Sign Language Recognition using Deep Learning

Vyom Pathak[1,2], **Dipali Patidar**[1,2], and **Shefali Mishra**[1,2]

[1]Department of Computer & Information Science & Engineering, University of Florida
[2]*{v.pathak, dipali.patidar, shefali.mishra}@ufl.edu*

Abstract

Vision-based gesture recognition is crucial in helping people with impaired speech and hearing. It is a challenging area of sign language translation that requires processing on a rather limited yet complex dataset. For this study, Word-Level American Sign Language (WLASL) video dataset was used that contains over 2000 glosses. It is one of the largest public ASL datasets to facilitate word-level sign recognition research. We have done a comparative study of various sign language recognition (SLR) systems using video/image-based Deep Learning models to understand their capacity to hold information. Isolated sign words recognition system has been implemented using pre-trained I3D weights by Carreira & others. Our results were similar to that in the reference study showing appearance-based model achieve up to 63.9% at top-10 accuracy on 2000 glosses, demonstrating the validity and challenges of our datasets.

1 Introduction

Based on a survey done by the CDC in 2019, about 1.7 per 1000 babies that were born that year were identified with a permanent hearing loss [15]. Moreover, in the United States, 2 to 3 out of every 1000 children in the United States are born with a measurable level of hearing loss in one or both ears [48]. One of the major problems faced by the deaf community is the communication gap with the more general hearing community. Most communication technologies have been developed to only support spoken or written forms of any language (which excludes the use of sign languages). With the arrival of modern communication technologies becoming an integral part of our life [2], deaf people have faced issues using these technologies.

Sign language as a structural form of communication system has been encouraged to help the speech-impaired and the deaf community in daily interactions [30]. The Fig. 1 shows different ASL sign gestures. Sign language consists of the usage of a different part of the body, like fingers, hands, arms, head, body, and facial expressions [8]. It accounts for five main parameters namely, hand shape, palm orientation, location, movement, and expression signals [30]. Fig. 2 shows some of these parameters like handshape, movement, location, and palm orientation each with one of the parameters which varies. For an accurate sign-word, all of these five parameters must be performed/interpreted correctly. A survey by the World Federation of the Deaf has reported that there are over 300 sign languages around the world that about 70 million deaf people use [53].

Because of the complex and intricate hand gestures in quick motions, body movements, and facial expressions, as well as the sheer amount of people using this language for day-to-day communication, Sign Language Recognition (SLR) is a very complex as well as important problem to tackle. Here, we look at automatically translating sign languages using vision technologies to text. With the advent of Deep Learning [31], GPUs i.e. compute power to boost these algorithms, along with the development of strong frameworks like TensorFlow [1], PyTorch [39], Keras [9], and MXNET [7]; we pose this problem using video/image-based Deep Learning algorithms.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

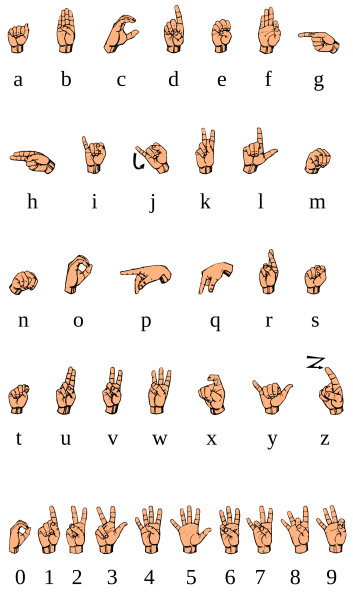


Figure 1: American Sign Language Hand Gestures for alphabet and numbers.

There have been several attempts at solving this problem using different types of features (hand pose, facial expressions, and body posture) [43]. In terms of the type of data, the algorithms are divided into RGB and Depth data as shown in Fig. 3 [43]. There have been strides in developing SLR systems in terms of recognition modality i.e. in both isolated (word-level sign language recognition) as well as dynamic (sentence-level sign language recognition) domains where the dynamic ones are the more complex because of their continuous nature [23, 43].

1.1 Feature Fusion

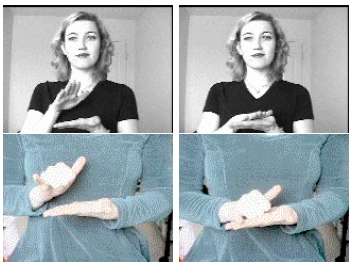
With the arrival of the precise depth sensor, the use of hand pose characteristics has become increasingly explored in recent years [5, 13, 14, 51]. Only hand characteristics are utilized to recognize sign language in this category. Following the identification of hands in input data, hand features are extracted using various deep learning architectures such as CNN, RBM, RNN, GAN, and so on [8, 13, 42, 51, 57]. While CNN is more than well suited for images, its not able to capture temporal information from video sequences. As a result, CNN is paired with another deep learning model, such as RNN, LSTM, or GRU, to make use of these models' capabilities in sequence feature extraction from visual input [43].

