INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multimodal Sign Language Recognition via Temporal Deformable Convolutional Sequence Learning

Katerina Papadimitriou , Gerasimos Potamianos

Electrical and Computer Engineering Department, University of Thessaly, Volos 38221, Greece
aipapadimitriou@uth.gr, gpotam@ieee.org

Abstract

In this paper we address the challenging problem of sign language recognition (SLR) from videos, introducing an end-to-end deep learning approach that relies on the fusion of a number of spatio-temporal feature streams, as well as a fully convolutional encoder-decoder for prediction. Specifically, we examine the contribution of optical flow, human skeletal features, as well as appearance features of handshapes and mouthing, in conjunction with a temporal deformable convolutional attention-based encoder-decoder for SLR. To our knowledge, this is the first use in this task of a fully convolutional multi-step attention-based encoder-decoder employing temporal deformable convolutional block structures. We conduct experiments on three sign language datasets and compare our approach to existing state-of-the-art SLR methods, demonstrating its superiority.

Index Terms: sign language recognition, OpenPose, optical flow, temporal deformable convolutions

1. Introduction

Automatic recognition of sign language (SL) constitutes an important human-computer interaction (HCI) technology, allowing communication for the speech and hearing impaired. Since SL is a non-vocal form of communication, information is delivered visually, involving the simultaneous use of hand movements in conjunction with facial expressions, hand shapes and orientations that complement each other. Among SL variants, recognition of fingerspelling [1–3], of isolated signs [4, 5], and of continuous signing [6, 7] have attracted significant interest.

SLR remains a challenging problem due to the variability in individual signing styles and gesture coarticulation. To address it, there have been numerous schemes proposed, differing in image data feature acquisition and/or temporal modeling (sequence learning). Concerning the former, most works have focused on either sensor-based or vision-based approaches. Specifically, some studies rely on input from hand gloves or motion capturing systems, facilitating hand tracking and, by extension, sign capturing, however at the expense of HCI naturalness [7, 8]. In the meantime, recent advances in computer vision and deep learning have re-ignited interest in vision-based methods for the extraction of spatio-temporal representations from SL videos. The most dominant SLR works rely on human skeletal data [4, 9, 10] generated by OpenPose [11] or motion tracking through optical flow estimation [12], while others explore their combination [4]. For image feature extraction, the majority of methods adopt deep learning models, with the most established ones being 2D [2, 4] or 3D [6, 13] convolutional neural networks (CNNs). To address sequence modeling, various works have approached SLR as a classification problem exploiting mainly recurrent neural networks (RNNs) and fully connected layers [4, 14] or connectionist temporal classification (CTC) [1, 2, 15]. Recently, SLR has been treated as a linguistic task resolved by automatic speech recognition approaches such as hidden Markov models [16, 17] and sequence-to-sequence models that adopt encoder-decoder modules relying mostly on RNNs and incorporating attention mechanisms [1, 6].

Regarding SLR in videos as an image-to-text translation task, here we introduce an end-to-end deep learning-based recognition scheme relying on spatio-temporal feature fusion and an attentional sequence learning model. Specifically, spatial features are obtained through the OpenPose detector [11] for skeletal data acquisition (body pose, hands, and face) that are also employed for hand and mouth region segmentation. Additionally, motion informative images are generated through optical flow using SpyNet [18]. Handshapes, mouthing, and optical flow feature representations are then obtained by the ResNet-18 model [19]. Subsequently, the features are fused and fed to an attention-based encoder-decoder module. To this end, instead of employing a typical RNN-based encoder-decoder, we introduce a novel fully convolutional encoder-decoder similar in spirit with the model in [3]. The key difference between [3] and our model is the substitution of standard convolutions with temporal deformable convolutional block structures [20] in the encoder-decoder. Note that the incorporation of temporal deformable convolutions (TDCs) to an encoder-decoder structure is also investigated in [21] for video captioning. Nevertheless, our model deviates from using a mean-pooling layer, capitalizing instead TDCs on both the encoder and decoder. In addition, our model adopts a different attention mechanism relying on a quadratic alignment function [3].

In summary, our work contributions are: (i) the development of a system that deviates from the SLR state-of-the-art approaches in the fused visual streams, where apart from skeletal data, we use both optical flow and feature representations of the hands and mouth; (ii) the design of a novel fully convolutional attention-based encoder-decoder architecture relying on TDCs; and (iii) the development of an end-to-end recognition scheme that can be evaluated on various SL forms (here, fingerspelling, isolated SL, and continuous SL).

