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*

1 Project Proposal

1.1 Introduction

Our team consists of three members; Vyom Pathak, Dipali Patidar, and Shefali Mishra. We intend to work on American Sign-Language Recognition from video using Deep Learning algorithms. We intend to work on the following research papers, and hopefully contribute to improving one of the techniques:

1. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison [27]
2. Transferring Cross-domain Knowledge for Video Sign Language Recognition [28]
3. ASL Recognition with Metric-Learning based Lightweight Network [17]
4. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language [18]
5. Continuous Sign Language Recognition through a Context-Aware Generative Adversarial Network [30]

The proposal is structured as follows, first we describe the motivation behind the problem we intend to solve in Section 1.2, then each of the described papers are summarized in Section 1.3, and finally in Section 1.4 we describe methods from each of these papers that we intend to work on/improve upon along with the dataset details.

1.2 Motivation & Background

Based on a survey done by the CDC in 2019, about 1.7 per 1000 babies that were born that that year were identified with a permanent hearing loss [12]. 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 [38]. 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 form 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 for daily interactions [25].Sign language consists of the usage of different part of the body, like fingers, hands, arms, head, body, and facial expressions [7]. It accounts for five main parameters namely, hand-shape, palm orientation, location, movement, and expressions signals [25]. For an accurate sign-word, all of these five parameters must be performed/interpreted correctly. A survey by World Federation of the Deaf has reported that there are over 300 sign languages around the world that about 70 million deaf people use.

Preprint. Under review.

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 an important problem to tackle. Here, we aim to automatically translate sign languages using vision technologies to text. With the advent of Deep Learning [26], GPUs i.e. compute power to boost these algorithms, along with the development of strong frameworks like TensorFlow [1], PyTorch [31], Keras [8], and MXNET [6]; we aim to solve this problem using video-based Deep Learning algorithms.

There have been several attempts at solving this problem using different types of features (hand pose, face expressions, and body posture) [33]. In terms of the type of data, the algorithms are divided into RGB and Depth data [33]. There has 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) domain where the dynamic ones are the more complex because of its continuous nature [19, 33].

Existing word-level SLR models have been developed on small-scale/private datasets with less that about 100 words. These methods include using hand-crafted features such as histogram of optical flow [29], and HOG-based features [4, 10, 32]. Hidden Markov Model [14, 35] 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 [15, 42] have also made great breakthroughs. Further more, after the development of WLASL dataset consisting of 2000 sign-poses, there have been developments in using pose-based TGCN method for solving the problem at hand [27]. With the advent of new technologies such as semi-supervision, 3D-convolution based algorithms have made strides in low-resource settings [28].

For the sentence-level SLR models, the largest known benchmark dataset is the RWTH-PHOENIX-Weather 2014 consisting of 1080 German language sign-poses [13]. For the American Sign Language, one of the benchmark dataset is the MS-ASL dataset consisting of 1000 sign-poses. Koller developed several methods on continuous SLR systems including iterative training using expected Maximization, using 2D-Convolution neural networks, hybrid 2D CNN with HMM models [20–23]. 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 [9, 24]. With the introduction of the Attention mechanism, for extracting important information from the embedded representation of the video; Transformer based architectures were developed [16, 43, 44].

We aim at improving the low-resource SLR systems by understanding and working with SOTA SLR methods developed for American Sign Language (ASL).

1.3 Paper Summary

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

WLASL 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 datasets consists of 2,000 words, 21,083 videos with 119 different signers, being the largest word-level ASL video dataset. The paper also compares the performance of different baseline systems using two approaches namely, Visual appearance based, and 2D Human poses based. 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) [27], and second is the 3D CNN system (I3D backbone) [5]. Both of these methods perform on-par with respect to the model's capacity to hold the information. For the 2D-human poses based system extracts the pose of the whole body-frame and then predicts the word accordingly. For this method, they compare a baseline RNN model (Pose-GRU) [27], and propose a novel Graph Convolution Neural Network based algorithm (Pose-GCN) [27], which out-performs the baseline in pose-based systems in terms of speed and accuracy. Pose-GCN stacks GCN layers with residuals. The I3D achieves the best recognition performance. They further compare all of these models, and the authors concludes with the future works where the research community working on isolated word detecting for the ASL language is able to create SOTA models with such a voluminous amount of data.