Face pose includes prosodic [intonation of the words] features as well as grammatical queues, thus use of these features along with the hand pose features proved to be useful in increasing the performance of the SLR models. However, very few models have been proposed using this method because it is challenging to track the human faces from video lie side-to-side movements, head-tilts, face being self-blocked due to signers' hands and hair ...etc [43]. The leading technique in this area was developed by Koller et. al. [25] which only looked at a part of the face to decrease the complexity of the problem.

Going even further, there are several models which leverage the features of the hands, face, as well as other parts of the human body. SLR models can be improved by using these fused characteristics to be more resistant to occlusions, extreme deformations, and appearance fluctuations [22, 37, 52]. This defers from the former approaches in the sense that if the face and the hand features are occluded, the body-pose feature can help learn the particular utterance.

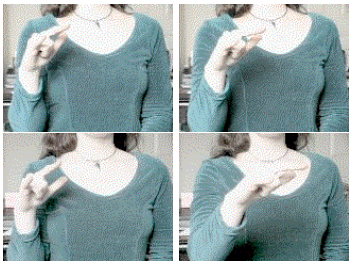
1.2 Visual Modality

RGB and Depth input data types are the two most common types of input data used in SLR systems. Depth-map inputs provide correct information about the distance between the corresponding object

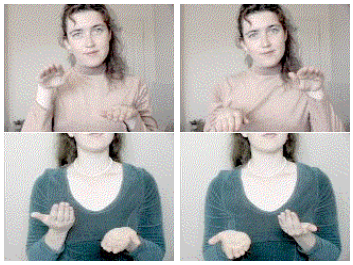


(a) ASL words school (upper image) and impossible (lower image) both have the same parameters of location, movement and palm orientation but they have different handshapes.

(b) ASL words apple (upper image) and onion (lower image) have the same handshape, movement, and palm orientation, but they have a different location which results in a different meaning.



(c) ASL words airplane (upper image) and fly (lower image) have the same parameters except for the movement. The former one has the repeated movement and the latter one has one continous movement.



(d) ASL words balance (upper image) and maybe (lower image) have the same parameters: handshape, location and movement, but the palm orientations of these signed words are different.

Figure 2: Different examples of variations in the main parameters accounting for Sign Language.

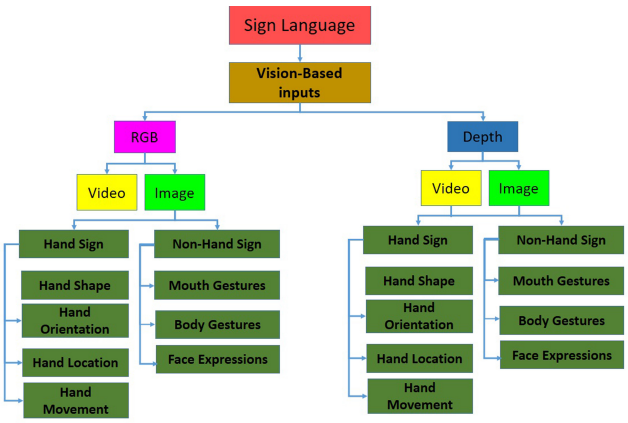


Figure 3: Vision-based sign language models classification [43]

and the image plane while RGB input data provides the high-resolution contents [43]. Other modalities include thermal modality, infrared (IR) thermal sensors, and the use of the whole skeleton as an encoded form of the joint sequences [43].

The camera is the most common device for capturing the input data modalities among several different devices, where different cameras support different qualities and formats of the data. Microsoft Kinect allows us to capture high-quality RGB as well as depth video streams simultaneously, making it one of the most widely used device [55]. Another device that tracks and detects hands, fingers, and finger-like objects are the Leap Motion Controller System (LMC) [41]. Using these types of sensors and cameras which capture the depth data, which employs 3D information decreasing the ambiguity of 2D information captured from traditional devices [34, 36].

