SIGN LANGUAGE RECOGNITION VIA DEFORMABLE 3D CONVOLUTIONS AND MODULATED GRAPH CONVOLUTIONAL NETWORKS

Katerina Papadimitriou, Gerasimos Potamianos

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

ABSTRACT

Automatic sign language recognition (SLR) remains challenging, especially when employing RGB video alone (i.e., with no depth or special glove-based input) and under a signer-independent (SI) framework, due to inter-personal signing variation. In this paper, we address SI isolated SLR from RGB video, proposing an innovative deep-learning framework that leverages multi-modal appearance- and skeleton-based information. Specifically, we propose three components for the first time in SLR: (i) a modified version of the ResNet2+1D network to capture signing appearance information, where spatial and temporal convolutions are substituted by their deformable counterparts, accomplishing both prevalent spatial modeling potential and motion-aware modeling adaptability; (ii) a novel spatio-temporal graph convolutional network (ST-GCN) that integrates a GCN variant, involving weight and affinity modulation for modeling diverse correlations between different body joints beyond the physical human skeleton structure, followed by a self-attention layer and a temporal convolution; and (iii) the "PIXIE" 3D human pose and shape regressor to generate 3D joint-rotation parameterization used for ST-GCN graph construction. Both appearance- and skeleton-based streams are ensembled in the proposed system and evaluated on two datasets of isolated signs, one in Turkish and one in Greek. Our system outperforms the state-of-the-art on the second set, yielding 53% relative error rate reduction (2.45% absolute), while it performs on par with the best reported system on the first.

***Index Terms*—** SI isolated sign language recognition, deformable 3D-CNN, ST-GCN, modulated GCN, "PIXIE"

1. INTRODUCTION

Automatic SLR from videos constitutes an important research problem, gaining considerable attention in recent years, enhancing accessibility for the hearing impaired, while also being incorporated in sign language (SL) learning applications [1–3]. Nevertheless, SLR remains an intricate task due to the multitude of strongly correlated manual and non-manual modalities, such as handshapes, shoulder motion, body leaning, head pose, mouthing patterns, eye gaze, and eyebrow movement, all contributing to sign formation [4], as well as the scarcity of data resources. Such issues are substantially more challenging for SLR under an SI setting [5–9], which is adopted in this work, due to the inherent articulation variability among signers.

Following deep-learning and computer vision advances, many recent SLR works have been combining appearance-based descriptors derived from RGB or optical flow frames and appropriate representations of the signer's skeletal information, achieving the state-of-the-art on various isolated SLR datasets [10–14]. For example, in [6] a 3D-CNN model is combined with a unified ST-GCN called G3D, in [7] a multi-modal setup is employed that includes 2D human pose landmarks and hand images, and in [8] an ensemble approach exploits multi-modal information from skeletal keypoints and features, as well as RGB, optical flow, and depth. Also, in our earlier work [15], the contribution of optical flow, human skeletal features, and appearance features of handshapes and mouthing is explored.

Here, motivated by the above, we propose a system focusing on SI isolated SLR relying on two main modalities: (i) a 3D-CNN model for capturing the spatio-temporal SL articulation dynamics and (ii) a ST-GCN to learn the spatial and motion correlation of the human skeletal joints (see also Fig. 1). Both aforementioned models are trained separately, and their outputs are combined through an ensemble module that exploits the last fully-connected layer outcome.

Most deep-learning SLR works rely on appearance-based spatio-temporal features, extracted by applying 2D-CNNs on RGB data [11, 16] and/or motion informative frames [15]. Lately, there have been works adopting 3D-CNNs for this purpose, since such models can capture spatio-temporal SL articulation correlations [8,12,14]. Specifically, in [14] an inflated 3D-CNN is deployed for extracting the spatio-temporal feature sequence to learn long-range temporal dependencies, while in [8] a pretrained ResNet2+1D is used that decouples the 3D-CNN spatial and temporal convolutions, yielding higher accuracy over other popular 3D-CNNs. Here, our first innovation is that we employ the ResNet2+1D network [17] as a backbone, but we substitute the spatial and temporal convolutions with their deformable counterparts. These are suitable for learning complex geometric transformations and inter-frame motions, by augmenting the convolution receptive field.