Specifically, we evaluate our introduced approach on three SL corpora: (i) The Polytropon Greek SL (GSL) corpus [22], achieving a significant word accuracy improvement of 6.2% absolute, compared to prior approaches [23] that rely only on handshape features and a convolutional attention-based encoder-decoder; (ii) the ChicagoFSWild dataset [2], improving signer-independent letter accuracy by 4.9% absolute over the best-performing previous frameworks that use an attention-based RNN encoder and CTC loss, but without involving any

This research was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: 2456).

http://dx.doi.org/10.21437/Interspeech.2020-2691

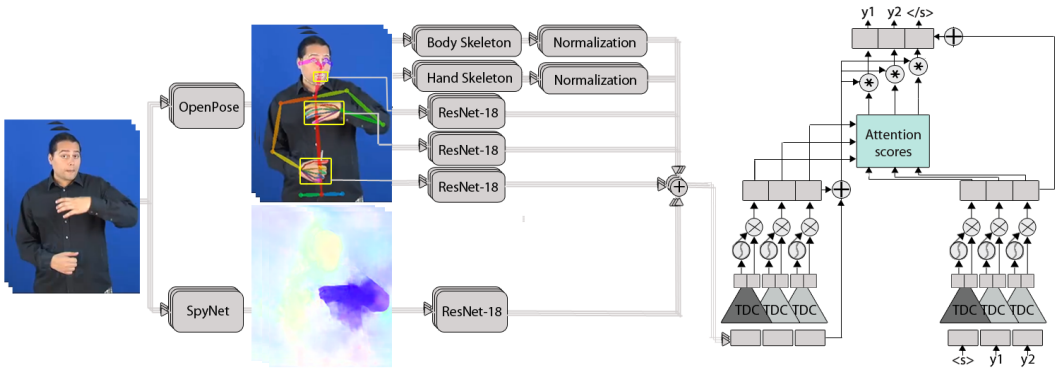


Figure 1: *Architecture of the introduced SLR model that generates natural language from SL videos through the fusion of spatio-temporal features (left) and a sequence learning model employing multi-step attention-based temporal deformable convolutions (right).*

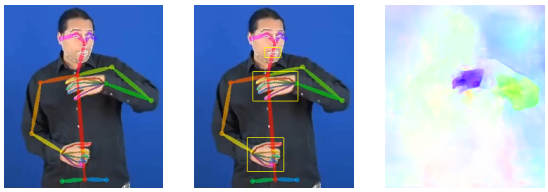
hand detection technique; and (iii) the RWTH-PHOENIX-Weather 2014T dataset [24], improving signer-dependent word accuracy by 1.2% absolute over the best earlier reported results obtained by a Transformer encoder-decoder [25].

2. Methodology

An overview of our system architecture is depicted in Fig. 1. As it can be observed, it comprises: (i) a feature extraction module that relies on human skeleton information (body, hand, and face), optical flow, as well as handshapes and mouthing feature representations, followed by their fusion; and (ii) a fully convolutional multi-step attentional encoder-decoder based on TDCs for the prediction task. All are detailed next.

2.1. Human Skeletal Features

In order to extract skeletal features, we employ the OpenPose framework [11], which is a human joint detector based on deep convolutional pose models. OpenPose provides a detailed representation of the human body, extracting in total 137 joint coordinates of the body pose, hands, and face. Specifically, it can estimate 25 body-pose keypoints, 21 keypoints of each hand, as well as 70 face keypoints, extracted as image coordinates (see also Fig. 2(a)). Since in the majority of SL videos only the upper body participates in signing, here we focus on 57 keypoints, discarding 10 body joints that correspond to invisible lower body parts of the signer, as well as all face keypoints. In addition, to obtain translation and scale invariance, we normalize all extracted human skeletal joints by converting the image to a local coordinate system with the neck keypoint being its origin, and further normalizing them based on the distance between the left and right shoulder keypoints. This yields 114-dimensional (dim) features, 30 of which correspond to the coordinates of the 15 kept body skeleton joints, and the remaining 84 to the hands (21 joints for each hand).



