Few-shot Learning for Sign Language Recognition with Embedding Propagation

Amjad Alsulami[1], Khawlah Bajbaa[1], Hamzah Luqman[1,2,*], Issam Laradji[3]

*Corresponding author
[1]King Fahd University of Petroleum & Minerals
[2]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, Dhahran, Saudi Arabia
{g20210113, g202115030, hluqman}@kfupm.edu.sa
[3]ServiceNow
issam.laradji@servicenow.com

Abstract- Sign language is a primary channel for the deaf and hard-hearing to communicate. Sign language consists of many signs with different variations in hand shapes, motion patterns, and positioning of hands, faces, and body parts. This makes sign language recognition (SLR) a challenging field in computer vision research. This paper tackles the problem of few-shot SLR, where models trained on known sign classes are utilized to recognize instances of unseen signs with only a few examples. In this approach, a transformer encoder is employed to learn the spatial and temporal features of sign gestures, and an embedding propagation technique is used to project these features into the embedding space. Subsequently, a label propagation method is applied to smooth the resulting embeddings. The obtained results demonstrate that combining embedding propagation with label propagation enhances the performance of the SLR system and achieved an accuracy of 76.6%, which surpasses the traditional few-shot prototypical network's accuracy of 72.4%.

Keywords- Sign language recognition; Sign language translation; Few-shot learning

1 Introduction

Sign language represents the main channel for deaf or vocal impairment people to communicate, exchange knowledge and express their feelings with others, and build social relationships (1). As technology advances, people with hearing impairments and deafness can communicate with their community more efficiently by translating sign language into natural languages and vice versa (2).

sign language recognition (SLR) is one of the most widespread critical problems addressed in computer vision (3). Despite most signs have clearly defined looks, they are slightly different from one another visually (4; 5). As a result, for SLR to be a comprehensive technique, it requires fundamental advancements in modeling and identifying fine-grained spatiotemporal patterns of hand movements (3). There are also other factors that affect

the performance of the recognition task, including variations in the visibility perspective (6), the development of sign languages over time (7), and regional differences in sign language (8).

SLR technique can be categorized into isolated and continuous SLR. Isolated SLR systems target word-level signs, whereas continuous SLR approaches recognize sign language sentences (9). Isolated SLR has been studied extensively in the literature compared with continuous SLR (2). One main issue with these approaches is the need for a large number of annotated samples per sign (10) (11) (12). Annotated samples of all signs in all languages of interest must be collected to satisfy this dependency. These samples must include signs expressed multiple times by multiple individuals per sign under different recording settings. Globally, more than 140 sign languages are spoken along with several dialects (13). Consequently, scaling up SLR is hindered by the demand for supervised examples. Recently, a few solutions have attempted to overcome this problem using few shot learning to recognize unseen signs with few annotated samples (14; 15; 16; 3). Few-shot learning is a technique to learn class discrimination from a limited number of labeled samples.

In this paper, we introduce a few-shot learning approach for SLR that is specifically designed to generalize well to unseen classes. Our approach accepts pose information of sign gestures and feeds them into a transformer encoder to extract a set of features encoding spatial and temporal information. We then transform these features from the features space to the embedding space by leveraging embedding propagation with label propagation techniques. The proposed approach has been evaluated using the WLASL-100 dataset and the obtained results demonstrate the effectiveness of combining embedding propagation with label propagation for few-shot learning for SLR.

This paper is arranged as follows. Section 2 begins with a review of the relevant literature. Then in Section 3, we present the few-shot SLR method, and the experimental work is presented in Section 4. Our conclusion and future work are presented in Section 5.

2 Related work

Sign language recognition (SLR). Several techniques have been developed in the last two decades to recognize sign language gestures(1; 2). The majority of these techniques focus mainly on tracking and recognizing signer's hands (17; 18; 19; 20). Hands motion represents the manual part of the sign language, whereas body movements and facial expressions represent the non-manual part of the sign language. Few studies in the literature that tried to simultaneously recognize manual and non-manual signs (21; 22; 23).