Due to the scarcity of SL data resources, data-hungry appearance-based methods struggle to attain their full potential. To mitigate this, recent studies leverage progress in vision-based extraction of whole-body keypoints and explore their integration into skeleton-based GCNs. Specifically, in [18] a ST-GCN approach is introduced based on skeletal data, capturing the dynamic aspects of SL cues in the spatial and temporal domains, while in [8] a spatial decoupling GCN is used, followed by an attention mechanism and a temporal convolution. Although decoupled GCNs employ different weight matrices for each graph node enhancing performance, the model size rises. Here, to rectify this, we adopt modulated spatial GCNs [19] that employ weight modulation to adjust the shared feature transformation for each node, maintaining a small model size. Another limitation of regular GCNs is that the graph is usually predefined following the human skeleton structure, while SL involves motion patterns beyond the natural body joint connections. To address this, we employ an affinity modulation technique that adds a learnable mask to the adjacent matrix. Moreover, we integrate an attention module, as well as a temporal convolution layer. This constitutes the second innovation of this paper.

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).

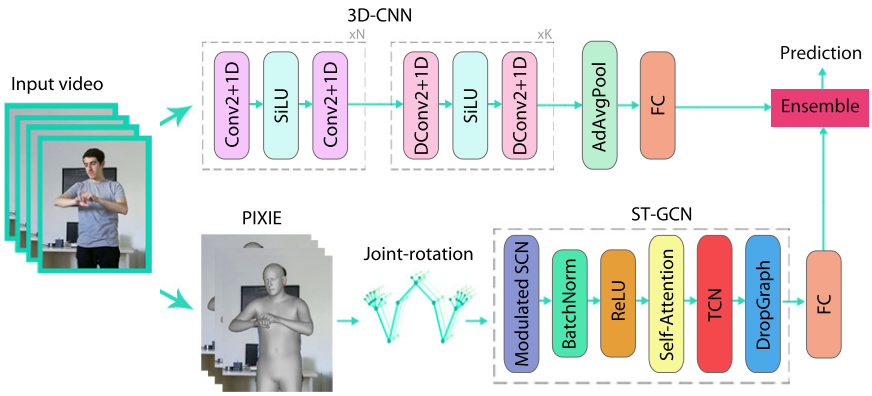

Fig. 1. Proposed isolated SLR model architecture: Two groups of streams are fused, namely a deformable ResNet2+1D network applied on the RGB image frames (appearance modality) and a ST-GCN model, which incorporates modulated spatial GCN (SCN), self-attention module, and temporal convolution (TCN) that operates on the 3D joint-rotation feature based skeleton graph (skeleton-modality).

For graph construction, most works rely on 2D skeletal joints [6, 8], while in [20] the ST-GCN employs 3D skeletal joints as graph features. Recently, in [21] we have introduced a model for continuous SLR that ensembles ST-GCNs based on both 2D and 3D skeletal joint graph features, as well as 3D joint-rotation parameterization. The third innovation of this paper regards the graph construction, where we utilize the 3D joint-rotation parameterization of the human skeleton, which instead of estimating via the "ExPose" human pose regression model [22], we deploy a recently introduced 3D human pose and shape regressor, the so-called "PIXIE" [23], that achieves state-of-the-art performance in many benchmarks.

In summary, our work contributions are: (i) the development of a novel version of the ResNet2+1D network that incorporates deformable spatial and temporal convolutions; (ii) the design of an innovative ST-GCN unit that relies on modulated GCNs; and (iii) the use of 3D joint-rotation parameterization, extracted via the "PIXIE" 3D pose and shape regressor, for graph construction. To date, none of the above have been investigated in conjunction with SLR.