1.3.2 Transferring Cross-domain Knowledge for Video Sign Language Recognition

Build up on the WLASL dataset, this paper proposes an algorithm to solve the SLR problem in the low-resource settings. In the WLASL dataset, there is a sub-set of dataset where only 10 examples per class is present, which entails the low-resource problem. The authors propose a transfer learning/semi-supervised learning based algorithm for LSR [28]. They aim at transferring knowledge between the News Data domain and isolated data domain. First they train a localizer for both the news an the isolated word signs data. After that, they use the features learned from this technique into a prototypical memory containing the information as a retrieval function for descriptor where the original video is trained using a Temporal Attention mechanism. The features from the original video were extracted using a I3D-backbone architecture [5]. Finally the output of the temporal attention mechanism is passed onto the classifier to predict the type of word. The results for the SOTA performance for the WLASL dataset.

1.3.3 ASL Recognition with Metric-Learning based Lightweight Network

On the flip-side of training a which gives on-par performance with large models as well as is light weight, 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 [17]. For training as well as testing the system they use the MS-ASL dataset [18]. Due to the low-resource setting, the metric learning approach is adopted where they use the AM-Softmax loss with auxiliary self-supervision loss [40]. 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 [11, 39]. They also provide an ablation study on different parameters for handling the tradeoff between the accuracy and the size of the model.

1.3.4 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 [18]. 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 in 4 subsets including 100, 200, 500 and 1000 most frequent words subsets called as ASL100, ASL200, ASL500 and ASL1000. The paper evaluates the existing three approaches 2D-CNN-LSTM [41], body key-point [45], CNN-LSTM-HMM [34] and 3D-CNN [36] as baselines. 3D-CNN baseline achieved good results in this challenging, uncontrolled data compared to other two baseline models and authors proposed it as as powerful network for sign language recognition. The experimental result presented in 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) [45] is doing relatively better compared to 2D-CNN mode.

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

The author here proposes a novel Sign Language Recognition Generative Adversarial Network (SLRGAN) [30]. The network architecture consists of generative model that learns the gloss of the signs by extracting the spatial as well as the temporal features from the video sequences. The discriminator then evaluates the quality of the generator's predictions by modeling the textual information at the sentence and gloss levels. At the end, sign language translation is performed by a transformer network [37], converting the sign language glosses into natural language text. They used 3 datasets to train and test the performance of the models which are: the RWTH-PHOENIX-Weather 2014 [13], the Greek Sign Language [3] and the Chinese Sign language dataset [16]. It 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 [30]. The authors also investigate the significance of the contextual information for both the deaf-to-deaf and deaf-to-hearing communication.

1.4 Methods & Datasets

Isolated American Sign Language Recognition using I3D Backbone For the isolated SLR system, we aim at implementing the I3D backbone (3D-CNN). This model is first trained on the ImageNet, and then fine-tuned on the Kinetics-400. They model the temporal and spatial by using the 3D ConvNet architecture, finetuned on the WLASL subset dataset. We attempt at using this model with the WLASL dataset, consisting of 21,083 video samples of about 14 hours of data with 119 different signers, of 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.

Continuous American Sign Language Recognition using GANs For the continuous SLR system, we aim at implementing a context-aware generative adversarial network. The SLRGAN is trained such that the video is passed on to a generator, then the generator predicts the sentence. This sentence is then passed onto the Discriminator, where the discriminator compares the prediction with the Ground truth and spits out the score given that if the pair is either fake or either real. The discriminator works at both Gloss as well as Sentence level. This score is used to train the whole structure. Now, the output of the video is a sentence with jumbled words (sign language gloss), which can be translated to natural language text by further passing them through a Transformer architecture [37]. The model will be trained and tested on 3 datasets namely the RWTH-PHOENIX-Weather 2014 [13], Greek Sign Language (GSL) Signer Independent (SI) [3] and the Chinese Sign Language dataset [16].

For both of these methods we will first try to reproduce the results on the given datasets, and then modify the architecture to imporve the performance of the model on low-resource settings.

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] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

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

[31] 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`.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.