There have been several attempts to develop SLR approaches based on deep learning in recent years. Camgoz et al. (24) proposed a transformer-based model for Continuous SLR and translation. The temporal information of the sentence's signs is learned in a unified way using a Connectionist Temporal Classification (CTC) loss. A previous study (25) proposed a progressive transformer to translate discrete speech sentences into continued 3D expression sequences. In this work (26), Tao et al. (26) employed a multi-view

augmentation of American sign alphabets to address incomplete occlusions and reduce the impact of perspective changes. The resulting augmented images are then fed into a simple convolution neural network (CNN). In another study (27), a CNN was used to combine several spatial and spectral constructions of images of hand gestures to create a method for the visual detection of fingerspelling in gestures. The proposed method creates spatiotemporal images of hand sign motions in Gabor spectral formats and then utilizes an improved CNN to categorize the gestures in the joint space into appropriate classes.

SAMSLR, a multi-modal skeleton-aware SLR framework, was proposed as a way to exploit multi-modal information for SLR (28). Huang et al. used a 3D-CNN to learn spatial-temporal aspects of sign gestures (29). A set of features were extracted from the signer's hands to highlight the significant changes in hand motions. A dataset consisting of 25 signs was used to evaluate the proposed approach and an accuracy of 94.2% was reported. Another system was developed for recognizing sign language alphabets and an accuracy of 98.9% was reported (29).

Using motion history images produced from color frames, authors in (30) proposed a model for isolated SLR. This technique was used to summarize the spatiotemporal information of each sign. A model that accepts RGB and motion history images was implemented as a movement-based spatial attention module combined with the 3D architecture. Using a late fusion technique, the model features are directly applied to the features of the 3D model. Albanie et al. (31) attempted to deal with the lack of annotated sign language data by detecting keywords in processed TV broadcasts. In 1,000 hours of video, 1000 signs are automatically localized through weakly aligned subtitles and keyword spotting. Authors in (32) offered an integrated framework for multiple instance learning in ongoing sign language movies.

Few-shot SLR In contrast to traditional supervised-based SLR, few-shot learning-based approaches recognize unexplored sign classes with either very few training samples (few-shot SLR) or no visual training samples (zero-shot SLR). Cornerstone Network (CN) is a few-shot learning model proposed by (14) that can mitigate the impact of support samples in unsuitable conditions. In this network, the mean with the bias of support samples are extracted from the input samples and used as an input features. Then, neural networks with clustering algorithms were used to learn the mapping from input space to the embedding space. As with the Siamese networks, the feature extraction network was trained in the same manner so that the features from the heterogeneous data are distributed as widely as possible. Similarly, Shovkoplias et al. (15) investigated several few-shot learning methods, such as Model-Agnostic, Meta-Learning, Matching Networks, and Prototypical networks, to classify electromyogram recordings of deaf and dumb gestures. Authors in (16) employed a pre-trained key-point predictor to keep only the information related to the body, hand, and face and discard other areas. This allows better comparison between vector embeddings as rich representations are learned from body key point sequences. Using k-nearest neighbors, cosine similarity, and Prototypical networks, the new input vector is classified by comparing its distance to a few examples of each class.

Bilge et al. (3) applied zero-shot learning to class sign language gestures without any annotated samples. In their work, semantic class representations are constructed from readily available textual sign descriptions derived from sign language dictionaries. These representations are used to map signs during the inference to their corresponding classes. Similarly, a zero-shot learning framework is used to develop spatiotemporal models of body and hand regions with the use of semantic class representations (33). RGB and depth modalities were used in this study. The approach includes two vision transformer models that identify body parts and segment them into 9 parts. Then, a set of visual features are extracted by the second transformer.

3 Methodology

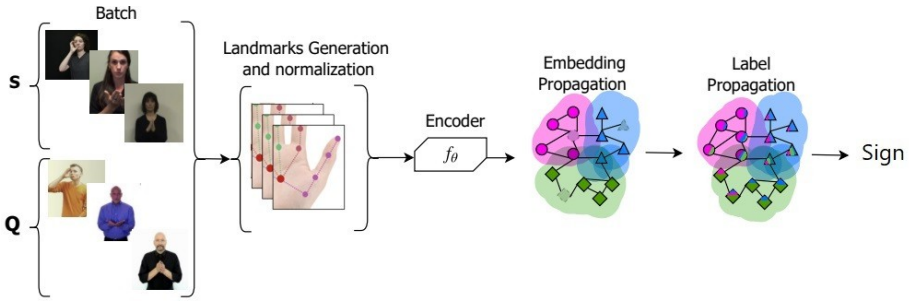

Figure 1: The proposed framework. Embedding and label propagations representations are taken from (34)

In this section, we present an overview of our proposed pipeline, illustrated in Figure 1. The pipeline's architecture integrates the transformer encoder (35) with embedding propagation (34). Initially, the transformer encoder model extracts features from each sign gesture. These features are subsequently mapped to embeddings via the embedding propagation component. We then evaluate two approaches for embedding smoothing, label propagation and prototypical network. Finally, the refined embeddings are input into a classifier to categorize each sign into its corresponding label.

