SPATIO-TEMPORAL GRAPH CONVOLUTIONAL NETWORKS FOR CONTINUOUS SIGN LANGUAGE RECOGNITION

Maria Parelli[1], *Katerina Papadimitriou*[2], *Gerasimos Potamianos*[2], *Georgios Pavlakos*[3], *Petros Maragos*[1]

[1] School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[2] Department of Electrical & Computer Engineering, University of Thessaly, Volos, Greece
[3] Electrical Engineering & Computer Sciences, University of California, Berkeley, CA, U.S.A.

ABSTRACT

We address the challenging problem of continuous sign language recognition (CSLR) from RGB videos, proposing a novel deep-learning framework that employs spatio-temporal graph convolutional networks (ST-GCNs), which operate on multiple, appropriately fused feature streams, capturing the signer's pose, shape, appearance, and motion information. In addition to introducing such networks to the continuous recognition problem, our model's novelty lies on: (i) the feature streams considered and their blending into three ST-GCN modules; (ii) the combination of such modules with bi-directional long short-term memory networks, thus capturing both short-term embedded signing dynamics and long-range feature dependencies; and (iii) the fusion scheme, where the resulting modules operate in parallel, their posteriors aligned via a guiding connectionist temporal classification method, and fused for sign gloss prediction. Notably, concerning (i), in addition to traditional CSLR features, we investigate the utility of 3D human pose and shape parameterization via the "ExPose" approach, as well as 3D skeletal joint information that is regressed from detected 2D joints. We evaluate the proposed system on two well-known CSLR benchmarks, conducting extensive ablations on its modules. We achieve the new state-of-the-art on one of the two datasets, while reaching very competitive performance on the other.

Index Terms— continuous sign language recognition, spatio-temporal graph convolutional networks, BiLSTM, CTC, ExPose

1. INTRODUCTION

Sign language (SL) constitutes a non-vocal form of language that encapsulates numerous manual and non-manual articulation cues, allowing communication for the deaf and hard-of-hearing. Its automatic recognition from videos has been attracting considerable attention in recent years, but nevertheless remains a challenging problem, due to the multitude, complexity, and strong spatio-temporal correlation of the SL articulators participating in signing, the natural inter-signer variability, and various difficulties of in-the-wild video processing. The problem becomes even more challenging in the continuous SL recognition (CSLR) case that constitutes the focus of this paper, where the goal is to predict a sequence of SL glosses from the signing video without prior knowledge of gloss-level segmentation [1–3].

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 2456).

To tackle CSLR, most works in the literature employ visual features that capture signing appearance information via 2D convolutional neural networks (CNNs) [2, 4], or 3D-CNNs to better model spatio-temporal dependencies [5], as well as 2D human-pose (skeletal) features that are derived from the OpenPose library [2] or deconvolutional neural networks [1]. Such features are used in conjunction with sequence learning techniques, typically recurrent neural networks (RNNs) [1, 3, 6], such as the bi-directional long short-term memory (BiLSTM) model [7], coupled with connectionist temporal classification (CTC) decoding [8]. In addition, many works treat CSLR as a neural machine translation task, adopting attention-based encoder-decoder models [2, 5, 6].

However, SL entails rich spatio-temporal structures that CNNs and RNNs in their native form do not capture well. Thus, some recent efforts in the literature have proposed the use of graph convolutional networks (GCNs), due to their adaptability to signing dynamics along both the spatial and temporal dimensions. Specifically, in [9], spatio-temporal GCNs (ST-GCNs) generate explicit feature maps from sequences of human skeleton graphs. Similarly, in [10], a scheme involving sequences of skeletal data with three different GCN modules and attention mechanisms is proposed. In addition, in [11], a ST-GCN architecture is used, in an attempt to unify the spatial and temporal features. Finally, in one of the most recent advances [12], an SL recognizer learns from multiple modalities using a GCN-based model. The various streams are trained separately, and their outputs are combined via an ensemble module that leverages the last fully-connected layer output. Note that all aforementioned approaches have only been applied to isolated SL recognition.