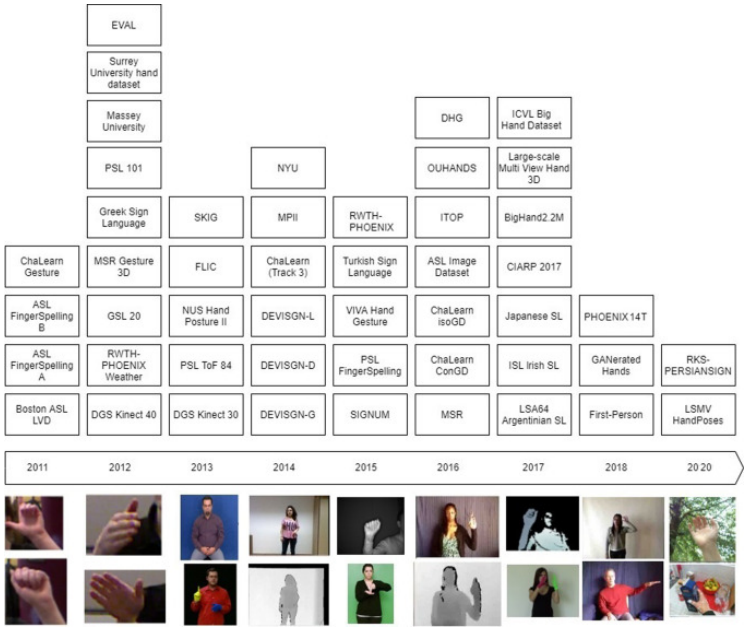


Figure 4: Evolution of Sign Language Datasets [43].

1.3 Static or Dynamic Data

Existing word-level SLR models have been developed on small-scale/private datasets with less than about 100 words. These methods include using hand-crafted features such as histogram of optical flow [35], and HOG-based features [4, 11, 40]. Hidden Markov Model [17, 45] has been used to model the temporal relationship from the video. 3D-Convolution for capturing spatial-temporal features instead of using separate information retrieval models [18, 54] has also made great breakthroughs. Furthermore, after the development of the WLASL dataset consisting of 2000 sign-poses, there have been developments in using the pose-based TGCN method for solving the problem at hand [32]. With the advent of new technologies such as semi-supervision, 3D-convolution-based algorithms have made strides in low-resource settings [33].

For the sentence-level SLR models, the largest known benchmark dataset is the RWTH-PHOENIX-Weather 2014 consisting of 1080 German language sign-poses [16]. For the American Sign Language, one of the benchmark datasets is the MS-ASL dataset consisting of 1000 sign-poses. Koller developed several methods for continuous SLR systems including iterative training using expected Maximization, using 2D-Convolution neural networks, hybrid 2D CNN with HMM models [24, 26–28]. To map the long-temporal dependencies of the video, Bi-GRU, LSTM, and Bi-LSTM were used with a Connectionist Temporal Classification loss function for sequence alignment [10, 29]. With the introduction of the Attention mechanism, for extracting important information from the embedded representation of the video; Transformer based architectures were developed [19, 56, 58].

The evolution of the sign and gesture dataset can be found in Fig. 4. Most of these datasets aim towards sign classification, rather than detection/spotting, having different qualities, constraints, complexities, and environments under which they are captured. Albeit SLR systems can use different languages as input data, American Sign Language (ASL) is one of the popular ones being used amongst other languages such as Indian, German, Dutch, Greek, Polish, Turkish, and Chinese [43].

We aim at comparing SLR systems methods developed for American Sign Language (ASL). We try to compare continuous (sentence-level]) as well as isolated (word-level) SLR models. For this task, we look at the following methods and implement the I3D backbone-based method [32] and compare its performance with the original paper for the WSASL [32] datasets.

1. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language [21]

2. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison [32]
3. Transferring Cross-domain Knowledge for Video Sign Language Recognition [33]
4. ASL Recognition with Metric-Learning based Lightweight Network [20]
5. Continuous Sign Language Recognition through a Context-Aware Generative Adversarial Network [38]

The paper is structured as follows, in Section 2 we describe each method and their respective dataset which we considered for the problem at hand. In Section 3 we show the results of the method that we implemented as well as discuss any nuances we gather from the same. Finally, we conclude our work in Section 5.

2 Methods & Datasets

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

MS-ASL is the first large-scale American Sign Language (ASL) data proposed by Microsoft [21]. This dataset consists of 25,000 annotated videos, over 200 signers, and signer independent sets. The dataset contains a large class count of 1000 signs recorded in challenging and unconstrained conditions. The dataset is divided into 4 subsets including 100, 200, 500, and 1000 most frequent words subsets called ASL100, ASL200, ASL500, and ASL1000. The paper evaluates the existing three approaches 2D-CNN-LSTM [52], body key-point [59], CNN-LSTM-HMM [44] and 3D-CNN [46] as baselines. 3D-CNN baseline achieved good results in this challenging, uncontrolled data compared to the other two baseline models and the authors proposed it as a powerful network for sign language recognition. The experimental result presented in the paper suggests that this data set is very difficult for 2D-CNN or at least LSTM could not propagate the recurrent information well. Body key-point-based approach (HCN) [59] is doing relatively better compared to 2D-CNN mode.

