A Two-Pronged Approach to ISLR Recognition

Farid Karimli, Jasmine Pham

April 30, 2024

Abstract

We propose VideoASL, an innovative approach to isolated sign language recognition (ISLR) from video. Using a two-pronged approach, our model uses video features with hand landmark locations to recognize large-scale American Sign Language (ASL). We aim to address challenges in dictionary retrieval, which are essential tools for language learners and users. Potential applications include live sign transcription, as well as creating a reliable ASL-to-English dictionary.

Introduction

Sign language serves as a vital mode of communication for individuals with hearing impairments, offering them a means to express themselves and interact with the world around them. However, there is a gap of communication between those proficient and those not proficient in ASL. Moreover, large-scale isolated sign language recognition (ISLR) from videos remains an ongoing challenge for several reasons:

- Dynamic nature of signs: ASL involves continuous movements and transitions between different hand shapes and positions.
- Variability among signers: Sign language users have different signing styles, speeds, and accuracies, which can introduce variability in hand shapes, movements, and durations.
- Background clutter from the videos can affect the visibility of hand gestures.

In response to this need, we propose VideoASL, an approach aimed to advance ISLR through the integration of masked autoencoders (MAE) and convolutional neural network (CNN) architecture.

Related Work

ISLR Recognition

Research on Isolated Sign Language Recognition (ISLR) has been on the rise. Earlier solutions involved using traditional machine learning techniques like Hidden Markov Models [1], Boosted Trees [2] and Support Vector Machines [3]. Newer solutions employ deep learning techniques, namely convolutional neural networks (CNNs) [4], transformers [5] and fitting models on human pose features [6].

However, these approaches did not achieve adequate accuracy on large-scale sign language vocabularies, largely because of the lack of such dataset previously. The ASL-Citizen dataset (2023) showed that it can significantly improve models' performance, up from around 30% recall-at-1 to around 63% using I3D [7], amongst other metrics. I3D employs two streams of 3-D CNNs, followed by feed-forward computation. The model achieved state-of-the-art performance on action recognition on the UCF101 and HMDB datasets.

Video Classification using Self-Attention

Three of the best performing models that use self-attention on video classification tasks are VideoMAE, Timesformer and Video Vision Transformer (ViVit) [8, 9, 10]. They process videos by first segmenting them into non-overlapping patches, which are then flattened and linearly embedded. The models employ a series of transformer blocks that consist of self-attention and MLP (multi-layer perceptron) layers. All three incorporate both temporal and spatial features in their self-attention mechanisms. Timesformer and Video Vision Transformer process the spatial and temporal features sequentially. VideoMAE uses a joint-space time transformer as its backbone (where time and space is handled concurrently), and masks portions of the input videos to force the model to pay increased attention to the portions that are visible.

We believe that the self-attention mechanisms will be very effective for sign language recognition as signs can have complicated sequences of hand movements and different signs could share some movement sequences. Self-attention would be able to quantify and weight different stages of signing.

Method

Model

Our methodology is inspired by [11], where the authors used two CNNs to extract features from ASL finger-spelling images. The two are trained in parallel; the first CNN receives a

processed version of the original image, and the second CNN gets an image of only hand landmark annotations. Figure 1 shows the diagram of the workflow.

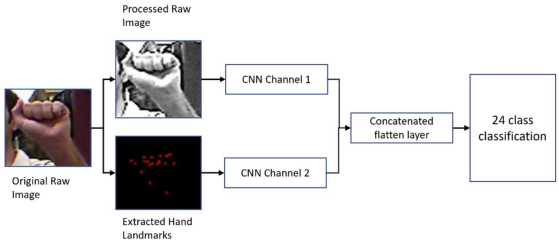

Figure 1: Workflow of the Multi-Headed CNN

After the feature vectors are extracted, they are concatenated and processed using a series a fully-connected layers, similar to traditional classification approaches for downstream tasks. We aim to imitate this workflow using video classification models, replacing the CNNs.

We will take an iterative approach, first focusing on training a video classification model on the entire dataset to find the best feature extractor. After that, we aim to add the second stream with hand landmarks positions using Mediapipe [12].