3.1 Transformer Model

A transformer-based model proposed by (35) is used in our pipeline as a feature extractor to learn body pose representations. The features are extracted using the transformer's encoder, while the decoder is replaced by the embedding propagation component. Each video frame undergoes standard pose estimation preprocessing, identifying head, body, and hand landmarks. To prevent model overfitting and enhance its generalization, the skeletal data is augmented during training, inspired by the techniques proposed in (35). Specifically, every joint coordinate in each frame is randomly rotated up to 13 degree

angle. These joint coordinates are then transformed into a new plane, giving the video a tilted appearance. Subsequently, the landmark is rotated relative to the current landmark as it passes through the keypoints of both hands. Following this, irrelevant spatial features are largely removed by normalizing the signer's body proportions, camera distance, and frame location, resulting in a vector of normalized body poses as input to the model. Each frame's pose vector consists of 54 joint locations, which are then encoded with positional information. The learned encoding is used with a dimension of 108 and is added elementwise to the pose vector. The input sequence is fed into the transformer's encoder layers, passing through a self-attention module and a two-layer feedforward network. The self-attention module comprises nine heads and six encoder layers.

3.2 Embedding Propagation

Embedding propagation is a technique to map features into a set of interpolated features called embeddings. In this work, we used the embedding propagation technique proposed in (34). This technique takes the extracted input features using the transformer encoder into the episodic data. Then, it produces a set of embeddings $\tilde{z}_i$ in two steps. First, for every pair of features (i, j), the distance is calculated as $d_{ij}^2 = z_i - z_j{}_2^2$ and the adjacency matrix as $A_{ij} = exp(-d_{ij}^2/\sigma^2)$ where σ^2 is a factor for scaling and $A_{ii} = 0$ for all i. Then, a Laplacian of the adjacency matrix is computed as follows:

$$L = D^{-1/2} * AD^{-1/2}, D_{ii} = \sum_j A_{ij} \tag{1}$$

Then, the propagator matrix is obtained as follows,

$$P = (I - \alpha L)^{-1} \tag{2}$$

where I is an identify matrix and $\alpha \in R$ is a factor for scaling, and the final embeddings are computed as follows,

$$\overline{z_i} = \sum_j P_{ij} Z_j \tag{3}$$

Embedding propagation removes unwanted noise from the feature vectors since the $\overline{Z_i}$ are now a weighted sum of their neighbors.

To perform manifold smoothing on the resulting embedding, we evaluated label propagation and prototypical network (36) techniques. The model optimization and classification are performed on the output of the smoothing technique.

4 Experimental work

Dataset. We utilized the Word-Level American Sign Language (WLASL) dataset to train and evaluate our proposed approach (37). WLASL is a dataset of American Sign Language comprising 100 distinct sign gestures, each performed by multiple signers, with

more than three signers executing each sign. The dataset includes pose information for all the signs. In our work, we divided the data into three sets: a base set with 90 gestures, a validation set with 5 gestures, and a novel class set with 5 gestures. The base and validation sets were used during the pretraining phase, while the novel set was used during the inference phase. During inference, we divided the novel set into support and query sets.

Experiments Setup. The models are optimized using an SGD optimizer during the training phase with a learning rate of 0.0001 selected empirically. Every time the model reaches a plateau, which occurs when the validation loss has not decreased for 10 epochs, we reduce the learning rate by a factor of 10.

Table 1: Recognition accuracies of the proposed system with different number of samples in the support set. The highest accuracy is bolded and the second highest score is underlined.