We evaluate our introduced approach on two popular isolated SLR benchmarks, the "AUTSL" [11] and "ITI GSL" [14] corpora, and we provide in-depth ablations that highlight our innovations. Comparing our method to state-of-the-art SI SLR systems, we achieve superior performance on the second dataset and very competitive results on the first.

2. OUR APPROACH

As already mentioned, our approach composes of two main modalities, a spatio-temporal 3D-CNN, which integrates deformable spatial and temporal convolutions for feature extraction from RGB videos, and an attention-based ST-GCN model that relies on modulated GCNs, as well as temporal convolutions to learn motion dynamics from the human skeleton. All are detailed next.

2.1. Deformable 3D-CNN

Due to the strong spatio-temporal correlation of the various SL articulators participating in signing, with each one carrying specific information content, their visual representation is a key aspect in SLR. Thus, for spatio-temporal visual feature learning, a 3D-CNN model is employed that decouples spatial and temporal convolutions of 3D-CNNs. Adopting the 18-layer ResNet2+1D network [17] as the underlying architecture, we replace the spatial and temporal convolutions with their deformable counterparts. Encouraged by the effective modeling of geometric variations obtained by deformable CNNs evaluated on challenging benchmarks in multiple domains, we choose them for learning human-related complex geometric transformations and inter-frame motions.

In ResNet2+1D, 3D convolutional kernels of dimension $t \times K \times K$ are replaced by spatial convolutional filters with dimensionality of $1 \times K \times K$ coupled with temporal convolution filters of size $t \times 1 \times 1$. In this work, instead of using regular convolutions operating on a fixed sampling grid, we append learned offsets to the grid of regular convolutional kernel enlarging the convolution receptive field. Specifically, we integrate deformable convolutions that operate on two steps: (i) a regular convolutional layer is applied predicting the position offsets Δp_n, where $n = 1, ..., |R|$ with R being the sampling grid of a regular convolution and (ii) the sampling grid is augmented by adding the predicted offsets Δp_n to the normal convolution operation.

Despite the robustness of deformable convolutions, their computational cost is higher than that of regular convolutions. For that purpose, we selectively apply them in the three last stages, instead of employing them in the entire network. Note that we performed experiments replacing the convolutions at all stages and at different stages of the ResNet2+1D model, concluding that adopting them at the last three stages of the network is the most accurate choice.

Finally, to further strengthen our model capacity, we replace the ReLU activation function with the SiLU one [24].

2.2. Attention-based Modulated ST-GCN

We introduce an ST-GCN that relies on a modulated GCN followed by a temporal convolution enhanced with an attention mechanism. For graph construction, we employ the 3D joint-rotation parameterization of the human skeleton. All are detailed next.

2.2.1. Graph Construction

GCNs are a generalized variant of CNNs, where filters operate on graph-structured data with nodes corresponding to human-body joints. Specifically, a graph is defined as $G = (V, E)$, where V denotes the node set with N human skeletal joints, and E accounts for the intra-skeleton edges. The edges can be represented by an adjacent matrix $\mathbf{A} \in \{0, 1\}^{N \times N}$, where 1 corresponds to direct connection between a joint pair and 0 to non-direct. Each joint i is related to a D-dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^D$, thus $\mathbf{X} \in \mathbb{R}^{D \times N}$ denotes a matrix that aggregates the features of all graph nodes.