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

WLASL [32] is the largest video dataset for Word-Level American Sign Language (ASL) recognition. This dataset was developed because of the lack of public and large sign language datasets. The dataset consists of 2,000 words, and 21,083 videos with 119 different signers, being the largest word-level ASL video dataset. Li et. al. [32] also compares the performance of different baseline systems using two approaches namely, Visual appearance-based, and 2D Human poses-based. For the testing phase, 4 different subsets were taken from the WLASL dataset. The video appearance-based system uses the whole-body video to predict the word. For this method, they compare two different types of baseline systems, which are the 2D CNN+RNN system (VGG-GRU) [32], and second is the 3D CNN system (I3D backbone) [6]. Both of these methods perform on par concerning the model's capacity to hold the information. The proposed model architectures are shown in Fig. 5.

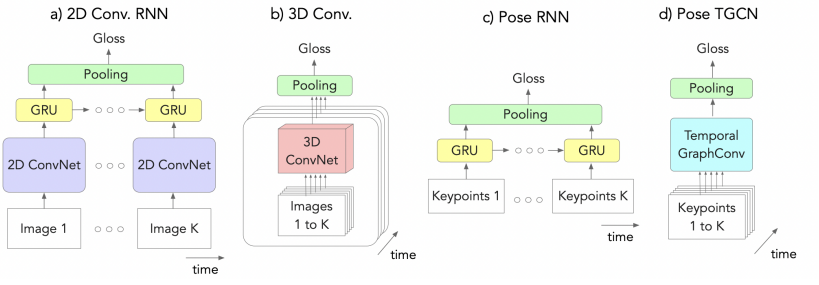


Figure 5: Baseline Architecture for the WLASL Dataset Study [32].

The 2D-human pose-based system extracts the pose of the whole body frame and then predicts the word accordingly. For the pose-based system, we first look at the baseline Pose-RNN model which

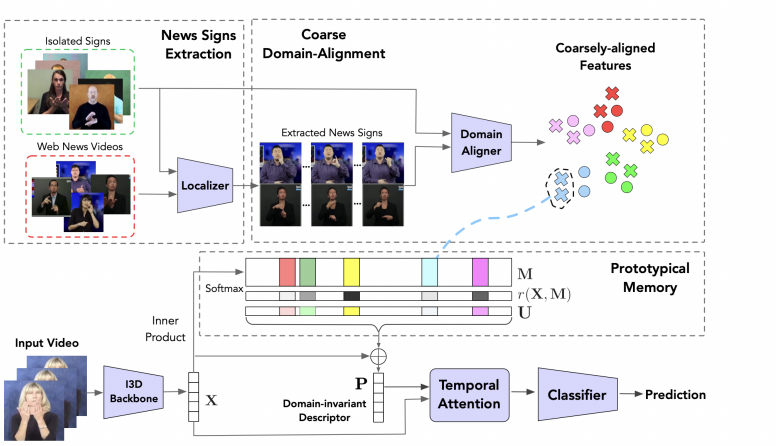


Figure 6: Overview approach of knowledge transfer of news signs to WSLR model [33].

has 55 body and hand 2D key points from the WLASL dataset which uses 13 upper-body joints and 21 joints from both right and left hands. These features are then concatenated with each other and are fed to stacked GRU layers. Cross-Entropy Loss was used for training and accuracy was used as a performance metric for the testing phase.

Secondly, a novel pose-based temporal graph convolution network was proposed (Pose-GCN). For this method, first, from N sequence frames, K dimensional key points are taken as input. For these key points, along with the motion of 2D joint angles, they incorporate temporal motion information as a representation of the trajectories of body key points. Then these features are passed onto the residual graph convolution block, stacking multiple GCN layers with a residual connection. Along the temporal dimension, average pooling results are taken as feature representations of pose trajectories. After that, a softmax layer following an average pooling layer is used for classification [32].

Finally, for the I3D network, they first consider the pre-trained model (trained on ImageNet and finetuned on Kinetics-400 dataset), which is further fine-tuned to model the temporal and spatial information of sign languages such as focusing on the hand shapes and orientations as well as arm movements. Here each image is passed onto the I3D backbone architecture followed by a pooling layer that predicts the given image. As the number of classes varies in the WLASL dataset, only the last classification layer concerning the number of classes [32].

The baseline RNN model (Pose-GRU) [32] with the proposed novel Graph Convolution Neural Network-based algorithm (Pose-GCN) [32], which out-performs the baseline in pose-based systems in terms of speed and accuracy. Amongst all these methods, the I3D model achieves the best recognition performance accuracy. However, given the size of the I3D model is larger than the Pose-GCN model, the Pose-GCN model performs comparably well.