Support set size	Without Embedding Propagation		With Embedding Propagation	
	Label Propagation	Prototypical Networks	Label Propagation	Prototypical Networks
1	72.2	67.2	70.8	68.6
5	72.4	73.4	**76.6**	72.2
10	69.8	65.4	68.8	<u>76.0</u>

Results and discussion. We evaluated the proposed model using various configurations by varying the number of samples in the support set. The results, presented in Table 1, demonstrate the impact of embedding propagation on the model's performance in SLR with limited samples. We evaluated system components with and without embedding propagation to highlight their effectiveness. As indicated in the table, an accuracy of 76.6% was achieved using the label propagation method combined with embedding propagation, compared to the same settings without embedding propagation. The second-highest accuracy, 76.0%, was obtained with prototypical networks with embedding propagation, marking an improvement of approximately 11% over the same settings without embedding propagation.

It is also evident that both smoothing techniques, label propagation and prototypical networks, performed effectively with the transformer model using a small number of samples in the support set. Although increasing the number of samples generally enhanced the performance of all techniques, some models exhibited overfitting, which may explain the performance decline when 10 samples were used in the support set.

5 Conclusions

In this paper, we proposed a few-shot learning method for SLR designed to generalize effectively to unseen classes. Our approach maps features in the input space to embedding space using embedding propagation combined with label propagation techniques. Initially, sign gesture features are extracted from the input frames using a transformer

encoder. These features are then mapped to the embedding space through an embedding propagation method, followed by label propagation to smooth these embeddings. We evaluated the proposed method using the WLASL-100 dataset, and the experimental results demonstrate the superiority of combining embedding propagation with label propagation compared to the prototypical network. For future work, we plan to evaluate our approach on different sign language datasets to further assess its generalization capabilities.

Aknowledgment

The authors would like to acknowledge the support received from the Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) under the SDAIA-KFUPM Joint Research Center for Artificial Intelligence Grant no. JRC-AI-RFP-14.

References

[1] E.-S. M. El-Alfy, H. Luqman, A comprehensive survey and taxonomy of sign language research, Engineering Applications of Artificial Intelligence 114 (2022) 105198.

[2] S. Alyami, H. Luqman, M. Hammoudeh, Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects, Information Processing & Management 61 (5) (2024) 103774.

[3] Y. C. Bilge, R. G. Cinbis, N. Ikizler-Cinbis, Towards zero-shot sign language recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 1–1doi:10.1109/TPAMI.2022.3143074.

[4] Y. Wu, T. S. Huang, Vision-based gesture recognition: A review, in: International gesture workshop, Springer, 1999, pp. 103–115.

[5] A. a. I. Sidig, H. Luqman, S. A. Mahmoud, Arabic sign language recognition using optical flow-based features and hmm, in: Recent Trends in Information and Communication Technology: Proceedings of the 2nd International Conference of Reliable Information and Communication Technology (IRICT 2017), Springer, 2018, pp. 297–305.

[6] C. Neidle, A. Thangali, S. Sclaroff, Challenges in development of the american sign language lexicon video dataset (asllvd) corpus, in: 5th workshop on the representation and processing of sign languages: interactions between corpus and Lexicon, LREC, Citeseer, 2012.

[7] C. Lucas, R. Bayley, Variation in sign languages: Recent research on asl and beyond, Language and Linguistics Compass 5 (9) (2011) 677–690.

[8] C. Valli, C. Lucas, Linguistics of American sign language: An introduction, Gallaudet University Press, 2000.

[9] R. Rastgoo, K. Kiani, S. Escalera, Sign language recognition: A deep survey, Expert Systems with Applications 164 (2021) 113794.

[10] N. Cihan Camgoz, S. Hadfield, O. Koller, R. Bowden, Subunets: End-to-end hand shape and continuous sign language recognition, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3056–3065.

[11] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural sign language translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7784–7793.

[12] S. Stoll, N. C. Camgöz, S. Hadfield, R. Bowden, Sign language production using neural machine translation and generative adversarial networks, in: Proceedings of the 29th British Machine Vision Conference (BMVC 2018), British Machine Vision Association, 2018.

[13] M. P. Lewis, F. Gary, Simons, and charles d. fennig (eds.). 2013. ethnologue: Languages of the world (2015).

[14] F. Wang, C. Li, Z. Zeng, K. Xu, S. Cheng, Y. Liu, S. Sun, Cornerstone network with feature extractor: a metric-based few-shot model for chinese natural sign language, Applied Intelligence 51 (10) (2021) 7139–7150.

