic methods to access the data set and limit the manual annotation work. While this was crucial to speed up the corpus generation and reach data set sizes suitable for deep learning, it increased noise in the transcriptions. As described in Section 3.1, we controlled this factor by manually verifying the labels of about one quarter of the data set. And the fact that we used most frequent words as well as manual touch-up for long samples decreased the chance of wrong labels but we know that the data set is not clean. There are challenges in this data set which make it more challenging compared to other computer vision data sets.

---

[2] Instructions and download links:

- One sample video may include repetitive act of a distinct signs.

- One word can sign differently in different dialects based on geographical regions. As an example, there are 5 common signs for the word *Computer*.

- It includes large number of signers and is a signer independent data set.

- The are large visual variabilities in the videos such as background, lighting, clothing and camera view point.

3.6. 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 compare to accuracy because of the unbalance test set inherited from the unbalance nature of the data set (Figure 3). 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. As we'll discuss later in subsection 5.2, ASL, just like any other language can have ambiguity which can be resolved in context. That is the main reason we pick top-five accuracy. On the other word, In order to have a good sign recognition technique, it could have good accuracy in first guess or at least on next few guesses.

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 [17, 22] 2) Extracting subject's body joints in the form of skeleton and using skeleton data for recognition [18, 73] 3) Using 3D convolution [10, 62, 46]. 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 person body bounding box as input image, so We extract person bounding box by SSD network [41] and release it for each video sample as part of MS-ASL data set. We employ the following data spacial and temporal augmentations during the training stage for all methods. For special augmentations, body bounding boxes are randomly scaled or translated by 10%, shaped into a square and re-sized to fixed 224×224 pixels. We picked 64 as our temporal window which is average number of frames for data set's sample videos. In addition,

Data set	Class	Subjects	Number of Videos				Duration	Videos per class	
			Train	Validation	Test	Total	[hours:min]	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 1. Showing statistics of the 4 proposed subsets of the MS-ASL signer independent sign language recognition data set.

the resulted video is randomly but consistently 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 fist or last frame. We use 40 epochs for all training process. 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 [54, 22] or just train stage [1] boosts the performance of prediction. Herein, we describe the methods used for determining baselines.

4.1. 2D-CNN

The high performance of 2D convolutional networks on image classication makes them the first candidate for video processing. This is achieved by extracting features from each frame of video independently. The first approach was to combine these features by simply pooling the predication, 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 [17], we picked LSTM [28] as our recurrent layer which records the temporal ordering and long range dependencies by encoding the states. We used VGG16 [55] 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 [35] as the state-of-the-art on PHOENIX2014 data set [26, 33]. This method use GoogleNets [59] as 2D-CNN with 2 bi-directional LSTM layers and 3 state HMM. We report it as Re-Sign in experimental result.

4.2. Body Key-Points

With the introduction of robust body key-points (so-called skeleton) detection [69], some studies try to solve human action recognition by body joints only [18, 73] or use body joints along with the image stream [13]. Since most body key-point techniques did not cover hand details, it was not rational to use it for sign language recognition task as
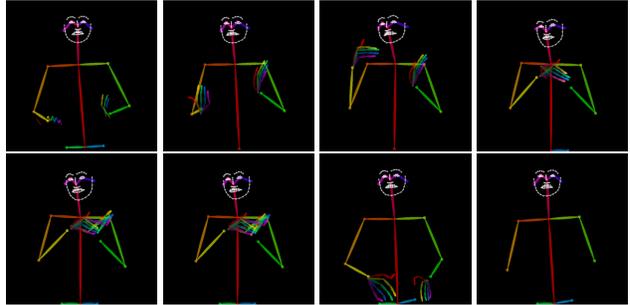

Figure 4. Extracted 137 body key-points for a video sample from MS-ASL by [9, 53] with label *Again*.

it relies heavily on the movement of fingers. But a recent work has covered hand and face key-points along with classical skeleton [53]. 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, 53]. 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 4 illustrates the extracted 137 body key-points for a video sample from proposed data set. Although this technique works well in normal cases, the hand key-points are not robust specially if the connection from body to the hand is not visible or hands are very close to camera.

We implemented hierarchical co-occurrence network (HCN) [73] 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. We train this network by Adam optimizer.