2.3 Transferring Cross-domain Knowledge for Video Sign Language Recognition

Building upon the WLASL dataset, Li et. al. [33] propose an algorithm to solve the SLR problem in the low-resource settings. In the WLASL dataset, there is a sub-set of the dataset where only 10 examples per class are present, which entails the low-resource problem. The authors propose a transfer learning/semi-supervised learning-based algorithm for LSR [33]. The main aim of the work is to transfer knowledge between the News Data domain and the isolated data domain (WSLR data). The proposed transfer learning architecture is described in Fig. 6.

First, a localizer is trained for both the news as well as the isolated word signs dataset. This step makes sure that the features from the news signs are extracted in a manner that correlates with the WSLR dataset. Then, these features are used jointly with the isolated word dataset to jointly train a classifier using sign samples from both domains making a coarse domain alignment. For both these steps, the I3D backbone network [6] is used whereas for the feature extraction step, the classification head is replaced with a polled feature map from the last inception sub-module. The knowledge of news signs can be further exploited while classifying isolated signs by adopting an external memory.

This entails encoding knowledge of news signs into a prototypical memory which is an array of prototype features from the embedding of the coarse-align model [33].

After both the domains are in-sync with each other by coarse alignment, we focus on learning domain-invariant descriptors using the prototypical memory buffer. Now to do this, the features from both the domains are correlated with each other by two different projection matrices into one common projection space. After that, this common embedding space is normalized via dot product finally acquiring the domain-invariant descriptors [33].

Now, to capture the salient temporal information from the isolated sign representations, a temporal attention mechanism is adopted which does this job by calculating the similarity between the domain invariant descriptors and the isolated sign representations. Then, the output from the attention mechanism is fed into a final classification layer which uses the binary cross-entropy layer for the given number of classes to predict the isolated sign language. This procedure makes the model robust by explicitly minimizing the influence of irrelevant gestures as well as making it concentrate on features from the salient temporal regions. The results are compared with the similar dataset distribution as the original WSLR dataset [32], where the proposed approach outperforms all the compared models from that paper.

2.4 ASL Recognition with Metric-Learning based Lightweight Network

On the flip-side of training, a model which gives an on-par performance with large models as well as is lightweight, the authors develop such an ASL gesture recognition model which is trained under the metric learning framework allowing the ASL signs to be recognized in a live stream of video [20]. For training as well as testing the system they use the MS-ASL dataset [21]. Despite the dataset being isolated ASL, the paper aims to build a model for continuous stream sign language recognition. As a backbone, they use a modified compute efficient 3D convolution model (MobileNet-V3). To further improve the robustness of the model, they introduce a two-way Spatio-temporal attention system using residual attention mechanism [12, 49].

The method proposed here processes the continuous input stream with a fixed-size window of 16 input frames with a constant frame rate of 15. The extracted frames are cropped according to the maximum bounding box containing the face and raised hands of a person taken over all frames in a sequence. The processed image sequence is resized to 224×224 generating the input dataset size of $16 \times 224 \times 224$. The author selected MobileNet-V3 as a base 2D model and extended it to spatial and temporal separable 3D convolutions. Two residual Spatio-temporal attentions layers were also added. The model architecture is described in fig 7. Due to the low-resource setting, the metric learning approach is adopted where they use the AM-Softmax loss with auxiliary self-supervision loss [50]. They also provide an ablation study on different parameters for handling the trade-off between the accuracy and the size of the model.

2.5 Continuous Sign Language Recognition through a Context-Aware Generative Adversarial Network

The author here proposes a novel Sign Language Recognition Generative Adversarial Network (SLRGAN) [38]. The network architecture consists of the generative model that learns the gloss of the signs by extracting the spatial as well as the temporal features from the video sequences that captures the content of the video. The discriminator which is a text modeling network evaluates the quality of the generator's predictions by modeling the textual information at the sentence and gloss levels to evaluate whether sentences or gloss information comes from the generator or original input data. The training is done in an adversarial manner, where the generator and discriminator play a min-max game to fool each other.

The proposed generator model by the author consists of four main parts which are presented below in Fig. 8. An input frame sequence is passed to a 2D-CNN followed by temporal convolution and pooling layers to map input frame sequences into Spatio-temporal features. The next layer is the BLSTM layer which takes the extracted feature as input to learn long-term dependencies over all time steps. The last layer is the softmax classifier predicting the output glosses. In the proposed CSLR task, input video is processed such that previous and succeeding signs of the video sequence have relationships with the current sign [38].