[15] G. Shovkoplias, M. Tkachenko, A. Asadulaev, O. Alekseeva, N. Dobrenko, D. Kazantsev, A. Vatian, A. Shalyto, N. Gusarova, Support for communication with deaf and dumb patients via few-shot machine learning, in: Proceedings 14th International Conference on ICT, Society and Human Beings (ICT 2021), the 18th International Conference Web Based Communities and Social Media (WBC 2021), 2021.

[16] S. Ferreira, E. Costa, M. Dahia, J. Rocha, A transformer-based contrastive learning approach for few-shot sign language recognition, arXiv preprint arXiv:2204.02803 (2022).

[17] S. Ravi, M. Suman, P. Kishore, K. Kumar, A. Kumar, et al., Multi modal spatio temporal co-trained cnns with single modal testing on rgb–d based sign language gesture recognition, Journal of Computer Languages 52 (2019) 88–102.

[18] K. M. Lim, A. W. C. Tan, C. P. Lee, S. C. Tan, Isolated sign language recognition using convolutional neural network hand modelling and hand energy image, Multimedia Tools and Applications 78 (14) (2019) 19917–19944.

[19] A. Wadhawan, P. Kumar, Sign language recognition systems: A decade systematic literature review, Archives of Computational Methods in Engineering 28 (3) (2021) 785–813.

[20] S. Aly, W. Aly, Deeparslr: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition, IEEE Access 8 (2020) 83199–83212.

[21] H. Luqman, E.-S. M. El-Alfy, Towards hybrid multimodal manual and non-manual arabic sign language recognition: Marsl database and pilot study, Electronics 10 (14) (2021) 1739.

[22] P. Kumar, P. P. Roy, D. P. Dogra, Independent bayesian classifier combination based sign language recognition using facial expression, Information Sciences 428 (2018) 30–48.

[23] A. Sabyrov, M. Mukushev, V. Kimmelman, Towards real-time sign language interpreting robot: Evaluation of non-manual components on recognition accuracy., in: CVPR Workshops, 2019.

[24] N. C. Camgoz, O. Koller, S. Hadfield, R. Bowden, Sign language transformers: Joint end-to-end sign language recognition and translation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10023–10033.

[25] B. Saunders, N. C. Camgoz, R. Bowden, Progressive transformers for end-to-end sign language production, in: European Conference on Computer Vision, Springer, 2020, pp. 687–705.

[26] W. Tao, M. C. Leu, Z. Yin, American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion, Engineering Applications of Artificial Intelligence 76 (2018) 202–213.

[27] H. Luqman, E.-S. M. El-Alfy, G. M. BinMakhashen, Joint space representation and recognition of sign language fingerspelling using gabor filter and convolutional neural network, Multimedia Tools and Applications 80 (7) (2021) 10213–10234.

[28] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, Y. Fu, Skeleton aware multi-modal sign language recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3413–3423.

[29] J. Huang, W. Zhou, H. Li, W. Li, Sign language recognition using 3d convolutional neural networks, in: 2015 IEEE international conference on multimedia and expo (ICME), IEEE, 2015, pp. 1–6.

[30] O. M. Sincan, H. Y. Keles, Using motion history images with 3d convolutional networks in isolated sign language recognition, IEEE Access 10 (2022) 18608–18618.

[31] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, A. Zisserman, Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues, in: European conference on computer vision, Springer, 2020, pp. 35–53.

[32] L. Momeni, G. Varol, S. Albanie, T. Afouras, A. Zisserman, Watch, read and lookup: learning to spot signs from multiple supervisors, in: Proceedings of the Asian Conference on Computer Vision, 2020.

[33] R. Rastgoo, K. Kiani, S. Escalera, Zs-slr: Zero-shot sign language recognition from rgb-d videos (2021). doi:10.48550/ARXIV.2108.10059.
URL https://arxiv.org/abs/2108.10059

[34] P. Rodríguez, I. Laradji, A. Drouin, A. Lacoste, Embedding propagation: Smoother manifold for few-shot classification, in: European Conference on Computer Vision, Springer, 2020, pp. 121–138.

[35] M. Boháček, M. Hrúz, Sign pose-based transformer for word-level sign language recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 182–191.

[36] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Advances in neural information processing systems 30 (2017).

[37] D. Li, C. Rodriguez, X. Yu, H. 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, 2020, pp. 1459–1469.