4.3. 3D-CNN

Recently, 3D convolutional networks have shown promising performance for video classification and action recognition including C3D network [62] and I3D network [10]. We applied C3D [62] released code from author as well as our own implemented version to our proposed data sets with and without pre-trained model, trained

Method	ASL100	ASL200	ASL500	ASL1000
VGG+LSTM	13.33%	7.56%	1.47%	-
HCN [73]	46.08%	35.85%	21.45%	15.49%
Re-Sign [35]	78.12%	-	-	15.01%
I3D [10]	81.76%	81.97%	72.50%	57.69%

Table 2. The average per class accuracy for method mentioned in section 4 on proposed ASL detests.

Method	ASL100	ASL200	ASL500	ASL1000
VGG+LSTM	33.42%	21.21%	5.86%	-
HCN [73]	73.98%	60.29%	43.83%	32.50%
I3D [10]	95.16%	93.79%	89.80%	81.08%

Table 3. The average per class top-five accuracy for method mentioned in section 4 on proposed ASL detests.

on Sport-1M [32]. 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 [30], which contains several 3D convolutional layers followed with 3D max-pooling layers and inflated Inception-V1 submodules. We started with pre-trained network trained on Imagenet [16] 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 3.6, we report two evaluation metrics: average per class accuracy and average per class top-five accuracy. The results are reported in Table 2 and Table 2 respectively. Although we did not optimize training parameters or implement complex technical details for each of these methods, we can still consider these results as a baseline for 2D-CNN, 3D-CNN 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 pass the recurrent information well. In video classification data sets such as UCF101 [56] or HMDB51 [38], the image itself carries context information regarding the classification while in MS-ASL there is minimum context information in a single image. Body key-point based approach (HCN) is doing relatively better compared to 2D-CNN but there is a 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 (I3D) did a pretty good job in this challenging, uncontrolled data set and we propose it as powerful network for sign language recognition.

5.1. 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 compute the average per class accuracy of the I3D model trained on *ASL200* on *ASL100* test set at 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 pre-trained model is not suitable for sign language recognition. Because its weights have been trained on irrelevant video classification task with classes such as kayaking, skydiving and car driving. We proposed the model trained on MS-ASL as a I3D pre-trained model for sign language recognition tasks.

5.2. 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, 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.
- *Not* labeled as *Nice* (.33)
- *Today* labeled as *Now* (.33): There are 2 version for *Today* one of them is signing *Now* twice.
- *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
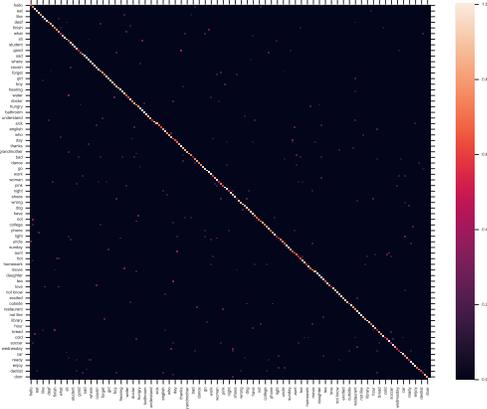
Figure 5. The confusion matrix obtained by comparison of the grand-truth labels and the predicted labels from models trained by I3D on *ALS200* data set.

Figure 6. The accuracy of 4 trained models based on frequency of training video samples.

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 came 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.

5.3. 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 6. 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 5.2.

5.4. 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 possi-
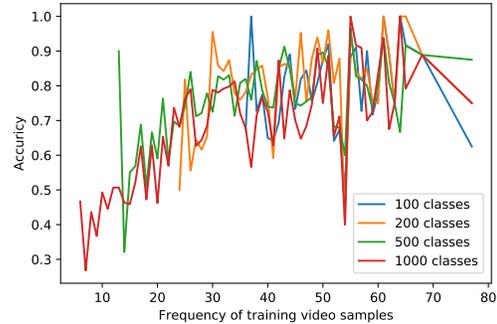
ble pairs and reported the average per class accuracy on Table 4. In this table we show subsets of the MS-ASL data set on the horizontal axis and the tested subsets on the vertical axis. Increasing the number of classes decreased the accuracy of either the train or the test phase. Doubling the size of 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 but it is inevitable.