(a) (b) (c)

Figure 2: *(a) Input image with super-imposed keypoints generated by OpenPose [11]; (b) input image marked with rectangular boxes enclosing the handshapes and the mouth region derived based on the human skeleton; and (c) optical flow image generated by SpyNet [18].*

2.2. Handshape and Mouthing Features

The most dominant SL information involves the signer hands (shape, orientation, and trajectory) and in a lesser degree any lip motion. For this purpose, we segment the two hands and mouth regions-of-interest (ROIs), based on the corresponding skeletal coordinates obtained by OpenPose (see Fig. 2(b)). This yields three ROIs, each of which is subsequently fed to a multi-layer 2D-CNN image feature learner. Specifically, a ResNet-18 network [19] is used for this purpose, pretrained on the ImageNet database [26]. Feature maps are generated by taking the output of the fully connected layer, yielding 512-dim features for each ROI. The network uses 3×3 convolutional kernels with stride 2. In all cases, input images are resized to the fixed size of the ResNet-18 network input layer (224×224 pixels).

2.3. Optical Flow Features

Another critical aspect in SLR is motion information, typically extracted as optical flow. To acquire it, the well-known SpyNet [18] model is employed, which combines classical optical flow algorithms with deep learning techniques. Once the optical flow is estimated, optical flow images are generated by colorizing velocity vectors according to their magnitude and orientation between adjacent frames (see also Fig. 2(c)). Such images, after resizing, are fed to a ResNet-18 feature learner, similarly to the hand and mouth ROIs, yielding 512-dim features.

2.4. Feature Fusion

The resulting feature streams are concatenated, generating a 2,162-dim feature vector (2×512 for the two hand ROIs, 512 for the mouth ROI, 512 for the optical flow image, and 114 for skeletal features), which is subsequently fed to the encoder-decoder module for prediction. Note that in the case of a missing stream, the corresponding features are set to zeros.

2.5. Attention-based Temporal Deformable Convolutional Encoder-Decoder

The concatenated feature vectors are subsequently fed to a multi-layer convolutional encoder-decoder complemented with attention (see Fig. 1) relying on TDCs [20]. Specifically, we exploit a fully convolutional attention-based architecture [3], but instead of using standard convolutional block structures to generate latent state representations we perform TDCs [21]. In the

Table 1: *Word accuracy (%) comparison of the proposed on all three datasets for SD SLR using various feature stream combinations.*

Features					Datasets		
Optical flow (512-dim)	Handshapes (1024-dim)	Mouth (512-dim)	Hand-skeleton (84-dim)	Body-skeleton (30-dim)	PGSL	ChicagoFSWild	RWTH-PT
	✓				89.03	87.12	70.38
	✓	✓			90.22	88.21	70.64
	✓	✓	✓	✓	93.31	91.20	74.62
	✓	✓		✓	92.10	91.89	73.36
	✓			✓	91.74	90.95	71.83
✓					91.56	88.35	72.24
✓	✓	✓			93.55	89.41	74.35
			✓	✓	85.68	84.69	69.41
✓	✓	✓		✓	94.96	91.89	74.24
✓	✓	✓	✓	✓	**95.31**	**92.63**	**76.30**

typical form of the proposed model, the l-th encoder layer reads in latent-representation sequential data and outputs a sequence of hidden states h^l, while the l-th decoder layer generates d^l hidden states and maps the latter to the desired output. Additionally, each layer is composed of a one-dimensional TDC sequence complemented with gated linear units (GLUs) [27]. Specifically, a GLU operates as a gating tool over the TDC output $H = [AB] \in \mathbb{R}^{2D}$ using $u(H) = A \otimes \sigma(B)$, where $u \in \mathbb{R}^D$ expresses which of A inputs associate with the current target element, and $\otimes$ denotes point-wise multiplication.

As already mentioned, both encoder and decoder employ TDC block structures to generate the latent state representations h^l and d^l. As with the deformable spatial convolutions [28], the idea behind TDC blocks is that the temporal samplings specified by the convolutional kernel k are augmented with auxiliary temporal offsets learned end-to-end along with the other network parameters. Specifically, the TDC is performed in two phases: (i) the temporal offsets computation through processing the sampled input features by an one-dimensional convolution, and (ii) the output features aggregation exploiting the temporal offsets via an additional one-dimensional convolution, as also schematically depicted in Fig. 3. For simplicity, we assume that $F = (h^{l-1}_{i-(k-1)/2}, ..., h^{l-1}_{i}, ..., h^{l-1}_{i+(k-1)/2})$ is a subsequence of input sequence h^{l-1} equal in length with odd kernel size k. Thus, a set of temporal offsets $\{\Delta f^n_i\}^k_{n=1} \in \mathbb{R}^k$ is generated by applying a one-dimensional convolution to the concatenated k elements of F, resulting in an offset field with the same size as the input sequence F. Subsequently, the output of the TDC block is produced by augmenting samples with temporal offsets $(h^{l-1}_{i-(k-1)/2+\Delta f^1_i}, ..., h^{l-1}_{i+\Delta f^{k/2}_i}, ..., h^{l-1}_{i+(k-1)/2+\Delta f^k_i})$ via bilinear temporal interpolation and by feeding them to an additional one-dimensional convolution.

