Multimodal Continuous Fingerspelling Recognition via Visual Alignment Learning

Katerina Papadimitriou, Gerasimos Potamianos

Department of Electrical & Computer Engineering, University of Thessaly, Volos, Greece

aipapadimitriou@uth.gr, gpotam@ieee.org

Abstract

Continuous fingerspelling recognition from videos is paramount for real-time sign language (SL) interpretation, enhancing accessibility. Despite deep learning progress, challenges persist, especially in signer-independent (SI) scenarios, due to signing variability. To address these, we propose a novel bimodal approach that integrates appearance and skeletal information, focusing solely on the signing hand. Our system relies on two basic modules: (a) a 3D-CNN model capturing spatial features, while adapting to motion variations and (b) a modulated spatio-temporal graph convolutional network (ST-GCN) based on 3D joint-rotation parameterization for skeletal feature modeling. Both modalities are combined with a BiGRU encoder and CTC decoding. To further enhance representation capacity, we introduce an alignment mechanism relying on two auxiliary losses. Through ensemble fusion and language model integration, our method achieves superior performance across three SI fingerspelling datasets.

Index Terms: fingerspelling recognition, 3D-CNN, ST-GCN, language model, alignment module

1. Introduction

Continuous fingerspelling recognition constitutes an essential component of SL processing, being indispensable to the effective communication within the deaf and hard of hearing community. Continuous fingerspelling involves the sequential articulation of alphabet letter signs mainly through intricate hand gestures, conveying lexical units that do not have dedicated signs, like names, technical terms, or foreign words. Despite the recent deep learning advancements in the fields of computer vision and human language technologies, as well as the acquisition of large continuous fingerspelling datasets [1–4], the problem of continuous fingerspelling recognition remains challenging. This primarily arises from the intricacy and resemblance of handshapes, the fast-paced hand motion, the natural inter-signer articulation variability, and the absence of letter level segmentation, all hindering performance of fingerspelling recognition systems. This paper focuses on addressing these issues, developing a robust system capable of accurately recognizing fingerspelling sequences from RGB video.

Fingerspelling recognition has gained considerable attention in recent years, with various approaches being proposed to solve this complex task. Early efforts focused primarily on handcrafted feature extraction techniques, followed by machine learning algorithms such as SVMs and HMMs, as proposed in [5]. In contrast, [6] proposed using HOG and Zernike moment features in conjunction with a deep belief network classifier. In addition, in [7] a scheme based on a hand-tracking device and an SVM classifier is presented. Further, the system in [8] uses LBP histogram features derived from both color and depth data in combination with an SVM classifier. With the advent of deep learning, convolutional neural networks (CNNs) have emerged as a powerful tool for feature learning, enhancing fingerspelling recognition performance. Most recent works commence with hand region segmentation [9–11], employing 2D-CNNs for hand appearance feature extraction, while others employ 2D-CNNs and spatial attention techniques to capture signing information from fingerspelling sequences [2, 12, 13]. Moreover, advances in computer vision techniques, such as pose estimation, have enhanced the accuracy and robustness of fingerspelling recognition systems. In addition, some studies have explored multimodal approaches that integrate skeletal and visual information to improve recognition performance, such as our previous works in [9, 10].

Here, we propose a system based on the signing hand solely, relying on two main modalities: (i) a 3D-CNN model to capture the spatio-temporal dynamics of hand articulation and (ii) an ST-GCN to learn the spatial and motion correlation of the hand skeletal joints (see also Fig. 1). Both models are trained separately and their outputs are combined during inference via an ensemble module, which is also coupled with a language model to further improve system performance. In particular, we present a deep learning-based fingerspelling recognition scheme that commences with the detection of the signing hand and its region segmentation, using the hand skeleton joints derived from the MediaPipe pose estimation framework [14]. Our first contribution is the use of 3D-CNN as a spatio-temporal visual feature learning module in fingerspelling SL recognition. Specifically, our system integrates an appearance modality that is based on the ResNet2+1D network [15], which decouples the spatial and temporal convolutions of the 3D-CNNs, being suitable for learning both intra- and inter-frame hand motion features. To achieve the full potential of the fingerspelling recognition system, we exploit advances in vision-based estimation of human body keypoints and explore their integration into skeleton-based GCNs, which constitutes the second contribution of this paper. Specifically, the modulated ST-GCN module, introduced in our previous work in [16], is applied. Note that graph construction relies on 3D joint-rotation parameterization of the hand skeleton inferred via the PIXIE hand pose regression model [17], which captures the dynamic aspects of hand in both spatial and temporal domains.