Motivated by the above, we propose a novel CSLR approach that relies on ST-GCNs, acquiring both spatial and temporal patterns from the signing videos. Compared to the earlier mentioned GCN-based methods for SL recognition, our work concerns the use of per-vertex feature vectors, generated by attaching visual latent representations to the skeleton graphs via a ST-GCN based ensemble module. In addition, we introduce an encoding model that relies on the combination of ST-GCNs with BiLSTMs, being (to our knowledge) the first to combine these two models to capture both short-term and long-term signing dynamics. Moreover, we introduce a fusion scheme, where three ST-GCN / BiLSTMs, operating in parallel on different feature streams, are aligned for gloss prediction via a guided CTC approach. We depict our proposed model architecture in Fig. 1, providing a detailed description in Section 3.

Further to the above, we add two novelties regarding CSLR visual features: The first constitutes the use of the "ExPose" approach [13] that extracts 3D pose and shape directly from RGB images of the signer. The second concerns the 3D body/hand skeleton regression, based on the corresponding 2D joints detected by Open-

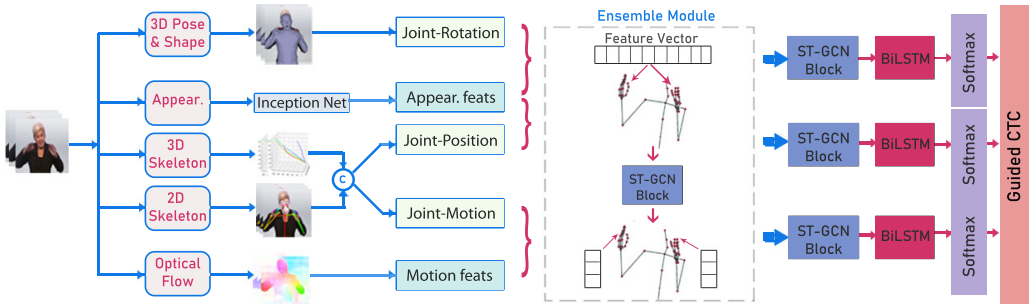

Fig. 1. Proposed CSLR model architecture: Groups of streams are fused via an ensemble module, producing three feature vectors that pass through a series of ST-GCN layers and a BiLSTM encoder followed by a linear softmax layer. The predictions are fused using guided CTC.

Pose [14], via a multi-layer neural network, extending our earlier similar work that was limited to the hand joints alone [15]. These features are considered together with traditional appearance and motion features, as well as 2D skeletal joints, as discussed in Section 2.

We evaluate our proposed system on two popular CSLR benchmarks, namely the "RWTH-PHOENIX Weather 2014T" corpus [16] and the "Chinese SLR dataset" [6], providing extensive ablations that highlight our contributions. Comparing our system against state-of-the-art CSLR methods on these corpora, we achieve superior performance on the second dataset, while reaching very competitive performance on the first. Details are provided in Section 4.

2. VISUAL FEATURES FOR CSLR

2.1. 3D Human Body Pose and Shape Representation

SL articulation occurs in the 3D space with numerous upper-body articulators contributing to sign formation. To model this process, we employ "ExPose" [13] that captures 3D human body shape, pose, and facial expression, operating directly on image pixels, being capable of reconstructing full expressive 3D humans from RGB images without relying on intermediate features (see also Fig. 2(b)).

Specifically, "ExPose" receives as input a bounding of the human body, which, after down-scaling to the network compatible resolution, is fed to a full body neural network similar to [17], yielding a rough body pose estimate. To obtain more accurate representations of the hands and face that are often hard to estimate in low-resolution images, "ExPose" approaches the reconstruction of the body, hands, and face separately, using part-specific models. In particular, the recovered pose joints from the body network are projected on the original image, and are subsequently employed for hands and face localization, as well as the generation of high-resolution image patches for each region-of-interest. Then, these patches are fed to hand- and face-specific networks that are pre-trained on high-quality hand and face images, refining the hand and face parameters.

