HANDS IN FOCUS: SIGN LANGUAGE RECOGNITION VIA TOP-DOWN ATTENTION

Noha Sarhan[⋆]*, Christian Wilms*[⋆]*, Vanessa Closius*[†]*, Ulf Brefeld*[†]*, Simone Frintrop*[⋆]

[⋆]Department of Informatics, University of Hamburg, Hamburg, Germany
[†]Institute of Information Systems, Leuphana University Lüneburg, Lüneburg, Germany

ABSTRACT

In this paper, we propose a novel Sign Language Recognition (SLR) model that leverages the task-specific knowledge to incorporate Top-Down (TD) attention to focus the processing of the network on the most relevant parts of the input video sequence. For SLR, this includes information about the hands' shape, orientation and positions, and motion trajectory. Our model consists of three streams that process RGB, optical flow and TD attention data. For the TD attention, we generate pixel-precise attention maps focusing on both hands, thereby retaining valuable hand information, while eliminating distracting background information. Our proposed method outperforms state-of-the-art on a challenging large-scale dataset by over 2%, and achieves strong results with a much simpler architecture compared to other systems on the newly released AUTSL dataset [1].

Index Terms— Sign language recognition, top-down attention, deep learning

1. INTRODUCTION

Sign Language (SL) is a form of non-verbal communication amongst the deaf or hard-of-hearing that uses a combination of manual and non-manual features to convey meaning. The former involve hands' shape, motion, orientation and place of articulation and are considered to be the dominant, distinguishing part of the sign morphology. The latter involve facial features, such as lip movements and eye contact, which merely convey emphasis and additional meaning. While automatic Sign Language Recognition (SLR) has been addressed for several years, a publicly available SL translation system does not yet exist, hindering deaf people's ability from working with technology or interacting with people who do not know SL. Research on SLR is mainly split into working on 1) isolated SLR (ISLR), where the task is to recognize trimmed sign gloss videos [2, 3, 4], or 2) continuous SLR, where a single video covers an entire sentence, which may require temporal localization [5, 6, 7]. In this work, we focus on ISLR.

In the past decade, SLR witnessed a surge with the success of deep learning [8, 4]. The inherent complexity of SLR drove researchers to explore different modalities including RGB, depth data, and skeletal joints [9, 4]. Including extra modalities does not enable the model to generalize to datasets that lack depth data. In order to capture the fine-grained features, researchers use multi-stream CNNs and complex ensemble methods to process various parts of the inputs in detail. As a consequence, this leads to inflated systems that may require significant computational resources and quickly become difficult to train.

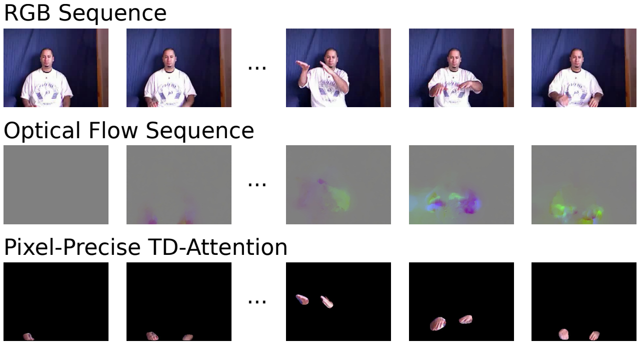

Fig. 1: An example of the three inputs sequences to our 3-stream model. Top: RGB sequence from the ChaLearn249 dataset [10], Middle and bottom: generated optical flow and TD Attention sequences.

Attention mechanisms are a way to selectively focus on relevant parts of the input data, similar to how humans can direct their focus to specific objects or features [11]. There are two main types of attention in this context: bottom-up (BU) and top-down (TD). BU attention refers to the process of automatically detecting salient features in the input data, without any prior knowledge or expectation. In contrast, TD attention involves using prior knowledge to guide attention to specific parts of the input data. In machine learning and computer vision, attention mechanisms have been used to direct the processing of the networks to the relevant parts, including SLR, where [12] use optical flow based motion attention, and [13, 14] use transformer-based attention.

In this paper, we propose a new ISLR system that incorporates TD attention. It comprises only three streams, relying solely on RGB data, making training more feasible in terms of computational resources than recent state-of-the-art models [16, 14]. Besides the commonly used RGB and optical flow data, we utilize pixel-precise TD attention to focus on