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

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 the I3D can successfully be used for sign language recognition as it has the suitable architecture for the task. We also estimated the effect of number of classes and number of training samples 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] M. Abavisani, H. Vaezi Joze, and V. Patel. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. *arXiv*, 2018. 6

[2] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. 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. 2

[3] A. Braffort, L. Bolot, E. Chételat-Pelé, A. Choisier, M. Delorme, M. Filhol, J. Segouat, C. Verrecchia, F. Badin, and N. Devos. Sign Language Corpora for Analysis, Processing and Evaluation. In *International Conference on Language Resources and Evaluation*, pages 453–456, Valletta, Malta, May 2010. ELRA. 2

[4] H. Brashear, T. Starner, P. Lukowicz, and H. Junker. Using multiple sensors for mobile sign language recognition. Georgia Institute of Technology, 2003. 1, 2

[5] C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural Sign Language Translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, June 2018. 3

[6] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. Using Convolutional 3D Neural Networks for User-Independent Continuous Gesture Recognition. In *IEEE International Conference of Pattern Recognition, ChaLearn Workshop*, Cancun, Mexico, Dec. 2016. 2

[7] N. C. Camgoz, A. A. Kindiroglu, S. Karabuklu, M. Kelepir, A. S. Ozsoy, and L. Akarun. BosphorusSign: A Turkish Sign Language Recognition Corpus in Health and Finance Domains. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016. 3

[8] P. Campr, M. Hrúz, and J. Trojanová. Collection and preprocessing of czech sign language corpus for sign language recognition. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008. 3

[9] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multiperson 2d pose estimation using part affinity fields. In *CVPR*, 2017. 6

[10] J. Carreira and A. 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. 5, 6, 7

[11] X. Chai, H. Wang, and X. 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. 3

[12] C. Charayaphan and A. E. Marble. Image processing system for interpreting motion in American Sign Language. *Journal of Biomedical Engineering*, 14(5):419–425, Sept. 1992. 2

[13] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid. Potion: Pose motion representation for action recognition. In *CVPR 2018*, 2018. 6

[14] H. Cooper and R. Bowden. Sign Language Recognition using Linguistically Derived Sub-Units... 2010. 2

[15] H. Cooper, N. Pugeault, and R. Bowden. Reading the signs: A video based sign dictionary. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference On*, pages 914–919. IEEE, Nov. 2011. 3

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. 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. 7

[17] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 5, 6

[18] Y. Du, W. Wang, and L. 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. 5, 6

[19] E. Efthimiou and S.-E. Fotinea. GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI. In *Universal Acess in Human Computer Interaction. Coping with Diversity*, Lecture Notes in Computer Science, pages 657–666. Springer, Berlin, Heidelberg, July 2007. 2

[20] E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos, and F. Lefebvre-Albaret. Sign Language technologies and resources of the Dicta-Sign project. In *Proc. of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. LREC*, pages 23–27, 2012. 2

[21] M. Fagiani, E. Principi, S. Squartini, and F. Piazza. A New Italian Sign Language Database. In H. Zhang, A. Hussain, D. Liu, and Z. Wang, editors, *Advances in Brain Inspired Cognitive Systems*, number 7366 in Lecture Notes in Computer Science, pages 164–173. Springer Berlin Heidelberg, Jan. 2012. 3

[22] C. Feichtenhofer, A. Pinz, and A. 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. 5, 6

[23] 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, Jan. 1993. 2

[24] Finish Association of the Deaf. Suvi Viittomat. http://suvi.viittomat.net/, 2015. 3

[25] J. Forster, C. Oberdörfer, O. Koller, and H. 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. 2

[26] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *International Conference on Language Resources and Evaluation*, pages 1911–1916, Reykjavik, Island, May 2014. 3, 6

[27] G. J. Grimes. Digital data entry glove interface device, Nov. 1983. US Patent. 2

[28] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[29] J. Huang, W. Zhou, H. Li, and W. Li. Sign language recognition using 3d convolutional neural networks. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015. 2

[30] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 7

[31] A. C. Kak. Purdue RVL-SLLL ASL Database for Automatic Recognition of American Sign Language. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, pages 167–172, Washington, DC, USA, 2002. IEEE Computer Society. 2

[32] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. 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. 6

[33] O. Koller, J. Forster, and H. Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, Dec. 2015. 3, 6