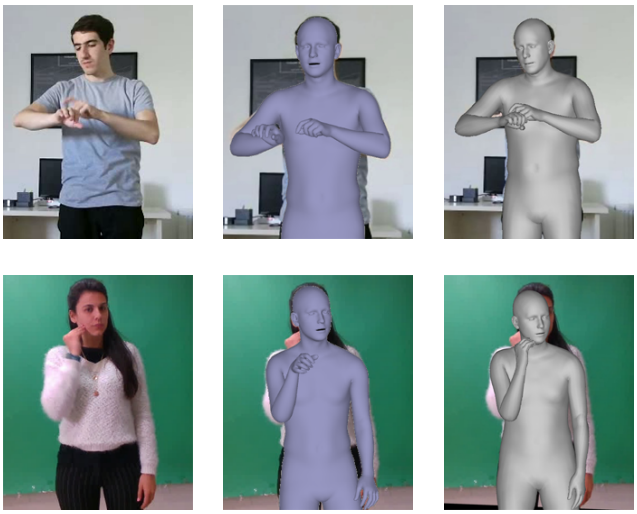

Fig. 2. 1st column: Sample frames from the AUTSL dataset [11] (top) and the ITI GSL database [14] (bottom); 2nd column: 3D body reconstruction via the "ExPose" regression model [22]; 3rd column: 3D human body reconstruction by the "PIXIE" estimator [23].

Following our previous work in [21], we adopt the 3D joint-rotation parameterization of the human pose as graph feature representations. Specifically, in [21] we used the "ExPose" 3D human reconstruction model [22] that relies on separate networks to regress body, face, and hand parameters. The problem is that such methods depend exclusively on the separate part predictions, where in many cases hands and face predictions can be wrong (see also Fig. 2 (2nd column)). To avoid this, here, we deploy a recently introduced model, the so-called "PIXIE" [23], which infers 3D body pose and shape parameterization using a moderator (see also Fig. 2 (3rd column)). "PIXIE" first applies encoders to extract features from cropped images of the body, face, and hands, while the moderator is trained to predict the confidence score for each part. The confidence score is exploited to compute the weighted average of the body and part features. Then, the fused features are fed to separate hand and face networks for parameter prediction. Hence, the final prediction includes information from both full body and part images. The "PIXIE" framework regresses parameters for the human shape and pose, as well as the facial expressions. Thus, shape and expression are described by 250 coefficients in total, while the whole body pose includes 55 joints with 6 degrees of freedom, i.e. 25 body pose joints including head, jaw, and neck pose, as well as 15 joints per each hand, yielding (6×55)-dimensional feature vectors.

2.2.2. ST-GCN Unit

The forward propagation rule of a spatial GCN layer given input $\mathbf{X}_{in}$ is implemented as follows:

$$\mathbf{X}_{out} = \sigma(\mathbf{W}\mathbf{X}_{in}\hat{\mathbf{A}}) \ , \tag{1}$$

where $\mathbf{X}_{out} \in \mathbb{R}^{D' \times N}$ is the output feature vector, $\sigma()$ denotes the activation function, $\mathbf{W} \in \mathbb{R}^{D' \times D}$ is the learnable weight matrix, and $\hat{\mathbf{A}}$ is the normalized affinity matrix. Each ST-GCN unit involves a spatial GCN that transforms and aggregates the node features and their neighbors followed by a temporal convolution, which operates in the temporal node neighborhood across adjacent frames.

One limitation of the spatial graph convolution function is that all nodes in the graph share the same feature transformation $\mathbf{W}$, obstructing the modeling of the correlation between different body joints that do not appertain to the same neighbor. To resolve this, most recent works adopt the decoupled graph convolution, where an independent learnable weight matrix is assigned to each node for transformation. Despite the enhanced performance accomplished via weight unsharing, the model size rises significantly. Here, we use a variant of graph convolution, called modulated GCN [19], which composes of two basic components: (i) weight modulation and (ii) affinity modulation. In weight modulation, the graph convolution function is supplemented by a learnable weight modulation vector $\mathbf{M} \in \mathbf{R}^{D' \times N}$ that is unique for each node i and is used to modulate the shared weight matrix. Thus, (1) is transformed as follows:

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

with $\odot$ denoting element-wise multiplication. In regular GCNs the adjacent matrix $\mathbf{A}$ is usually predefined following the human skeleton structure, but SL involves motion patterns beyond the natural body joint connections, as in Fig. 2 (1st column). The affinity modulation technique addresses this issue by adding a learnable mask $\mathbf{Q} \in \mathbb{R}^{N \times N}$ to matrix $\mathbf{A}$, i.e. $\mathbf{A}' = \mathbf{A} + \mathbf{Q}$. To prevent overfitting, we deploy symmetry regularization on affinity modulation, producing a symmetric affinity matrix as introduced in [19].