Once a one-layer encoder with kernel width k generates hidden states h^l_j relating to k inputs, stacking multiple layers on top of each other results in states that are related to more inputs than previously. Meanwhile, the feed-forward convolutional structure in the encoder enables the parallelization within the input sequential data providing fast computation. In addition, our model is equipped with a multi-step attention mechanism [29, 30] relying on a quadratic attentional scoring function introduced in [3]. To enable deeper convolutional networks, residual functions concerning each TDC input and layer output are appended [19].

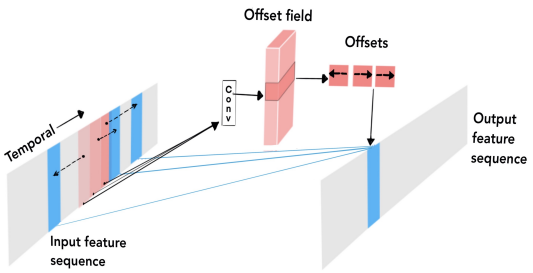


Figure 3: *An illustration of TDC (figure modified from [21]).*

3. Experimental Evaluation

3.1. Datasets and Experimental Framework

We conduct experiments on the following three databases:

Polytropon GSL corpus (PGSL): This dataset [22] contains three repetitions of 3,600 sentences performed by a single signer. Corpus annotations are based on the ELAN video annotation tool [31, 32] and are provided at both the signed sentence and signed word levels. The corpus signed vocabulary consists of 2,664 words, and here we use 103 unique words that appear between 30 to 110 times. These yield 5,414 isolated word video clips segmented based on the annotation time-stamps. Experiments are conducted using the setup of [23].

Chicago Fingerspelling in the Wild (ChicagoFSWild): This dataset [2] contains clips of fingerspelling sequences collected from online videos and annotated through ELAN. It consists of 7,304 fingerspelling clips performed by 160 signers with a vocabulary of 3,553 unique fingerspelled words. For the signer-dependent (SD) case, we use 103 unique fingerspelled words performed by 143 signers appearing between 10 to 130 times yielding 3,076 videos based on ELAN annotation time-stamps. SD evaluation is conducted using ten-fold cross-validation. On the other hand, for the signer-independent (SI) setting we use the dataset existing partitions (5452, 984, and 867 videos for training, validation, and testing, respectively).

RWTH-PHOENIX-Weather 2014T (RWTH-PT): This dataset [24] is an extension of the corpus in [33], extracted from the German TV station PHOENIX news and weather forecast, consisting of both sign-gloss annotations and spoken language translations for SL videos. Here, we use only the gloss level annotations for the SLR task. The dataset includes 8,257 German SL videos from 9 unique signers with gloss-level annotations of a vocabulary of 1,066 unique signs. For SD evaluation we utilize the existing split of the dataset, which consists of 7096, 519, and 642 samples for training, validation, and testing, respectively. For SI evaluation, we employ 9-fold cross-validation.

Table 2: *Word accuracy (%) comparison of state-of-the-art encoder-decoder models on the three datasets under the SD and SI frameworks. The number of model parameters are also shown (in millions).*

Experimental Paradigm $\longrightarrow$		SD			SI	
Model	# Parameters (M)	PGSL	ChicagoFSWild	RWTH-PT	ChicagoFSWild	RWTH-PT
ALSTM	0.699	91.83	90.63	72.12	56.25	44.86
AGRU	0.815	91.42	90.04	71.81	59.32	44.15
ACNN	1.373	92.59	91.12	73.98	60.94	46.99
Transformer	7.819	89.37	88.55	70.87	54.40	42.32
CNN + GRU	1.028	90.83	91.84	73.45	60.88	47.55
TDCNN + GRU	2.013	91.95	92.34	74.63	61.97	48.18
Proposed	3.145	**95.31**	**92.63**	**76.30**	**62.50**	**50.90**