The "ExPose" algorithm parameterizes face, body, and hands shape, as well as facial expressions using 10 coefficients, while for the body pose it extracts 53 joints (22 body-pose joints, 15 joints per hand, and 1 for the jaw), with rotation representation dimension equal to 6. In total, it yields a 338-dimensional (dim) feature vector.

2.2. 2D Human Pose Detection

To detect and track the 2D SL articulation, we employ the OpenPose library [14], which infers 2D human skeletal data from monocular images. In particular, OpenPose takes as input single images and produces human skeletal joint locations in the 2D image pixel coordinate system (see also Fig. 2(a)). OpenPose estimates in total 137 human skeletal keypoints, including 70 facial landmarks, 25 body-pose points, and 21 hand-pose joints for each hand. Here, we disregard 80 skeletal joints, i.e. all facial ones and 10 body-pose keypoints corresponding to the lower limbs area, thus yielding a 114-dim feature vector. For scale invariance, we normalize the joints by treating the image frame as a local coordinate system with the neck being its origin and the distance between the left and right shoulders set to unity. Note that, for frames where OpenPose fails, the missing features are substituted by the previous existing ones.

2.3. 3D Human Body and Hand Skeleton Regression

Since knowledge of the trajectory of the arm joints in the 3D space can also yield meaningful information, we extend our earlier work [15] to obtain 3D human body joint keypoints, by "lifting" 2D skeletal joint locations to the 3D space. More precisely, we employ the extracted 2D skeleton of the body (as discussed in Section 2.2), and we retrain a model similar to [19], so that it can "lift" upper-body 2D pose to the 3D space. The model input is a set of 17 2D pose keypoints, and the output is an estimation of the 3D upper body pose, with the neck joint being the origin in both 2D and 3D coordinate systems, so as to ensure translation invariance. For training the model, we use the Human 3.6 M dataset [20, 21], which contains accurate 3D human joint positions. Since in the majority of SL corpora the lower body parts are rarely visible, during training we set the occluded keypoints to zero, so that the model can adapt to upper body poses. In the same manner, we generate the coordinates of hand joints in the 3D space, using the 21 2D keypoints of each

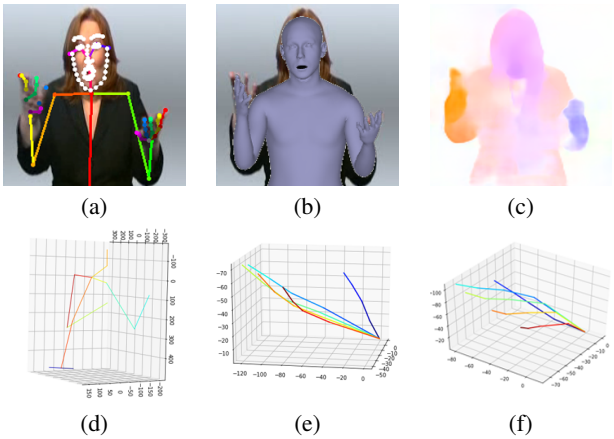

Fig. 2. (a) Sample frame of the RWTH-PHOENIX Weather 2014T corpus [16] with super-imposed 2D skeletal joints obtained by OpenPose [14]; (b) 3D body reconstruction via the ExPose regression model [13]; (c) Motion informative image derived from the SpyNet optical flow model [18]; and (d)-(f) 3D body, left, and right hand skeleton representations obtained by 2D-to-3D regression [19].

hand (derived from OpenPose), following our work [15]. This yields 177-dim features, 126 of which correspond to the hand 3D joints and 51 to the body 3D keypoints (see also Fig. 2(d)-(f)).