The GCN is followed by self-attention, involving a spatial, a temporal, and a channel attention module, all three connected in cascade to enhance the nodes context representation. Further, a temporal convolution is used in order to learn the relational patterns between consecutive frames. Since the dropout layer does not enhance GCN performance, here, in order to avoid overfitting, a DropGraph module [25] is added, where one node is dropped together with its neighbor node set. In the introduced model, ten such ST-GCN units are utilized, followed by a global average pooling layer on both spatial and temporal domains before the fully-connected layer.

2.3. Multi-modal Fusion

Finally, we apply an ensemble module, where the posteriors returned from the last fully-connected layers of the two different modalities are appropriately fused. More precisely, we assign different weights to each modality in accordance with their individual performance and sum them up, producing the final probability scores.

3. EXPERIMENTAL FRAMEWORK

3.1. Datasets

As already mentioned, the performance of the proposed system is evaluated on two publicly available isolated SLR corpora.

The *AUTSL dataset* [11] contains 36,302 RGB+D videos of 226 Turkish isolated signs that are performed by 43 signers and recorded with 20 different backgrounds. Here, we employ the RGB-only stream, available at 30 Hz and 512×512 resolution, and we adopt the official SI data split, comprising 28,142 training videos (31 signers), 4,418 validation ones (5 signers), and 3,742 test videos (7 signers).

The *ITI GSL database* [14] contains RGB+D videos of 15 continuous Greek SL (GSL) dialogues performed by 7 different signers, 5 times each. It includes temporal gloss annotations, thus allowing extraction of 40,826 isolated sign videos (with vocabulary size equal to 310). Here, we use the RGB stream (30 Hz rate, 648×480 resolution), and perform SI SLR via 7-fold cross-validation (one test signer per fold, with SLR models trained on the remaining 6).

Table 1. SLR accuracy (%) on both datasets using various 3D-CNNs on the appearance stream alone.

CNN Models	AUTSL	ITI GSL
C3D [27]	81.95	85.66
I3D [14]	87.64	89.11
P3D [28]	90.57	92.14
R3D [29]	92.04	94.03
ResNet2+1D + ReLU [17]	93.26	95.89
ResNet2+1D + SiLU	93.85	95.98
ResNet2+1D (pretrained) + SiLU [8]	94.77	96.51
Ours	**95.39**	**97.12**

Table 2. SLR accuracy (%) on both datasets using ST-GCN variants on the skeletal stream only.

ST-GCN Variations	AUTSL	ITI GSL
w/o Attention	93.88	94.85
w/o Modulated GCN	94.59	95.04
w/o DropGraph	95.12	95.79
w Decouple GCN	95.17	95.84
Ours	**95.32**	**96.14**

Table 3. SLR accuracy (%) on both datasets using various streams for the skeleton graph construction (skeletal stream alone).

Streams	AUTSL	ITI GSL
2D Joint-position	94.96	95.46
3D Joint-position	95.10	95.68
2D Joint-motion	92.54	93.11
3D Joint-motion	93.24	93.57
Joint-rotation ("ExPose")	95.15	95.74
Joint-rotation ("PIXIE")	**95.32**	**96.14**

Table 4. Ours vs. literature results on both datasets. Notation: appearance (A), hand appearance (HA), skeleton (S), optical flow (F).

Dataset	Model	Modalities	Acc. (%)
AUTSL	VTN-PF [7]	A + HA + S	92.92
	MS-G3D [6]	A + S	96.15
	Ours	A + S	96.67
	SAM-SL [8]	A + S + F	**98.42**
ITI GSL	I3D + BiLSTM [14]	A	89.74
	OpenHands [31]	S	95.40
	Ours	A + S	**97.85**

3.2. Implementation Details

For the 3D-CNN, we crop the upper body using the 3D joints generated by the MediaPipe framework [26] and resize it to 256×256. We employ ResNet2+1D-18 as the underlying model, pretrained on the Kinetics dataset. To further enhance SLR performance, we pretrain our model on the Chinese SL dataset [10]. During finetuning, we use the Adam optimizer with an initial learning rate of 0.0001 and weight decay of 0.0001. We employ the cross-entropy loss function with label smoothing, and set the mini-batch size to 16.