3.2. Evaluated System Details

We compare our approach to a number of alternative encoder-decoder (enc-dec) sequence models:
Attentional LSTM enc-dec (ALSTM): The feature vectors are fed to a one-layer LSTM [34] encoder-decoder with hidden dimensionality equal to 128. Training is conducted with an initial learning rate of 0.001 decreased by a factor of 0.3.
Attentional GRU enc-dec (AGRU): The model comprises of a 2-layer attention-based GRU encoder-decoder with 128 hidden units. During training an initial learning rate of 0.003 decreased by a factor of 0.3 is employed.
Attentional CNN enc-dec (ACNN) [3]: The model constitutes a 2-layer multi-step attention-based CNN encoder-decoder with kernel width 5 and 128 hidden units. An initial learning rate of 0.001 decreased by a factor of 3.0 is used.
Transformer enc-dec (Transformer): Here, a 6-layer transformer with 8 heads for self-attention and 2048-dimension hidden transformer feed-forward is employed. An initial learning rate of 0.3 decreased by a factor of 0.3 is employed and parameter initialization is carried out using Xavier [35].
CNN enc & attention-based GRU dec (CNN+GRU): The model comprises of an attention-based 3-layer convolutional encoder with kernel width 5 and a one-layer GRU decoder. The size of hidden states is fixed at 128 and an initial learning rate of 0.001 reduced by 0.3 is used.
TDCNN enc & attention-based GRU dec (TDCNN + GRU): Feature vectors are fed to a 3-layer temporal deformable convolutional encoder with kernel width 5 and a 2-layer GRU decoder. An initial learning rate of 0.001 decreased by a factor of 0.3 and 128 hidden units are employed.
Attention-based TDCNN enc-decoder (Proposed): The model comprises of a 3-layer TDCNN block structure with kernel width 5 and 128 hidden units. An initial learning rate of 0.001 decreased by a factor of 0.1 is used.

All models are trained employing the Adam optimizer [36] with a dropout rate of 0.3. To achieve a better matching of a target element, the beam search strategy [37] with beam width of 2 in decoding is applied. Finally, the mini-batch size is fixed to 128. All models are implemented in PyTorch [38] and the training is carried out using GPU acceleration.

Table 3: *Comparative evaluation of various model variations in word accuracy (%) for the SD case, with L being the number of layers, KW the kernel widths, and BW the beam widths.*

Model details			Datasets		
L	KW	BW	PGSL	ChicagoFSWild	RWTH-PT
1	3	2	90.77	89.22	65.49
2	3	3	93.84	91.92	68.77
3	3	3	94.61	91.89	71.56
1	5	2	92.53	90.25	68.04
2	5	2	94.87	92.07	71.23
3	5	4	95.22	92.10	72.12
3	5	2	**95.31**	**92.63**	**76.30**

3.3. Results

First, in Table 1, we evaluate the combination of different visual stream feature representations in conjunction with our proposed SLR model. All results are reported in word accuracy (%) for the SD case. It can be observed that the fusion of all feature streams yields the best results. Further, the optical flow seems to constitute a more powerful representation than the skeletal features, while using just the latter yield the lowest accuracy.

Next, Table 2 compares the proposed SLR model against state-of-the-art models on all datasets under both SD and SI experimental paradigms. It should be noted that only the multi-signer datasets are considered for the SI case. It can be observed that the proposed model turns out superior to the considered alternatives in terms of word accuracy, revealing the power of exploiting long-range contextual relations using TDCs on the encoder-decoder structure that can boost the performance of the prediction task. It can be readily seen that the results for all models in the SI case are much lower than those in the SD case. This is primarily due to the video quality of both datasets, which is rather low compared to studio data, as well as signing variability among subjects. We also evaluated the performance of the proposed model in terms of letter accuracy (%) on the ChicagoFSWild dataset, improving over the best reported results of [2] from 45.1% to 50.0% under the SI experimental paradigm. Note also, that compared to previous approaches [23, 25], our model yields word accuracy improvements from 75.12% to 76.30% on the RWTH-PT dataset in the SD case and from 89.10% to 95.31% on the PGSL corpus.

Finally, Table 3 reports word accuracy results of a number of model variations regarding the number of layers, the kernel widths, and the beam width used during decoding, showcasing that deeper architectures benefit model performance.

4. Conclusions

In this paper we proposed an end-to-end model for effective sign-based language recognition, relying on spatio-temporal feature extraction and fusion followed by a fully convolutional attention-based encoder-decoder that exploits temporal deformable convolutions. We highlighted how the incorporation of multiple representation streams and temporal deformable convolutions improves feature learning performance. The performance evaluation on SLR state-of-the-art databases for sign-based communication under both SD and SI settings demonstrated that the proposed model outperforms other sequence learning architectures.