2.4. Appearance and Optical Flow Features

In addition to the pose and shape features, we also consider an appearance representation based on the Inception Net [22]. In particular, after rescaling to a 299×299-pixel resolution, we feed the entire frame to the Inception Net, yielding 1024-dim features.

Further, an additional visual representation of interest is the optical flow, as it effectively captures motion information of the numerous SL articulators. To this end, we utilize the SpyNet model [18] for motion informative image generation (see Fig. 2(c)). To learn motion representations, we feed such images (after appropriate rescaling to 227×227 pixels) to a 2D-CNN model that follows the AlexNet architecture [23] and is pre-trained on the ImageNet corpus [24]. This process also yields 1024-dim features.

3. CSLR SEQUENCE LEARNING MODEL

3.1. Recognition Pipeline

Inspired by [25], we address the challenging problem of CSLR from videos by applying a skeleton-based action recognition method that relies on ST-GCNs. Such models operate on a sequence of skeleton graphs, where each node corresponds to a joint in the human body and hands. Formally, an undirected graph $G = (V, E)$ is constructed with the node set $V = \{v_{it} | \, i = 1,...,N, t = 1,...,T\}$, which contains all human joints, i, in a skeleton sequence. The edge set is composed of the intra-skeleton edges $E = \{(u_{it}, u_{jt}) | (i, j) \in H\}$, where H represents the set of edges that follow the structure of the human body and inter-frame edges, which depict the trajectory of a joint in the time domain. Each ST-GCN unit comprises a graph convolution, followed by a temporal convolution. Both are complemented with batch normalization and a ReLU layer.

The ST-GCN encoder can effectively capture short-term motion. However, modeling long-term motion dynamics is a vital aspect in CSLR. Thus, we leverage the power of LSTMs that can process long-range dependencies and model state transitions. We implement a BiLSTM encoder, which allows us to exploit future context, as well as previous one, at the same time. Our architecture consists of four stacked ST-GCN blocks with 256 channels and a temporal kernel size of 5, followed by a global average pooling layer. The output is a $D \times T$ vector, constituting a latent intermediate representation for each frame, with D set to 256. The ST-GCN output is fed to a 2-layer BiLSTM encoder with 256 hidden units. The final output passes through a fully-connected softmax layer to produce the prediction. Inference is performed via a CTC beam search decoder with beam size equal to 5.

3.2. Ensemble Module

Inspired by [26], we extend graph convolutions by effectively attaching visual latent representations to each skeletal joint. The modified ST-GCN network uses the skeletal coordinates of each joint, along with the input frame features (appearance or optical flow), and at each layer it aggregates information from the spatio-temporal neighborhood of each node to obtain a per-vertex feature representation (see also Fig. 1). Motivated by the success of complementary skeleton representations fusion in many computer vision tasks, we apply multi-stream ST-GCN and BiLSTM-based encoding on different combinations of the input feature streams. More precisely, we generate three feature maps:

(A) In the first, $v_{it}^{(JP)} = (x_{it}^{(2D)}, y_{it}^{(2D)}, x_{it}^{(3D)}, y_{it}^{(3D)}, z_{it}^{(3D)})$, i.e. each graph/skeleton vertex is represented by its 2D/3D skeletal-joint positions. Appearance features are then embedded.

(B) In the second, $v_{it}^{(JM)} = v_{it+1}^{(JP)} - v_{it}^{(JP)}$, i.e. each node is represented by its 2D and 3D joint-motion vector. Optical flow features are then embedded to it.

(C) In the last, the 3D joint-rotation parameterization of "ExPose" is embedded on each node/joint and combined with shape and expression coefficients, as well as appearance features.

The feature streams are then fed to three ST-GCN/BiLSTM/CTC models, and their decoding scores are added using the posterior fusion scheme to obtain the final prediction, as detailed next.