Initially, we chose the Video Vision Transformer (ViVit) for our feature extractor. After extensive testing, we found that ViVit did not perform well. This may be due to vision transformer models requiring a large number of examples per label for adequate results on downstream tasks, which ASL-Citizen did not provide. The transition to using VideoMAE marked a significant improvement, and we used VideoMAE moving forward.

We believe that the attention modules of VideoMAE will be able to effectively model relationships between different stages of signing motion, particularly for complex signs. Figure 2 displays the architecture of our approach.

Dataset

	Train	Val	Test
Signers	35	6	11
Videos	40,154	10,304	32,941
Signer Distribution	60% F, 40% M	83% F, 170% M	55% F, 45% M

Table 1: Video distribution of ASL-Citizen Dataset

We use the ASL Citizen dataset [13]. This dataset represents the first large-scale, continuous ASL dataset, emphasizing conversational ASL within a diverse range of contexts,

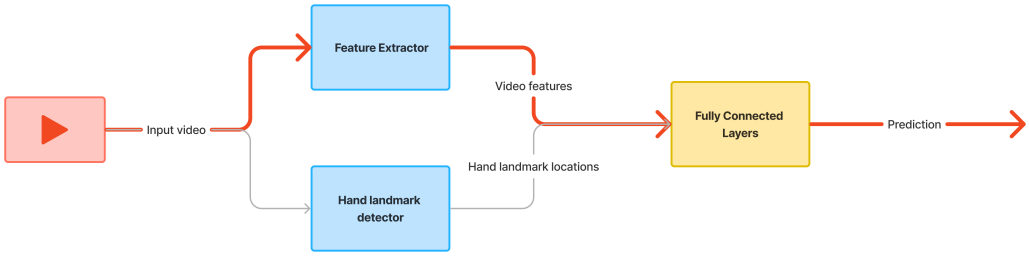

Figure 2: Proposed Architecture

surpassing previous datasets that primarily focused on single-sign instances. It has about 84,000 video recordings of over 2700 ASL signs. Its rich collection of real-life, continuous sign language interactions provides a more realistic and comprehensive foundation for developing and testing our ASL recognition algorithms.

The crowd-sourced aspect of MS-ASL-Citizen introduces a better variability in signers. This diversity ensures that our model is exposed to a broad spectrum of signing variations, which is critical for developing a robust and inclusive ASL recognition system. The richness and diversity of the MS-ASL-Citizen dataset, therefore, would enhance the potential accuracy and applicability of our project outcomes in real-world scenarios.

The videos are split into training, validation, and test sets. It's worth noting that the splits are signer independent, making it more aligned with real-world cases, where users are unlikely to be previously seen in training data.

	I3D	VideoMAE			
Sign Amount	2731	100	500	1000	2731
Training Size	40,154	1,611	15,081	20,081	40,154
Validation Size	10,304	302	3,430	3430	10,304
Test Size	32,941	1,191	12,025	7025	32,941
Test Accuracy	0.631	**0.853**	**0.880**	**0.904**	TBD
Test Top 5	0.861	**0.970**	**0.963**	**0.972**	TBD
Test Top 10	0.908	**0.972**	**0.980**	**0.981**	TBD

Table 2: Results of fine-tuning VideoMAE on ASL Citizen dataset

Evaluation Results

To evaluate the models' performance, we used accuracy and recall-at-k as our metrics, which are widely used in information retrieval contexts. We present our results of fine-tuning VideoMAE on the ASL Citizen dataset (our first step) and compare them to the performance of the I3D model developed by the dataset creators in Table 2. We trained the model for 10 epochs, using the Adam optimizer with a learning rate of 0.00001 (10^{-5}). For time-efficiency reasons, we evaluate VideoMAE on subsets of the signs. So far, we have been able to train the model on subsets of 100, 500, and 1000 signs. We see that on those subsets, VideoMAE achieves remarkable accuracy. We do not have the performance results of the I3D model on subsets of the dataset, so these results cannot be compared directly. However, they show that VideoMAE is able to effectively classify ASL from video with high accuracy. We are currently in the process of training the model on the entire dataset of 2731 signs. We were not able to incorporate the hand landmark locations branch of our proposed architecture in time.