[34] O. Koller, S. Zargaran, and H. Ney. Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017. 2

[35] O. Koller, S. Zargaran, and H. Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6, 7

[36] O. Koller, S. Zargaran, H. Ney, and R. Bowden. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *British Machine Vision Conference*, York, UK, Sept. 2016. 2

[37] O. Koller, S. Zargaran, H. Ney, and R. Bowden. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *International Journal of Computer Vision*, 126(12):1311–1325, Dec. 2018. 2

[38] H. Kuehne, H. Jhuang, R. Stiefelhagen, and T. Serre. Hmdb51: A large video database for human motion recognition. In *High Performance Computing in Science and Engineering 12*, pages 571–582. Springer, 2013. 7

[39] A. Kuznetsova, L. Leal-Taixé, and B. 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. 2

[40] R.-H. Liang and M. 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. 1, 2

[41] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 5

[42] 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. 2, 3

[43] S. Matthes, T. Hanke, A. Regen, J. Storz, S. Worseck, E. Efthimiou, A.-L. Dimou, A. Braffort, J. Glauert, and E. Safar. Dicta-Sign–building a multilingual sign language corpus. In *Proc. of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon (LREC), European Language Resources Association*. Istanbul, 2012. 2

[44] S. A. Mehdi and Y. N. 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. 1, 2

[45] R. E. Mitchell, T. A. Young, B. BACHELDA, and M. A. Karchmer. How many people use asl in the united states? why estimates need updating. *Sign Language Studies*, 6(3):306–335, 2006. 1, 2

[46] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. 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. 5

[47] S. Nayak, K. Duncan, S. Sarkar, and B. Loeding. Finding Recurrent Patterns from Continuous Sign Language Sentences for Automated Extraction of Signs. *Journal of Machine Learning Research*, 13:2589–2615, Sept. 2012. 3

[48] E.-J. Ong, O. Koller, N. Pugeault, and R. Bowden. Sign Spotting using Hierarchical Sequential Patterns with Temporal Intervals. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1938, Columbus, OH, USA, June 2014. 3

[49] C. Oz and M. 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. 2

[50] V. Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos. Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–6, June 2011. 3

[51] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. 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. 3

[52] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4

[53] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 6

[54] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances*

in neural information processing systems, pages 568–576, 2014. 6

[55] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[56] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7

[57] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion-Based Recognition*, pages 227–243. Springer, 1997. 2

[58] T. E. Starner. Visual recognition of american sign language using hidden markov models. Technical report, Massachusetts Inst Of Tech Cambridge Dept Of Brain And Cognitive Sciences, 1995. 2

[59] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[60] S. Tamura and S. Kawasaki. Recognition of sign language motion images. *Pattern Recognition*, 21(4):343–353, 1988. 2

[61] S. Theodorakis, V. Pitsikalis, and P. Maragos. Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. *Image and Vision Computing*, 32(8):533–549, 2014. 00003. 3

[62] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 5, 6

[63] D. Uebersax, J. Gall, M. Van den Bergh, and L. 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. 1, 2

[64] W. Vicars. American sign language. *ASLU*, 2012. 4

[65] V. Viitaniemi, T. Jantunen, L. Savolainen, M. Karppa, and J. Laaksonen. S-pot–a benchmark in spotting signs within continuous signing. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, May 2014. 3

[66] U. von Agris, M. Knorr, and K.-F. Kraiss. The significance of facial features for automatic sign language recognition. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference On*, pages 1–6. IEEE, 2008. 2

[67] 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. 2

[68] R. Y. Wang and J. Popović. Real-time hand-tracking with a color glove. *ACM transactions on graphics (TOG)*, 28(3):63, 2009. 1, 2

[69] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 6

[70] R. Wilbur and A. 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. 2

[71] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. American sign language recognition with the kinect. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 279–286. ACM, 2011. 1, 2

[72] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Deutsche Arbeitsgemeinschaft für Mustererkennung Symposium*, volume 3663 of *Lecture Notes in Computer Science*, pages 401–408, Vienna, Austria, Aug. 2005. 2

[73] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *AAAI*, volume 2, page 6, 2016. 5, 6, 7

[74] J. E. Zinza. Master asl. *Sign Media Inc*, 2006. 4