3.3. Posterior Fusion Scheme

CTC models tend to emit spiky posterior distributions, where most activations are dominated by high-confidence blanks. Consequently, different CTC-LSTM models suffer from non-aligned spike timings, which renders posterior fusion of softmax scores ineffective. To address this, we adopt a similar approach to [27], where a technique is proposed to explicitly guide the CTC spike timings of speech recognition models to be aligned to those of a pre-trained CTC model (the guiding model). More specifically, during CTC model training, we add a loss term that guides the spikes from the model being trained to occur at the same time as those from the guiding model. During training, the posteriors for each time index predicted by the guiding model are converted to a mask $M(X)$ by setting 1 at the output symbol with the highest posterior, 0 at other symbols, and the blank symbol at each time index. During training, the posterior probabilities $P(X)$ are predicted by the guided model. Through an element-wise multiplication of mask M and the posteriors, we obtain the masked posteriors $\hat{P}(X) = M(X) \odot P(X)$. We synchronize the spike timings of the guided CTC model to those of the guiding one, by maximizing the summation of masked posteriors.

4. EVALUATION

4.1. Datasets and Experimental Framework

We consider two popular CSLR benchmarks in our evaluation:

- The *RWTH-PHOENIX Weather 2014T* dataset (PH2014T) [16] that contains German SL videos of broadcast weather forecasts by 9 signers (6F, 3M), recorded at low frame resolution (210×260 pixels) and a 25 Hz frame rate. The corpus includes 8,257 German SL sequences with a 1,066-gloss vocabulary. In our experiments, we use the official multi-signer split, comprising 7,096 training videos, 519 validation, and 642 testing ones.
- The *Chinese SLR dataset (CSL)* [6], containing studio-quality video of daily-life communication in Chinese SL, recorded by a Microsoft Kinect at high frame resolution (1280×720 pixels) and a 30 Hz rate. The corpus consists of 100 signing sentences (178-gloss vocabulary), expressed by 50 signers (25F, 25M) and performed by each signer 5 consecutive times, yielding $25k$ clips. Here, we adopt the official signer-independent setup (Split I), comprising $20k$ training clips (40 signers) and $5k$ testing ones (10 signers). To avoid tuning on test data, we allocate $5k$ training clips for validation.

It should be noted that the second benchmark corresponds to an easier CSLR task, due to the superior video quality, smaller gloss vocabulary, fixed language content, and larger number of signers.

Table 1. Comparison of our proposed model to the literature on the RWTH-PHOENIX Weather 2014T dataset (PH2014T, left) and the Chinese SLR corpus (CSL, right). Systems are listed in decreasing gloss error rate (%). The following notation is used in the feature stream column: Appearance features based on full frame (FF), hands (H), mouth region (M), and face (F). "Glosses" refers to embeddings.

Model	Feature streams	PH2014T	Model	Feature streams	CSL
SFD-SGS-SFL [4]	FF + Glosses	26.10	LS-HAN [6]	FF + H + Glosses	17.30
Bi-ST-LSTM-A [5]	H + Articulations position	24.68	DenseTCN [28]	FF	14.30
Transformer-CTC [29]	FF	24.59	CTF [30]	FF	11.20
BiLSTM-CTC [3]	FF + Glosses	24.30	Align-iOpt [31]	FF	6.10
CNN-LSTM-HMM [32]	Glosses + H/M	24.10	BiLSTM + CTC [3]	FF + Glosses	2.40
Att-TDCNN [2]	H/M + 2D skel. + Flow	23.70	SLRGAN [33]	FF + Glosses	2.10
Proposed	FF + Flow + 2D/3D Pose	21.34	TMC-BiLSTM-CTC [1]	FF + H + F + Pose	2.10
TMC-BiLSTM-CTC [1]	FF + H + F + Pose	**21.00**	**Proposed**	FF + Flow + 2D/3D Pose	**1.48**