For sequence learning, most schemes in the literature rely on RNNs, typically BiLSTMs [18], followed by CTC alignment [1,2,11], while others use attentional encoder-decoders [9, 10, 19]. Here, a 3D-CNN model and a ST-GCN module serve as the spatio-temporal feature learner of each video frame, while a BiGRU encoder [20] learns their temporal relations (see also Fig. 1). Like most sequence-to-sequence prediction problems,

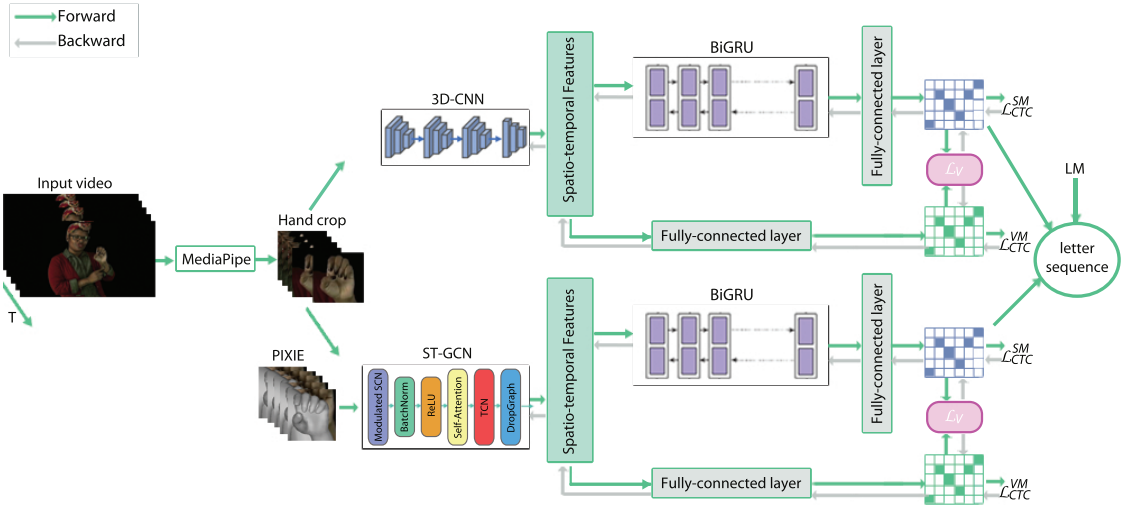

Figure 1: *Architecture of the introduced fingerspelling recognition system that generates letters from a series of signing hand images through a bimodal framework that relies on 3D-CNN based appearance and ST-GCN based skeletal features. A BiGRU-based sequence learning model coupled with CTC decoding is employed. Two auxiliary loss functions are also incorporated during training and a language model (LM) is integrated during inference.*

there is no one-to-one alignment between the input and output sequences, as a portion of frames can be associated with a fingerspelled letter. To this end, we integrate an alignment module, combining the CTC loss with a knowledge distillation loss [21, 22] and a visual module loss function, which aligns the probability distributions generated by the sequence learning model and the visual module, enhancing the representation capacity of our model. This comprises the third innovation of this paper.

To summarize, the main contributions of this approach lie in the design of a novel bimodal continuous fingerspelling recognition system that integrates 3D-CNN based appearance and modulated ST-GCN based skeletal feature modeling. In addition, we integrate a visual alignment loss function and a knowledge distillation loss during training, enhancing the representation capacity of both the visual and sequential modules. To date, none of the above have been investigated in conjunction with continuous fingerspelling recognition in the literature.

We evaluate our introduced approach on the ChicagoFSWild and the ChicagoFSWild+ datasets, as well as a Greek fingerspelling corpus, and we provide in-depth ablations that highlight our innovations. Comparing our method to state-of-the-art fingerspelling recognition systems, our system outperforms the state-of-the-art on both American SL sets by a significant amount (7.01% and 7.37% absolute reduction in error rate in the first and second corpus, respectively). We also report the first-ever results on the Greek fingerspelling corpus.