For graph construction, a 55-node skeleton graph is applied with 6-dimensional skeleton features corresponding to the rotation representation dimension, which is estimated by "PIXIE". The initial learning rate is 0.1 and the weight decay is set to 0.001. Adam optimization and batch size 16 are used.

Both modalities are trained for 200 epochs, saving the checkpoint of each one with the lowest validation error. Then the posteriors are weighted and summed as: $p_{\text{fused}} = 1.0\, p_{\text{app}} + 0.9\, p_{\text{skel}}$. The system is implemented in PyTorch, and the experiments are performed on an Nvidia RTX 3090 GPU.

3.3. Evaluation Results

Our isolated SLR model is evaluated on both datasets in terms of sign recognition accuracy (%). First, in Table 1, we compare the introduced deformable 3D-CNN model against state-of-the-art 3D-CNN variations. Our model achieves 95.39% and 97.12% accuracies on AUTSL and ITI GSL, respectively. In both cases it outperforms all 3D-CNN alternatives considered. Next, we conduct ablations on the proposed ST-GCN, depicted in Table 2. Clearly, the modulated GCN and the attention mechanism seem to be the most important contributors. Further, in Table 3 we evaluate the ST-GCN, when different features are employed for graph construction. Note that the 2D skeleton is extracted using the HRNet whole-body pose estimator [30], while for the 3D skeleton regression the MediaPipe holistic model [26] is employed. It is obvious that the joint-rotation parameterization derived via the "PIXIE" framework achieves the highest recognition accuracies.

Finally, in Table 4 we evaluate model performance against the literature, when both modalities are considered and their outputs are fused. In the case of ITI GSL, our model achieves 97.85% SLR accuracy, outperforming the state-of-the-art by 2.45% absolute, corresponding to a 53% relative error reduction. In the case of AUTSL, our model yields 96.67% accuracy, trailing the state-of-the-art result (98.42%) of [8].[1] Note, however, that multiple modalities were considered in that work, including appearance, optical flow, skeleton, and skeletal features, at the expense of increased complexity and computational cost. We can actually beat this result slightly, reaching 98.45%, if we also incorporate optical flow in our ensemble module (derived from the RAFT model [32]).

Revisiting the tabulated results, a comparison of the "Ours" entries of Table 1 (appearance stream alone) to the corresponding entries of Tables 2 or 3 (skeletal stream only) shows that the 3D-CNN appearance module outperforms the skeletal ST-GCN one (AUTSL: 95.39% vs. 95.32%, ITI GSL: 97.12% vs. 96.14%). Nevertheless, both modules perform well, and their fusion improves performance further (Table 4, "Ours"), namely by 28% relative error reduction on AUTSL and 25% on ITI GSL (over appearance only).

4. CONCLUSIONS

In this work, we focused on the challenging problem of SI isolated SLR, investigating the contribution of fusing two different modalities that operate on visual representations of appearance and human pose to capture signing activity. In particular, we explored the integration of deformable convolutions in the ResNet2+1D network for augmenting the convolution receptive field. Further, we introduced a novel ST-GCN model relying on modulated GCNs, an attention mechanism, and temporal convolutions to capture local and global human skeletal joint dynamics. For graph construction we explored the utility of 3D human joint-rotation parameterization, estimated by the "PIXIE" approach. Finally, we assembled both modalities in the proposed system, significantly outperforming the state-of-the-art on the ITI GSL corpus and reaching competitive performance on the popular AUTSL dataset.

[1] Slightly better performance (98.53%) is reported in [8] when incorporating depth video to the system, but here we focus on RGB video only.