4.2. Implementation details

Our model contains approximately $5M$ trainable parameters. It is trained for 100 epochs, keeping the one with the lowest validation error. The Adam optimizer is employed in its training, using batch size 4, learning rate 10^{-3}, and weight decay 10^{-3}. Plateau learning scheduling is applied with a decrease factor of 0.7 when the validation score does not improve for 8 evaluation steps. Concerning the fusion scheme, we employ the model trained on stream (A) of Section 3.2 (joint-position and appearance) as the guiding model. We also add a label smoothing term equal to 0.025 that aims to penalize low-entropy distributions. The model is implemented in PyTorch, and the experiments are performed on an Nvidia RTX 2080Ti GPU.

4.3. Evaluation Results

Our CSLR model is evaluated quantitatively in terms of gloss error rate (GER, %). First, in Table 1, we provide a comparison against state-of-the-art models on both benchmarks considered. As shown there, our CSLR model achieves a GER of 21.34% and 1.48% on the RWTH-PHOENIX Weather 2014T dataset and the CSL corpus, respectively. In the first case, it outperforms most results in the literature, coming very close to the state-of-the-art (21.00% GER) of [1], where multi-modal appearance features are combined with 2D pose feature maps using a temporal multi-cue (TMC) module, and subsequently fed to a BiLSTM-CTC model for sequence prediction. In the second case, though, our model achieves the state-of-the-art result, significantly outperforming the best alternative by a 30% relative GER reduction (1.48% vs. 2.10%).

Table 2. Gloss error rate (GER, %) on the RWTH-PHOENIX Weather 2014T dataset of various feature combinations in conjunction with our proposed model.

Feature streams	GER (%)
2D skeleton	51.10
2D skeleton + Appearance	23.16
2D skeleton + Appearance + Optical Flow	22.28
3D skeleton	53.72
3D skeleton + Appearance	23.35
3D skeleton + Appearance + Optical Flow	22.37
"ExPose" parameters (Rotation)	50.25
Rotation + Appearance + Optical Flow	22.14
Joint-position + Appearance (A)	23.03
Joint-motion+ Optical Flow (B)	23.15
Rotation + Appearance (C)	22.96
A + B	22.04
A + C	21.75
A + B + C	**21.34**

Further, in Table 2, we evaluate our system on the RWTH-PHOENIX Weather 2014T dataset, when various modality combinations are considered. As it may observed, our network yields competitive performance when all three streams are considered, revealing the benefit of explicitly combining spatio-temporal dynamics and different skeletal representations. When our CSLR model relies exclusively on skeletal structures, we report inferior accuracy, demonstrating that incorporating additional visual feature representations is crucial. As deduced from Table 2, "ExPose" parameters seem to be a robust representation, achieving the highest accuracy compared to the 2D and 3D skeleton. Fusing 2D and 3D pose information boosts system performance, indicating that 2D and 3D skeletal data are complementary to each other.

We also investigate the contribution of combining ST-GCNs with BiLSTMs in the encoder, comparing the performance of the proposed ST-GCN/BiLSTM/CTC model against two variations of it on the RWTH-PHOENIX Weather 2014T dataset. First, we evaluate our model without the inclusion of the BiLSTM encoder. This degrades GER to 22.42%, showing the BiLSTM benefit and confirming our intuition that modeling both short-term and long-term dynamics is important in CSLR. We also consider a baseline 3-layer BiLSTM encoder alone. Such model yields a GER of 24.04%, thus validating the power of ST-GCNs. Finally, we evaluate our approach without the guiding method. This increasing GER by 3.5% absolute, confirming the benefit of synchronizing the CTC spikes of the three model streams.