2. Methodology

Continuous fingerspelling recognition involves the task of predicting a sequence of letters $\mathbf{l} = (l_1, l_2, ..., l_N)$ from a sequence of T image frames $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$. To address this, we propose the framework illustrated in Fig. 1. As it may be observed, our fingerspelling recognition system comprises: (i) a preprocessing phase, which focuses on signing hand detection and segmentation; (ii) an appearance modality relying on a 3D-CNN based visual module followed by a BiGRU encoder; and (iii) a skeletal based modality relying on an ST-GCN module a BiGRU encoder. An alignment module, relying on two auxiliary loss functions, is also integrated. Further details follow.

2.1. Preprocessing

Since the hand typically serves as the primary articulation in fingerspelling signing, our model relies on input from the signing hand region only. Thus, an essential stage of our system lies in effective detection and tracking of the manual articulation. For that purpose, we adopt the MediaPipe human pose detector [14], exploiting the 21 hand keypoints to delineate and segment the hand region. Afterwards, the x and y landmarks of the hands, whose values fall within the range of $[0.0, 1.0]$, are normalized to the image plane, using the image width and height.

Given that the non-signing hand typically exhibits limited motion, we leverage the hand trajectory to distinguish it from the signing hand. In particular, we examine the stability of the wrist joints across successive frames, designating the hand as non-signing if its wrist joint remains "stable" for more than 10 consecutive frames, with the Euclidean distance between joint coordinates of hand position transitions (adjacent frames) being less than 8. Subsequently, we determine the maximum and minimum values of the corresponding x and y landmarks of the signing hand to facilitate hand region cropping. Note that in case of MediaPipe failure, any missing keypoints are filled with the previously detected ones.

2.2. Appearance Features

To capture visual feature representations from the previously generated hand image sequences, we utilize a 3D-CNN architecture. As depicted in Fig. 1, we adopt the ResNet2+1D network [15] that decomposes the 3D-CNNs into spatial convolutions for frame-wise feature regression and temporal convolutions to capture short-term dynamics of hand posture and motion across adjacent frames. Our model consists of five (2+1)D convolutional blocks, integrating both spatial and temporal convolutions, followed by a global average pooling layer that operates across both spatial and temporal dimensions.

In particular, to process the sequence of hand images with a length of T, we first rescale them to the appropriate input layer size of the network, i.e., 112×112 pixels. Subsequently, we

separate the initial sequence into T' subsequences via a sliding window of 8 consecutive frames with a stride of $s = 4$, ensuring proper pre-padding at the video's end, where $T' = \lceil T/s \rceil$. Notably, there is a 4-frame overlap (half-clip overlap) between subsequences. Each hand image subsequence is then fed into the ResNet2+1D network, yielding spatio-temporal features. The model is pre-trained on the Kinetics dataset [23]. To further improve model performance, we pretrain our model on the Chinese SL dataset [24]. Feature maps are derived from the last pooling layer, resulting in 512-dimensional features.

2.3. Skeletal Features

To improve the efficacy of our system, we integrate spatio-temporal representations derived from signing hand skeletal data. For this purpose, we employ an ST-GCN architecture [16], comprising a modulated GCN that is followed by a temporal convolution, enriched with an attention mechanism. As depicted in Fig. 1, the model commences with graph construction based on the PIXIE 3D joint-rotation parameterization of the signing hand pose. Specifically, the PIXIE model [17], utilizing a moderator to estimate the 3D hand and shape parameterization, yields 15 joints with 6 degrees of freedom for the signing hand. The resulting graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is propagated into the GCN unit, where $\mathcal{V}$ denotes the set of nodes corresponding to J hand skeletal joints, and $\mathcal{E}$ represents intra-skeleton structure edges. Thus, each node i is associated with a D-dimensional feature vector $\mathbf{q}_i \in \mathbb{R}^D$ corresponding to the 3D joint-rotation parameterization inferred from the PIXIE model.