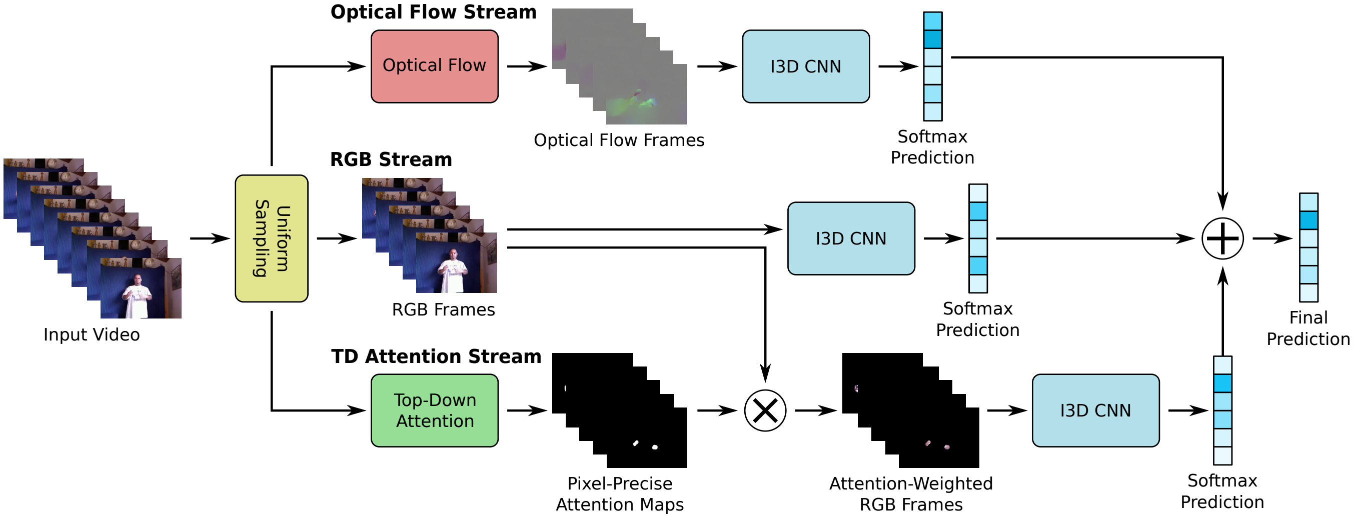

Fig. 2: Our three-stream model for SLR utilizing top-down (TD) attention. The middle stream is fed full-frame RGB data, the upper stream uses optical flow frames, and third TD attention stream (bottom) combines pixel-precise attention maps with the RGB images. All streams use I3D CNN modules [15] to generate per-stream predictions. The predictions of all three streams are then averaged in a late fusion step to compute the final prediction.

the relevant regions in the input inspired by the human visual system. Fig. 1 shows the inputs to our three streams. We leverage the knowledge that the hands' shape, position, orientation and motion trajectory are crucial to recognizing a SL gesture, and hypothesize that TD attention is likely more relevant compared to BU attention. We use a state-of-the-art hand detector [17] as a TD generator and extract pixel-precise attention maps that focus on both hands, while maintaining all relevant information for SLR, that is not possible with coarse body poses or hand crops.

Recent work on SLR has utilized hand crops [14] to focus on the hands. The drawbacks here are that such crops lose valuable information of where the hands are positioned with respect to each other, and to the body. By applying our pixel-precise attention map to remove the background, as opposed to cropping out the hands, we preserve that information, and also maintain information about the motion trajectory of the hands in a sequence of consecutive frames, in addition to hand shape and orientation. Our results show that our simple model, with few streams focusing on important details outperforms other systems or achieves results that are on par with more complex systems.

To summarize, the contributions of this work are:

- We leverage application knowledge to extract task-specific information utilizing a TD attention stream that attends to the most relevant features for SLR.
- We evaluate the benefit of pixel-precise segmentation versus hand crops, and the use of BU attention versus TD attention for SLR to support our hypothesis above.
- We outperform state-of-the-art by more than 2% on the challenging ChaLearn249 IsoGD dataset using only RGB data. We also obtain an accuracy of 97.93% on the recently released AUTSL dataset [1].

2. PROPOSED METHOD

In this section, we thoroughly explain our approach including the TD attention module and the implementation details.

2.1. Network Architecture

The primary idea of our proposed method is to utilize top-down attention to focus on the signers' hands, since they are considered the most relevant for SLR. We achieve this by introducing a novel, three-stream model depicted in Fig. 2. Our model relies only on input RGB data, without using depth data or other modalities, to allow the model to generalize to other datasets that might not have this input.

In our model, we opted for the stream-based design by Sarhan *et al.* [3] as it outperforms others on SLR tasks and is easily adaptable. Each stream uses Inflated 3D (I3D) CNNs, where the 2D $k \times k$ filters and pooling kernels are inflated into 3D kernels by adding a new dimension t and become cubic through this transformation $t \times k \times k$, spanning t frames. The first stream in our proposed model takes the full-frame RGB sequence as input. For the second stream, optical flow is first computed from the RGB input image using Dual TV-L1 algorithm [18]. Including RGB and optical flow streams has been successful in several approaches for ISLR [3, 2, 16, 19].

In the third stream, we utilize task-specific knowledge to apply top-down attention to the RGB sequence. The RGB frames F^i are passed to a TD attention-based module, which generates attention maps A^i focusing on the task-relevant areas. The RGB input frames are then combined with the generated attention maps via element-wise multiplication:

$$X^i = A^i \otimes F^i. \tag{1}$$

We refer to the resulting sequence X^i as attention-weighted RGB frames. This is then fed to the I3D CNN module of the TD attention stream.