Due to resource restrictions, instead of training the model on all 2731 labels (signs) we explored training the model incrementally on one subset of the signs at the time. Unfortunately, the model exhibited significant forgetting, so we abandoned this approach.

IMPORTANT: We realized too late that there is possible data leakage in our dataset splits. We found the distributions of videos per label to be rather strange, with around 15-24 videos per sign in the training split and 13-15 videos in the testing split. We moved some videos for each sign from the testing split to the training split, removing them from the testing split. The validation split was left untouched. Although we're not sure if this constitutes direct data leakage, we believe that the testing results may be inflated. We did not have time to evaluate the problem in time.

Conclusion and Future Work

VideoASL represents a novel contribution to the field of isolated sign language recognition from video data. Our approach is engineered to enhance dictionary retrieval systems—a critical resource for learners and practitioners of sign language. By leveraging the expansive ASL Citizen dataset, VideoASL on its own is able to effectively classify large scale ASL from video with high accuracy, albeit on subsets of the signs. Training on the entire set of signs is pending, as is supplementing video features with hand landmark locations. We hope that once we finalize our model, we will be able to beat the performance of the dataset creators on isolated sign language recognition. Ultimately, VideoASL aims to be an innovative and practical solution that will significantly benefit the hearing-impaired community.

In the future, the model could be made more robust by evaluating it on completely new

signers. Also, we believe the model can be used for real-time sign recognition.

References

[1] Christian Vogler and Dimitris Metaxas. Adapting hidden markov models for asl recognition by using three-dimensional computer vision methods. In 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, volume 1, pages 156–161. IEEE, 1997.

[2] Ong, E. & Bowden, R. A boosted classifier tree for hand shape detection. *Sixth IEEE International Conference On Automatic Face And Gesture Recognition, 2004. Proceedings.*. pp. 889-894 (2004)

[3] Monteiro, C., Mathew, C., Gutierrez-Osuna, R. & Shipman, F. Detecting and Identifying Sign Languages through Visual Features. *2016 IEEE International Symposium On Multimedia (ISM)*. pp. 287-290 (2016)

[4] Rao, G., Syamala, K., Kishore, P. & Sastry, A. Deep convolutional neural networks for sign language recognition. *2018 Conference On Signal Processing And Communication Engineering Systems (SPACES)*. pp. 194-197 (2018)

[5] Boháček, M. & Hrúz, M. Sign Pose-based Transformer for Word-level Sign Language Recognition. *2022 IEEE/CVF Winter Conference On Applications Of Computer Vision Workshops (WACVW)*. pp. 182-191 (2022)

[6] Selvaraj, P., Nc, G., Kumar, P. & Khapra, M. OpenHands: Making Sign Language Recognition Accessible with Pose-based Pretrained Models across Languages. *Proceedings Of The 60th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers)*. pp. 2114-2133 (2022,5), https://aclanthology.org/2022.acl-long.150

[7] Carreira, J. & Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset.. *Computer Vision And Pattern Recognition (CVPR)*. pp. 4724-4733 (2017)

[8] Tong, Z., Song, Y., Wang, J. & Wang, L. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. (2022)

[9] Bertasius, G., Wang, H. & Torresani, L. Is Space-Time Attention All You Need for Video Understanding?. (2021)

[10] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M. & Schmid, C. ViViT: A Video Vision Transformer.. *IEEE International Conference On Computer Vision (ICCV)*. pp. 6816-6826 (2021)

[11] Pathan, Refat Khan, et al. "Sign Language Recognition Using the Fusion of Image and Hand Landmarks through Multi-Headed Convolutional Neural Network." Nature News, Nature Publishing Group, 9 Oct. 2023.

[12] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C., Yong, M., Lee, J., Chang, W., Hua, W., Georg, M. & Grundmann, M. MediaPipe: A Framework for Building Perception Pipelines. *ArXiv.* **abs/1906.08172** (2019)

[13] Desai, Aashaka and Berger, Lauren and Minakov, Fyodor O and Milan, Vanessa and Singh, Chinmay and Pumphrey, Kriston and Ladner, Richard E and Daumé III, Hal and Lu, Alex X and Caselli, Naomi and Bragg, Danielle *ASL Citizen: A Community-Sourced Dataset for Advancing Isolated Sign Language Recognition*