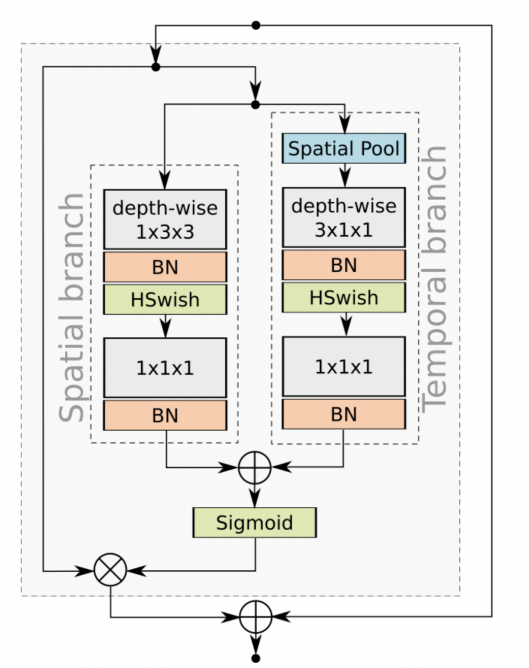


Figure 7: Block-scheme of residual spatio-temporal attention module. "Spatial Pool" block carries out global average pooling over spatial dimensions [20].

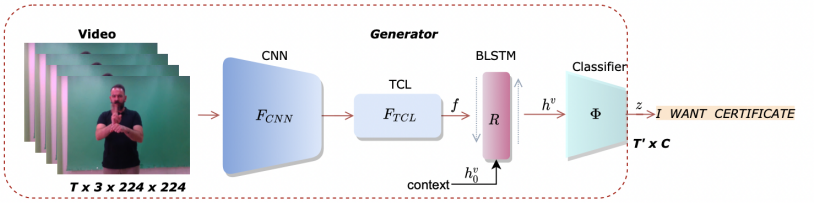


Figure 8: The proposed generator extracts spatio-temporal features from a video and predicts the signed gloss sequences [38].

The discriminator classifies the glosses predicted by the generator with input glosses. This module as illustrated in Fig. 9 has three separate submodules; word embedding layer, gloss-level, and sentence-level module in parallel and fully connected layer. The input to the discriminator is passed through a word embedding layer which is fully connected to map the input to the lower dimension by learning linear projection. The embedding output is passed through the gloss-level submodule having a fully connected layer, a global average pooling layer, and a sigmoid function to produce a score for each gloss. Simultaneously same embedding input is passed through a sentence-level submodule consisting of a BLSTM layer and a fully connected layer with sigmoid function activation to obtain a sentence-level score. The entire score of the discriminator is calculated by combining glass-level and sentence-level scores passing it through the fully connected layer [38].

The model can cover sign language conversation by initializing the hidden state of the BLSTM layer of the generator with context information from the previous sequence or video sequence for Deaf-to-hearing and Deaf-to-Deaf dialogues, respectively. In the end, the sign language translation is performed by a transformer network [47], converting the sign language glosses into natural language text. 3 datasets were used to train and test the performance of the models which are: the RWTH-PHOENIX-Weather 2014 [16], the Greek Sign Language [3] and the Chinese Sign language dataset [19]. The proposed model outperforms previous models on the RWTH-PHOENIX-Weather 2014 baseline, GSL baseline, and CSL dataset baselines with WER of about 23.4%, 2.26% and 2.1% respectively [38]. The authors also investigate the significance of the contextual information for both the deaf-to-deaf and deaf-to-hearing communication [38].

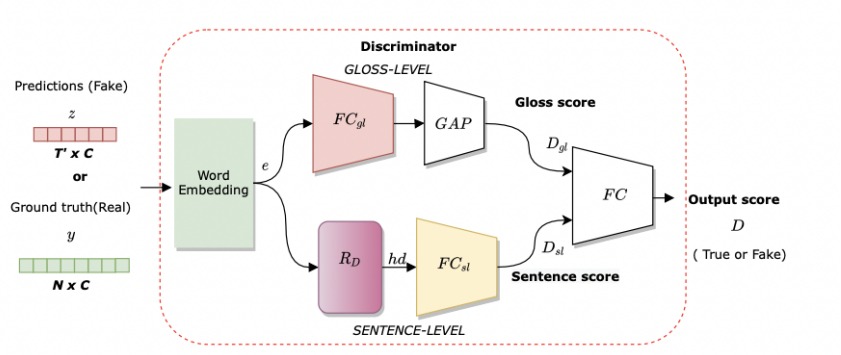


Figure 9: The proposed discriminator aims to distinguish between the ground truth and predicted glosses by modeling text information at both the gloss and sentence levels [38].