5. REFERENCES

[1] G. Potamianos, K. Papadimitriou, E. Efthimiou, S. E. Fotinea, G. Sapountzaki, and P. Maragos, "SL-ReDu: Greek sign language recognition for educational applications. Project description and early results," in *Proc. PETRA*, 2020.

[2] C. K. Mummadi, F. P. P. Leo, K. D. Verma, S. Kasireddy, P. M. Scholl, and K. V. Laerhoven, "Real-time embedded recognition of sign language alphabet fingerspelling in an IMU-based glove," in *Proc. iWOAR*, 2017.

[3] C. Ong, I. Lim, J. Lu, C. Ng, and T. Ong, "Sign-language recognition through gesture & movement analysis (SIGMA)," in *Mechatronics and Machine Vision in Practice 3*, J. Billingsley and P. Brett, Eds., pp. 235–245. Springer, 2018.

[4] K. Antzakas and B. Woll, "Head movements and negation in Greek sign language," in *Gesture and Sign Language in Human-Computer Interaction*, I. Wachsmuth and T. Sowa, Eds., vol. LNCS-2298, pp. 193–196. Springer, 2002.

[5] P. M. Ferreira, D. Pernes, A. Rebelo, and J. S. Cardoso, "Signer-independent sign language recognition with adversarial neural networks," *Int. J. Machine Learning*, 11(2): 121–129, 2021.

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

[7] M. De Coster, M. Van Herreweghe, and J. Dambre, "Isolated sign recognition from RGB video using pose flow and self-attention," in *Proc. CVPRW*, 2021, pp. 3436–3445.

[8] 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.

[9] A. Moryossef, I. Tsochantaridis, J. Dinn, N. C. Camgöz, R. Bowden, T. Jiang, A. Rios, M. Müller, and S. Ebling, "Evaluating the immediate applicability of pose estimation for sign language recognition," in *Proc. CVPRW*, 2021, pp. 3429–3435.

[10] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese sign language recognition with adaptive HMM," in *Proc. ICME*, 2016.

[11] O. M. Sincan and H. Y. Keles, "AUTSL: A large scale multimodal Turkish sign language dataset and baseline methods," *IEEE Access*, 8: 181340–181355, 2020.

[12] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proc. WACV*, 2020, pp. 1448–1458.

[13] S. Albanie, G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, and A. Zisserman, "BOBSL: BBC-Oxford British sign language dataset," *CoRR*, arXiv:2111.03635, 2021.

[14] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. Th. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Trans. Multimedia*, 24: 1750–1762, 2022.

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

[16] 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.

[17] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. CVPR*, 2018, pp. 6450–6459.

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

[19] Z. Zou and W. Tang, "Modulated graph convolutional network for 3D human pose estimation," in *Proc. ICCV*, 2021, pp. 11457–11467.

[20] M. Al-Hammadi, M. A. Bencherif, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, W. Abdul, Y. A. Alohali, T. S. Alrayes, H. Mathkour, M. Faisal, M. Algabri, H. Altaheri, T. Alfakih, and H. Ghaleb, "Spatial attention-based 3D graph convolutional neural network for sign language recognition," *Sensors*, 22(12): 4558, 2022.

[21] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, "Spatio-temporal graph convolutional networks for continuous sign language recognition," in *Proc. ICASSP*, 2022, pp. 8457–8461.

[22] 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.

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

[24] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, 107: 3–11, 2018.

[25] 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. ECCV*, 2020, p. 536–553.

[26] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," *CoRR*, arXiv:1906.08172, 2019.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. ICCV*, 2015, pp. 4489–4497.

[28] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. ICCV*, 2017, pp. 5534–5542.

[29] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. J. Belongie, and Y. Cui, "Spatiotemporal contrastive video representation learning," in *Proc. CVPR*, 2021, pp. 6960–6970.

[30] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. CVPR*, 2019, pp. 5686–5696.

[31] P. Selvaraj, G. Nc, P. Kumar, and M. Khapra, "OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages," in *Proc. ACL*, 2022, pp. 2114–2133.

[32] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. IJCAI*, 2021, pp. 4839–4843.