As already mentioned, we deploy a modulated graph convolution layer [25], which relies on weight modulation and affinity modulation. In the case of weight modulation, a unique learnable weight modulation matrix $\mathbf{L} \in \mathbb{R}^{D' \times J}$ is incorporated into the graph convolution function for each node i, with the forward propagation rule of the GCN layer being formulated as follows:

$$\mathbf{Q}_{out} = \sigma((\mathbf{L} \odot (\mathbf{W}\mathbf{Q}_{in}))\hat{\mathbf{A}}),$$

where $\odot$ denotes element-wise multiplication, $\mathbf{Q}_{in} \in \mathbb{R}^{D \times J}$ represents the input feature vector, $\mathbf{Q}_{out} \in \mathbb{R}^{D' \times J}$ denotes the updated feature vector, $\sigma()$ indicates the activation function, $\mathbf{W} \in \mathbb{R}^{D' \times D}$ represents the learnable weight matrix, and $\hat{\mathbf{A}}$ is the symmetrically normalized affinity matrix. Note that the adjacency matrix $\mathbf{A} \in \{0, 1\}^{J \times J}$ represents the edges, where a value of 1 indicates a direct link between a pair of joints, while 0 indicates no direct connection. Through the affinity modulation technique, a learnable mask $\mathbf{B} \in \mathbb{R}^{J \times J}$ is added to matrix $\mathbf{A}$, resulting in $\mathbf{A}' = \mathbf{A} + \mathbf{B}$.

The GCN unit is coupled with self-attention modules, including spatial, temporal, and channel attention. Moreover, a temporal convolution, which operates on the temporal neighborhood of nodes, is incorporated to learn the relational patterns between successive frames. To mitigate overfitting, a DropGraph module [26] is applied. Our system incorporates ten such ST-GCN units, coupled with a global average pooling layer for both spatial and temporal domains.

2.4. Sequence Learning

To capture long-term dependencies, both appearance and skeletal features are individually propagated into a 4-layer Bi-GRU [20] model, where each layer has a hidden state dimensionality of 512. Subsequently, a dense fully-connected layer coupled with a softmax activation function are applied to generate the predicted probability scores for each letter label. The CTC loss function $\mathcal{L}_{CTC}^{SM}$ is then employed, aligning the probability distribution of the sequence learning model with the sequence letter labels. In addition, for each modality two auxiliary losses are incorporated. In particular, the appearance features extracted from the ResNet2+1D network, as well as the skeletal features inferred from the ST-GCN model are individually fed into a fully-connected layer paired with softmax, generating posteriors. Moreover, a CTC loss function ($\mathcal{L}_{CTC}^{VM}$) is employed for aligning visual features with the target letter sequence. Subsequently, the posteriors of the sequence learning model $\mathbf{D}_{SM}$ are aligned with those derived from the visual models $\mathbf{D}_{VM}$ through the KL-divergence loss function, formulated as: $\mathcal{L}V = \text{KL}(\text{softmax}(\mathbf{D}_{VM}), \text{softmax}(\mathbf{D}_{SM}))$. The KL-divergence loss function facilitates alignment between short-term visual predictions and long-term context predictions. Notably, during training, the three loss functions are linearly combined as: $\mathcal{L}_T = \mathcal{L}_{CTC}^{SM} + \mathcal{L}_{CTC}^{VM} + 0.5\mathcal{L}V$.

2.5. Ensemble Module

The two modalities are trained separately, while during inference they are fused using an ensemble module. Specifically, the posteriors returned from the final fully-connected layers of each modality are fused appropriately. For fusion, distinct weights are assigned to each modality based on their individual performance during validation, and, subsequently, they are summed up to generate the final probability scores. In addition, we incorporate a language model, which provides a probability estimate for each potential succeeding letter given the preceding ones. Note that for language model training, we employ a one-layer LSTM network with 512 hidden units. In particular, the posteriors of each modality and the language model probability are weighted and summed as: $p_{\text{fused}} = 1.0p_{\text{app}} + 0.9p_{\text{skel}} + 0.6p_{\text{LM}}$.