3 Experiments & Discussions

For the isolated SLR system, we aim at reproducing the I3D backbone (3D-CNN) described by Li el. al.[32]. The I3D model architecture is defined in Fig. 5. This model is first trained on the ImageNet and then fine-tuned on the Kinetics-400. They model the temporal and spatial information for the sign language by using the 3D ConvNet architecture, finetuned on the WLASL subset dataset. We attempt to first train this model with the WLASL dataset, consisting of 21,083 video samples of about 14 hours of data with 119 different signers, with an average video length of 2.41 seconds. It consists of about at max 2000 glosses, where on average there are 10.5 video samples per gloss.

The dataset is divided into 4 categories according to the vocabulary size i.e. WLASL100, WLASL300, WLASL1000, WLASL2000, which consists of gloss sizes of 100, 300, 1000, and 2000 respectively. The I3D model was implemented in Pytorch and our code for the same is available on github [1]. The model was trained for 200 epochs on the Hipergator with 1 A100 GPU. We performed early stopping when the validation accuracy became stagnant. The samples of gloss were split into three sets of training, validation, and testing with a ratio of 4:1:1, ensuring that each split has at least one sample per gloss.

After finetuning the I3D bone for each subset of the WSLR dataset, we evaluated the models with the original papers' results. We compared our I3D results with Pose-GRU, Pose-TGCN, VGG-GRU, and I3D [32]. We used the mean scores of top-K classification accuracy with $K = \{1, 5, 10\}$ overall sign instances. Table 1 shows the performance of the I3D model compared with the baseline models based on poses and image appearances.

Table 1: Top-1, Top-5, Top-10 accuracy (%) acheived by each mode (row-ise) on th four WLASL subset dataset

Method	WLASL100			WLASL300			WLASL1000			WLASL2000		
	top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-5	top-10
Pose-GRU	46.51	76.74	85.66	33.68	64.37	76.05	30.01	58.42	70.15	22.54	49.81	61.38
Pose-TGCN	55.43	78.68	87.6	38.32	67.51	79.64	34.86	61.73	71.91	23.65	51.75	62.24
VGG-GRU	25.97	55.04	63.95	19.31	46.56	61.08	14.66	37.31	49.36	8.44	23.58	32.58
I3D (Original Paper)	65.89	84.11	89.92	56.14	79.94	86.98	47.33	76.44	84.33	32.48	57.31	66.31
I3D (Our Result)	**67.07**	**84.58**	**90.25**	**56.24**	**78.39**	**84.33**	**47.49**	**75.66**	**82.96**	**30.45**	**55.33**	**63.90**

From the results, we see that the I3D implementation outperforms all the other pose-dataset-based model results. As the image appearance-based baseline was trained in an end-to-end manner for SRL and thus we see that errors that reside in spatial features were reduced during training. From Table 1, the I3D model achieves very good results on low-resource settings i.e. WLASL100 and WLASL300. For the larger dataset i.e. WLASL2000, we see that all the models achieve similar performance which is very close to practical word-level classification scenarios. Therefore, recognition performance on small vocabulary datasets does not reflect model performance on large vocabulary datasets, and

[1] https://github.com/dxli94/WLASL

large-scale sign language recognition is very challenging. This also indicates that, given low resource settings, I3D is the recommended method.

In future work, we believe that the I3D model can be further improved for low-resource settings by leveraging an attention-based mechanism. We also believe that this work can be extended to sentence level as well as story level SLR systems. Finally, we can also scale this proposed approach for online devices (building models with low latency as well as requiring low compute power) by using knowledge distillation.

4 Acknowledgment

The authors gratefully acknowledge the GPU resources provided by the HiPerGator research computing group at the University of Florida. We thank Dongxu Li for promptly responding to our request for the complete WLASL dataset and pre-trained model.

5 Conclusion

In this paper, we studied various deep learning models for Sign Language Recognition Systems taking a wide range of input formats including RGB, Depth data, Thermal data, feature fusion, visual modality, and type of detection i.e. static or dynamic. Concretely we explored four approaches along with their model architecture. We also studied a benchmark comparison of MS-ASL dataset for existing SLR state-of-the-art network architectures We implemented the 3D convolution neural network system (I3D) and evaluated its performance on four categories of WLSR datasets divided on vocabulary size namely WLASL100, WLASL300, WLASL1000, WLASL2000. We demonstrated that I3D implementation outperforms the other pose-dataset-based models by a large margin. We believe that the I3D model can be further improved for low-resource settings by using attention-based mechanisms. We also believe that this work can be extended to sentence level as well as story level SLR systems. Finally, through knowledge distillation, we plan to scale this proposed approach for online devices.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[2] B Acton and J Koum. Whatsapp. *Yahoo*, 2009.

[3] Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*, 2(2), 2020.

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

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

[8] Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10(1):131–153, 2019.

[9] François Chollet et al. Keras. `https://keras.io`, 2015.

[10] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3056–3065, 2017.

[11] HM Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13:2205–2231, 2012.