5. 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 Workshop*, 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. Papadimitriou and G. Potamianos, "End-to-end convolutional sequence learning for ASL fingerspelling recognition," in *Proc. of the Annual Conference of the International Speech Communication Association*, 2019, pp. 2315–2319.

[4] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "A deep learning approach for analyzing video and skeletal features in sign language recognition," in *Proc. of the IEEE International Conference on Imaging Systems and Techniques*, 2018, pp. 1–6.

[5] H. Wang, C. Xiujuan, X. Hong, G. Zhao, and X. Chen, "Isolated sign language recognition with Grassmann covariance matrices," *ACM Transactions on Accessible Computing*, vol. 8, no. 4, 2016.

[6] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4160–4169.

[7] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, "A modified LSTM model for continuous sign language recognition using Leap Motion," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 7056–7063, 2019.

[8] B. G. Lee and S. M. Lee, "Smart wearable hand device for sign language interpretation system with sensors fusion," *IEEE Sensors Journal*, vol. 18, no. 3, pp. 1224–1232, 2018.

[9] F. Nugraha and E. C. Djamal, "Video recognition of American sign language using two-stream convolution neural networks," in *Proc. of the International Conference on Electrical Engineering and Informatics*, 2019, pp. 400–405.

[10] S. Ko, J. Son, and H. Jung, "Sign language recognition with recurrent neural network using human keypoint detection," in *Proc. of the Conference on Research in Adaptive and Convergent Systems*, 2018, pp. 326–328.

[11] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4645–4653.

[12] D. R. Kartika, R. Sigit, and Setiawardhana, "Sign language interpreter hand using optical-flow," in *Proc. of the International Seminar on Application for Technology of Information and Communication*, 2016, pp. 197–201.

[13] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. of the International Conference on Computer Vision*, 2017, pp. 5534–5542.

[14] L. Pigou, S. Dieleman, P. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *Proc. of the European Conference on Computer Vision*, 2015, pp. 572–578.

[15] Z. Yang, Z. Shi, X. Shen, and Y. Tai, "SF-Net: Structured feature network for continuous sign language recognition," *CoRR*, arXiv:1908.01341, 2019.

[16] C. Vogler and D. Metaxas, "Parallel hidden Markov models for American Sign Language recognition," in *Proc. of the IEEE International Conference on Computer Vision*, 1999, pp. 116–122.

[17] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, "Speech recognition techniques for a sign language recognition system," in *Proc. of the Annual Conference of the International Speech Communication Association*, 2007, pp. 2513–2516.

[18] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2720–2729.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] P. Lei and S. Todorovic, "Temporal deformable residual networks for action segmentation in videos," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6742–6751.

[21] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao, and T. Mei, "Temporal deformable convolutional encoder-decoder networks for video captioning." in *Proc. of the AAAI Conference on Artificial Intelligence*, 2019, pp. 8167–8174.

[22] E. Efthimiou, K. Vasilaki, S.-E. Fotinea, A. Vacalopoulou, T. Goulas, and A.-L. Dimou, "The POLYTROPON parallel corpus," in *Proc. of the International Conference on Language Resources and Evaluation*, 2018.

[23] 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. of the PErvasive Technologies Related to Assistive Environments Conference*, 2020.

[24] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7784–7793.

[25] N. C. Camgöz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," *CoRR*, arXiv:2003.13830, 2020.

[26] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[27] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. of the International Conference on Machine Learning*, 2017, pp. 933–941.

[28] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. of the IEEE International Conference on Computer Vision*, 2017, pp. 764–773.

[29] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. of the International Conference on Machine Learning*, 2017, pp. 1243—-1252.

[30] S. Chollampatt and H. T. Ng, "A multilayer convolutional encoder-decoder neural network for grammatical error correction," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2018.

[31] "ELAN (Version 5.8) [Computer software]," 2019, Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. [Online] https://archive.mpi.nl/tla/elan.

[32] O. Crasborn and H. Sloetjes, "Enhanced ELAN functionality for sign language corpora," in *Proc. of the Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 2008, pp. 39–43.

[33] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computing*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. of the International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

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

[37] M. Freitag and Y. Al-Onaizan, "Beam search strategies for neural machine translation," *CoRR*, arXiv:1702.01806, 2017.

[38] 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.