3. Experimental Framework

The performance of the introduced model is evaluated on three publicly available datasets: (a) the *Chicago fingerspelling in the wild dataset (ChicagoFSWild)* [1] and (b) the *Chicago fingerspelling in the wild dataset+ (ChicagoFSWild+)* [2] in American SL, as well as (c) a *Greek SL (FGSL) corpus* [4]. Specifically, for the first two datasets, we employ their official SI splits with no signer overlap between the different sets. Particularly, in the instance of the ChicagoFSWild dataset 5,455 videos are employed for training (87 subjects), 981 for validation (37 signers), and 868 videos for test (36 informants). On the other hand, the ChicagoFSWild+ dataset contains 50,402 training videos (216 signers), 3,115 validation videos (22 signers), and 1,715 test sequences (22 signers). In addition, for our experiments on the FGSL corpus, we deploy the official SI split, where a 7-fold cross-validation is used, with each fold containing training data (80% of the fold) and validation data (20% of the fold) from 18 signers, whereas testing is performed on the remaining 3 signers. The process repeats over all 7 folds to cover all signers.

Regarding system training, both modalities are trained for 50 epochs with a mini-batch size fixed to 2. Training is conducted using the Adam optimizer [27] with an initial learning rate of 0.0001, decreased by a factor of 0.5 in each iteration. Training data augmentation is applied through random cropping and horizontal flipping. Regarding the language model, training is conducted via the Adam optimizer with initial learning rate of 0.001, decayed by a factor of 0.1, while the model is trained employing for each dataset the training and the validation an-

Table 1: *Ablation study concerning the auxiliary loss functions, the language model (LM), and the different modalities. The evaluation is conducted in terms of letter accuracy (LAcc, %) on all three datasets.*

Backbone Models	$\mathcal{L}_{CTC}^{SM}$	$\mathcal{L}_{CTC}^{VM}$	$\mathcal{L}_{V}$	LM	ChicagoFSWild	ChicagoFSWild+	FGSL
3D-CNN & BiGRU (Appearance)	✓				47.27	66.58	82.90
	✓	✓			60.32	71.05	92.05
	✓	✓	✓		61.60	71.30	92.30
	✓	✓	✓	✓	62.57	72.05	92.49
ST-GCN & BiGRU (Skeleton)	✓				48.15	66.28	81.23
	✓	✓			54.37	69.12	89.70
	✓	✓	✓		56.88	69.54	89.98
	✓	✓	✓	✓	58.02	70.01	90.13
Appearance & Skeleton	✓				51.13	70.50	86.77
	✓	✓			63.24	72.23	91.94
	✓	✓	✓		64.08	72.64	92.89
	✓	✓	✓	✓	**64.85**	**73.57**	**93.14**

notations. The system is implemented in PyTorch [28], and the experiments are carried out on an Nvidia RTX 3090 GPU.

4. Experimental Results

Here, we present the experimental results pertaining to the evaluation of the proposed methodology, conducted quantitatively on the datasets outlined in Section 3 in terms of letter accuracy (LAcc, %). Initially, we conduct a comparative analysis between the proposed and its various adaptations, highlighting the advantages of integrating both appearance and skeletal streams, as well as the enhancements achieved through the incorporation of the auxiliary losses and the language model. Comparing the fourth row entries of Table 1 (appearance stream alone) to the corresponding entries of the eighth row (skeletal stream only) shows that the 3D-CNN based appearance module provides superior performance over the skeletal ST-GCN. Nevertheless, both modules perform well, and their fusion improves performance further (last row), i.e. 2.28% absolute reduction on ChicagoFSWild, 1.52% absolute reduction on ChicagoFSWild+, and 0.65% on FGSL (over appearance only). In addition, the integration of both auxiliary losses into our model benefits system performance. It is also worth noting that incorporating the language model leads to further gains in LAcc. To emphasize this, in Fig. 2 we illustrate the sequence prediction results obtained by the introduced fingerspelling recognition model, as well as several variations of it, when applied to a sample video of the FGSL corpus. Note that our model turns out superior to the considered alternative relying exclusively on a 2D-CNN (ResNet-18) image feature learner as evaluated on the ChicagoFSWild corpus, yielding LAcc of 55.78%. Further, we assess the performance of the modulated ST-GCN using the 3D skeletal feature representations derived from the MediaPipe regression model for graph construction on the ChicagoFSWild corpus, resulting in lower recognition accuracy (53.46% vs. 58.02%). Finally, we evaluate model performance on the same corpus when substituting the BiGRU with a BiLSTM encoder, resulting in lower LAcc (62.34% vs. 62.57%).