[12] Naina Dhingra and Andreas Kunz. Res3atn-deep 3d residual attention network for hand gesture recognition in videos. In *2019 International Conference on 3D Vision (3DV)*, pages 491–501. IEEE, 2019.

[13] Endri Dibra, Thomas Wolf, Cengiz Oztireli, and Markus Gross. How to refine 3d hand pose estimation from unlabelled depth data? In *2017 International Conference on 3D Vision (3DV)*, pages 135–144. IEEE, 2017.

[14] Bardia Doosti. Hand pose estimation: A survey. *arXiv preprint arXiv:1903.01013*, 2019.

[15] Centers for Disease Control, Prevention, et al. 2014 annual data early hearing detection and intervention (ehdi) program. 2016.

[16] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3785–3789, 2012.

[17] Kirsti Grobel and Marcell Assan. Isolated sign language recognition using hidden markov models. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 1, pages 162–167. IEEE, 1997.

[18] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Sign language recognition using 3d convolutional neural networks. In *2015 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2015.

[19] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[20] Evgeny Izutov. Asl recognition with metric-learning based lightweight network. *arXiv preprint arXiv:2004.05054*, 2020.

[21] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.

[22] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–433, 2018.

[23] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.

[24] 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*, 141:108–125, 2015.

[25] Oscar Koller, Hermann Ney, and Richard Bowden. Deep learning of mouth shapes for sign language. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 85–91, 2015.

[26] Oscar Koller, R Bowden, and H Ney. Automatic alignment of hamnosys subunits for continuous sign language recognition. Technical report, University of Surrey, 2016.

[27] Oscar Koller, Hermann Ney, and Richard Bowden. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3793–3802, 2016.

[28] 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*, 126(12):1311–1325, 2018.

[29] 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*, 42(9):2306–2320, 2019.

[30] Jim G Kyle, James Kyle, Bencie Woll, G Pullen, and F Maddix. *Sign language: The study of deaf people and their language*. Cambridge university press, 1988.

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[32] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020.

[33] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6205–6214, 2020.

[34] Ita Lifshitz, Ethan Fetaya, and Shimon Ullman. Human pose estimation using deep consensus voting. In *European Conference on Computer Vision*, pages 246–260. Springer, 2016.

[35] Kian Ming Lim, Alan WC Tan, and Shing Chiang Tan. Block-based histogram of optical flow for isolated sign language recognition. *Journal of Visual Communication and Image Representation*, 40:538–545, 2016.

[36] Manuel J Marin-Jimenez, Francisco J Romero-Ramirez, Rafael Munoz-Salinas, and Rafael Medina-Carnicer. 3d human pose estimation from depth maps using a deep combination of poses. *Journal of Visual Communication and Image Representation*, 55:627–639, 2018.

[37] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[38] Ilias Papastratis, Kosmas Dimitropoulos, and Petros Daras. Continuous sign language recognition through a context-aware generative adversarial network. *Sensors*, 21(7):2437, 2021.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[40] Lionel Pigou, Mieke Van Herreweghe, and Joni Dambre. Gesture and sign language recognition with temporal residual networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3086–3093, 2017.

[41] Leigh Ellen Potter, Jake Araullo, and Lewis Carter. The leap motion controller: a view on sign language. In *Proceedings of the 25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration*, pages 175–178, 2013.

[42] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications*, 150:113336, 2020.

[43] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021.

[44] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017.

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

[46] 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.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[48] Betty Vohr. Overview: Infants and children with hearing loss-part i. *Mental retardation and developmental disabilities research reviews*, 9(2):62–64, 2003.

[49] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[50] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[51] Min Wang, Xipeng Chen, Wentao Liu, Chen Qian, Liang Lin, and Lizhuang Ma. Drpose3d: Depth ranking in 3d human pose estimation. *arXiv preprint arXiv:1805.08973*, 2018.

[52] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[53] Wikipedia contributors. List of sign languages — Wikipedia, the free encyclopedia, 2022. URL `https://en.wikipedia.org/w/index.php?title=List_of_sign_languages&oldid=1077124451`. [Online; accessed 24-April-2022].

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

[55] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.

[56] Zhihao Zhang, Junfu Pu, Liansheng Zhuang, Wengang Zhou, and Houqiang Li. Continuous sign language recognition via reinforcement learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 285–289. IEEE, 2019.

[57] Lihong Zheng, Bin Liang, and Ailian Jiang. Recent advances of deep learning for sign language recognition. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2017.

[58] Mingjie Zhou, Michael Ng, Zixin Cai, and Ka Chun Cheung. Self-attention-based fully-inception networks for continuous sign language recognition. In *ECAI 2020*, pages 2832–2839. IOS Press, 2020.

[59] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.