5. CONCLUSION

In this work, we focused on the challenging task of CSLR from RGB videos, proposing a ST-GCN based sequence learning model that operates on multiple visual representations of the signing activity, capturing signer pose, shape, appearance, and motion information. These feature streams are combined into three ST-GCN modules, which are followed by BiLSTMs and an appropriate fusion scheme via a guiding CTC approach. Further, we investigated the utility of 3D human pose and shape parameterization via the "ExPose" approach, as well as 3D skeletal joint information inferred from detected 2D joints via OpenPose. Our ablations demonstrated the benefit of all modules of the proposed architecture. Compared to the state-of-the-art, our system achieved competitive performance on the popular RWTH-PHOENIX Weather 2014T dataset and set the new state-of-the-art on the Chinese SLR corpus (Split I setup).

6. REFERENCES

[1] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," in *Proc. AAAI*, 2020, pp. 13009–13016.

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

[3] I. Papastratis, K. Dimitropoulos, D. Konstantinidis, and P. Daras, "Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space," *IEEE Access*, 8: 91170–91180, 2020.

[4] Z. Niu and B. Mak, "Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition," in *Proc. ECCV*, 2020, pp. 172–186.

[5] Q. Xiao, X. Chang, X. Zhang, and X. Liu, "Multi-information spatial-temporal LSTM fusion continuous sign language neural machine translation," *IEEE Access*, 8: 216718–216728, 2020.

[6] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. AAAI*, 2018, pp. 2257–2264.

[7] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, 45(11): 2673–2681, 1997.

[8] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.

[9] C. C. Amorim, D. Macêdo, and C. Zanchettin, "Spatial-temporal graph convolutional networks for sign language recognition," in *Proc. ICANN*, 2019, pp. 646–657.

[10] L. Meng and R. Li, "An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network," *Sensors*, 21(4): 1120, 2021.

[11] M. Vázquez-Enríquez, J. L. Alba-Castro, L. Docío-Fernádez, and E. Rodríguez-Banga, "Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks," in *Proc. CVPRW*, 2021, pp. 3457–3466.

[12] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proc. CVPRW*, 2021, pp. 3408–3418.

[13] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *Proc. ECCV*, 2020, pp. 20–40.

[14] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. CVPR*, 2017, pp. 4645–4653.

[15] 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. ECCVW (SLRTP)*, 2020, pp. 249–263.

[16] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. CVPR*, 2018, pp. 7784–7793.

[17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. CVPR*, 2018, pp. 7122–7131.

[18] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. CVPR*, 2017, pp. 2720–2729.

[19] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. ICCV*, 2017, pp. 2659–2668.

[20] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Patt. Anal. Mach. Intell.*, 36(7): 1325–1339, 2014.

[21] C. Ionescu, F. Li, and C. Sminchisescu, "Latent structured models for human pose estimation," in *Proc. ICCV*, 2011, pp. 2220–2227.

[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI*, 2017, pp. 4278–4284.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NeurIPS*, 2012, pp. 1097–1105.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[25] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI*, 2018, pp. 7444–7452.

[26] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proc. CVPR*, 2019, pp. 4496–4505.

[27] G. Kurata and K. Audhkhasi, "Guiding CTC posterior spike timings for improved posterior fusion and knowledge distillation," in *Proc. Interspeech*, 2019, pp. 1616–1620.

[28] D. Guo, S. Wang, Q. Tian, and M. Wang, "Dense temporal convolution network for sign language translation," in *Proc. IJCAI*, 2019, pp. 744–750.

[29] N. C. Camgöz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. CVPR*, 2020, pp. 10020–10030.

[30] S. Wang, D. Guo, W.-g. Zhou, Z.-J. Zha, and M. Wang, "Connectionist temporal fusion for sign language translation," in *Proc. ACM MM*, 2018, pp. 1483–1491.

[31] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proc. CVPR*, 2019, pp. 4160–4169.

[32] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos," *IEEE Trans. Patt. Anal. Mach. Intell.*, 42(9): 2306–2320, 2020.

[33] I. Papastratis, K. Dimitropoulos, and P. Daras, "Continuous sign language recognition through a context-aware generative adversarial network," *Sensors*, 21(7): 2437, 2021.