REF: E L A F I
HYP (w $\mathcal{L}_{CTC}^{SM}$ only): A T E F I
HYP (w/o $\mathcal{L}_{CTC}^{VM}$): E P A K I
HYP: (w/o $\mathcal{L}_{V}$): E P E F I
HYP: (w/o LM): E L E F I
HYP: E L A F I

Figure 2: *Predictions generated by the proposed, as well as several variations of it against the reference (ground truth), applied to frame sequences of the FGSL corpus [4]. Wrong letter predictions are colored in red and correct predictions in green.*

Next, Table 2 reports the evaluation comparison of the introduced against state-of-the-art models on the ChicagoFSWild fingerspelling dataset. It can be observed that the proposed outperforms the state-of-the-art scheme that relies on optical flow based spatial attention. Further, our approach exhibits significant gains over the state-of-the-art model on the ChicagoFSWild+ fingerspelling dataset (Table 3), which is based on visual attention to informative frame regions. Finally, the performance on the FGSL dataset is significantly better than on the other datasets, due to the studio-like recording setup of the former vs. the in-the-wild nature of the latter. Our whole system (51M parameters) takes on average 85ms per frame during inference.

5. Conclusions

In this work, we presented a novel deep learning approach addressing continuous fingerspelling recognition. Integrating a ResNet2+1D model for spatio-temporal appearance feature learning and a modulated ST-GCN relying on 3D joint-rotation parameterization features for robust skeletal feature modeling, our system achieves both prevalent spatial modeling potential and motion-aware modeling adaptability. Sequence learning relies on a BiGRU encoder, aligned with the target letter sequence via the CTC loss function. We also integrated two auxiliary loss functions, enhancing our system representation capacity. During inference, modality fusion and language model integration leads to superior performance. Our model outperformed the current state-of-the-art on two American sets, whereas we reported the first-ever results on a Greek corpus.

Table 2: *Letter accuracy (LAcc, %) comparison of state-of-the-art on the ChicagoFSWild dataset.*

Model	Feature streams	LAcc (%) ↑
R-CNN-Att [2]	Full Frame	45.10
CNN-Att [10]	Hand/Mouth & 2D/3D Skel.	47.93
FG-Transformer [12]	Full Frame	48.36
Siam-LSTM [29]	Full Frame	48.00
Iterative-LM [30]	Full Frame	49.60
Flow-ResNet-BiLSTM [31]	Full Frame	57.84
Ours	Hand	**64.85**

Table 3: *Letter accuracy (LAcc, %) comparison of state-of-the-art on the ChicagoFSWild+ dataset.*

Model	Feature streams	LAcc (%) ↑
R-CNN-Att [2]	Full Frame	46.70
RNN-Att [13]	Full Frame	66.20
Ours	Hand	**73.57**

6. Acknowledgements

This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "1st Call for H.F.R.I. Research Projects to support Faculty Members & Researchers and the procurement of high-cost research equipment grant" (Project "SL-ReDu", Project Number HFRI-FM17-2456).

7. References

[1] B. Shi, A. M. D. Rio, J. Keane, J. Michaux, D. Brentari, G. Shakhnarovich, and K. Livescu, "American sign language fingerspelling recognition in the wild," in *Proc. of the IEEE Spoken Language Technology*, 2018, pp. 145–152.

[2] B. Shi, A. M. D. Rio, J. Keane, D. Brentari, G. Shakhnarovich, and K. Livescu, "Fingerspelling recognition in the wild with iterative visual attention," in *Proc. of the IEEE International Conference on Computer Vision*, 2019, pp. 5399–5408.

[3] K. Prajwal, H. Bull, L. Momeni, S. Albanie, G. Varol, and A. Zisserman, "Weakly-supervised fingerspelling recognition in British sign language videos," in *Proc. of the British Machine Vision Conference (BMVC)*, 2022.

[4] K. Papadimitriou, G. Sapountzaki, K. Vasilaki, E. Efthimiou, S.-E. Fotinea, and G. Potamianos, "SL-REDU GSL: A large Greek sign language recognition corpus," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing Workshop on Sign Language Translation and Avatar Technology (ICASSPW-SLTAT)*, 2023, pp. 1–5.

[5] S. Upendran and A. Thamizharasi, "American sign language interpreter system for deaf and dumb individuals," in *Proc. of the International Conference on Control, Instrumentation, Communication and Computational Technologies*, 2014, pp. 1477–1481.

[6] Y. Hu, H. F. Zhao, and Z. G. Wang, "Sign language fingerspelling recognition using depth information and deep belief networks," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 06, p. 1850018, 2017.

[7] L. Quesada, G. López, and L. Guerrero, "Automatic recognition of the American sign language fingerspelling alphabet to assist people living with speech or hearing impairments," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 4, pp. 625–635, 2017.

[8] C. S. Weerasekera, M. H. Jaward, and N. Kamrani, "Robust ASL fingerspelling recognition using local binary patterns and geometric features," in *Proc. of the International Conference on Digital Image Computing: Techniques and Applications*, 2013, pp. 1–8.

[9] K. Papadimitriou and G. Potamianos, "Multimodal sign language recognition via temporal deformable convolutional sequence learning," in *Proc. of the Interspeech*, 2020, pp. 2752–2756.

[10] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, "Exploiting 3D hand pose estimation in deep learning-based sign language recognition from RGB videos," in *Proc. of the ECCV (SLRTP) 2020 Workshops*, 2020, pp. 249–263.

[11] B. Shi, D. Brentari, G. Shakhnarovich, and K. Livescu, "Fingerspelling detection in American sign language," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4164–4173.

[12] K. Gajurel, C. Zhong, and G. Wang, "A fine-grained visual attention approach for fingerspelling recognition in the wild," *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3266–3271, 2021.

[13] S. KruthiventiSS, G. Jose, N. Tandon, R. Biswal, and A. Kumar, "Fingerspelling recognition in the wild with fixed-query based visual attention," in *Proc. of the ACM International Conference on Multimedia*, 2021, pp. 4362–4370.

[14] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for perceiving and processing reality," in *Proc. of the Workshop on Computer Vision for AR/VR at IEEE CVPR*, 2019.

[15] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.

[16] K. Papadimitriou and G. Potamianos, "Sign language recognition via deformable 3D convolutions and modulated graph convolutional networks," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[17] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, "Collaborative regression of expressive bodies using moderation," in *Proc. of the 3DV*, 2021, pp. 792–804.

[18] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[19] B. Shi and K. Livescu, "Multitask training with unlabeled data for end-to-end sign language fingerspelling recognition," *CoRR*, vol. abs/1710.03255, 2017.

[20] C. Yu, L. Tianrui, J. Zhen, and Y. Chengfeng, "BGRU: A new method of Chinese text sentiment analysis," *Journal of Physics: Conference Series*, vol. 13, no. 06, pp. 973–981, 2019.

[21] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *ArXiv*, vol. abs/1503.02531, 2015.

[22] K. Papadimitriou and G. Potamianos, "Multimodal locally enhanced Transformer for continuous sign language recognition," in *Proc. of the INTERSPEECH 2023*, 2023, pp. 1513–1517.

[23] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about Kinetics-600," *CoRR*, vol. arXiv:1808.01340, 2018.

[24] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese sign language recognition with adaptive HMM," in *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[25] Z. Zou and W. Tang, "Modulated graph convolutional network for 3D human pose estimation," in *Proc. of the ICCV*, 2021, pp. 11 457–11 467.

[26] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling GCN with DropGraph module for skeleton-based action recognition," in *Proc. of the ECCV*, 2020, pp. 536–553.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. arXiv:1412.6980, 2014.

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. of the NIPS-W*, 2017.

[29] P. Pannattee, W. Kumwilaisak, C. Hansakunbuntheung, and N. Thatphithakkul, "Novel American sign language fingerspelling recognition in the wild with weakly supervised learning and feature embedding," in *Proc. of the International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2021, pp. 291–294.

[30] W. Kumwilaisak, P. Pannattee, C. Hansakunbuntheung, and N. Thatphithakkul, "American sign language fingerspelling recognition in the wild with iterative language model construction," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[31] A. E. Kabade, P. Desai, C. Sujatha, and G. Shankar, "American sign language fingerspelling recognition using attention model," in *Proc. of the IEEE International Conference for Convergence in Technology (I2CT)*, 2023, pp. 1–6.