To generate the TD attention maps, we use Hand-CNN model [17], which predicts hand masks and orientation. Its architecture builds on the Mask R-CNN network for instance segmentation [20]. It comprises a CNN for extracting features, a region proposal network, a region-of-interest classifier, a bounding box regressor, a CNN for mask prediction, and an attention mechanism to integrate contextual cues into the detection process.

The resulting TD attention maps are masks, which are then applied to the RGB frames, suppressing the irrelevant background information, and maintaining a pixel-precise representation of both hands. We note that this differs from purely cropping the hands as it retains important information about the relative positions of the hands with respect to each other, in addition to positional information, preserving the motion trajectory of the hands.

Finally, all three streams share no parameters, and we opt for late fusion, where the output softmax predictions are averaged together to yield a final prediction, as done in [3].

2.2. Implementation Details

Preprocessing. Since the gesture videos in the dataset are of various lengths, they are uniformly downsampled to a fixed size of 40 frames per video. Afterwards, we crop the frames around the center to a spatial size of 224 × 224.

Training. The I3D CNNs were initialized with the corresponding, pre-trained weights from Kinetics dataset [21], a large-scale human action recognition dataset. A classifier layer, initialized with random weights, was appended to each of the streams, and each underwent separate training. As an optimizer, we opted for Adam and a mini-batch size of 4. Training was stopped using early-stopping technique, and as a loss function, we computed the categorical cross-entropy.

Regularization. Besides early stopping, we used several regularization techniques to prevent overfitting: for data augmentation, the videos are augmented during training using replicas of the video frames, which are either shifted along both the x- and y-axes, or have their brightness changed. A dropout of 0.5 is applied before classification. Batch normalization is used after each convolutional layer.

3. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present quantitative and qualitative results of our TD attention-based SLR approach. We evaluate our method on two popular, large-scale, singer-independet, ISLR RGB-D benchmark datasets, namely ChaLearn249 IsoGD dataset [10], and AUTSL (Ankara University Turkish Sign Language Dataset) [1]. For this paper, we only rely on the RGB data, without including depth data, as clarified in Sec. 1. For evaluation, we adhere to the pre-defined data split provided by the dataset and calculate the classification accuracy as the evaluation metric. For fair comparison, we compare to methods that only rely on RGB data.

Table 1: Comparison of our proposed model with the state-of-the-art methods on ChaLearn249 IsoGD Dataset.

Method	**Accuracy**	
	Validation	**Test**[1]
TD-SLR (ours)	**67.13** %	**70.91** %
Hybrid Attn-I3D-SLR [3]	65.02 %	68.89 %
2SCVN-RGB-Fusion [22]	62.72 %	-
I3D-SLR [2]	62.09 %	64.44 %
8-MFFs-3flc (5 crop)[23]	57.40 %	-
XDETVP [24]	51.31 %	-
SYSU_ISEE [4]	47.29 %	-
3DDSN [22]	46.08 %	-
ASU [19]	45.07 %	-

3.1. Results on ChaLearn249 IsoGD Dataset

The challenging chaLearn249 IsoGD dataset has nearly 50,000 videos spanning 249 different gestures, performed by 21 individuals. Table 1 shows the quantitative results of our proposed method in comparison to other state-of-the-art methods. Our proposed TD-SLR model beats state-of-the-art by over 2% increase on each of the validation and test sets, achieving an accuracy of 67.13% for validation, and 70.91% for the test set. By comparison to the Hybrid Attn-I3D-SLR [3], a method that solely relies on RGB and flow data, the results clearly demonstrate the significance of incorporating hand information in our proposed method.

3.2. Results on AUTSL Dataset

We also evaluated our approach on the more recent AUTSL dataset. It comprises 226 signs that are performed by 43 different signers, and has 38,336 videos samples. In Table 2, we compare our results with the baseline, recent methods, and the top winners of the ChaLearn 2021 challenge. We also highlight the extra modalities used by other approaches besides RGB and optical flow data. This includes adding extra hand and/or face information via hand crops or finger joint locations. Our results surpass the baseline [1] by a vast 48.49%. We also outperform the transformer-based models [14] and [13] by over 5% and over 2%, respectively.

[1] Some methods lack results on the test set as the ChaLearn249 dataset was part of a competition, during which the test set was not yet available.

Table 2: Comparison of our proposed model with the state-of-the-art methods on AUTSL Dataset with the used additional modalities per method.

Method	**Accuracy**	**Additional modalities**		
		Hands	**Face**	**Skeleton**
SAM-SLR [16]	98.42%	x	x	x
S3D [25]	98.34%	x	x	x
TD-SLR (ours)	**97.93%**	**x**		
USTC-SLR	97.62%	x	x	x
jalba [25]	96.15%	x	x	x
VLE-trans. [13]	95.46%	x	x	x
VTN-PF [14]	92.92%	x		x
RGB-MHI [12]	93.53%	x		
Baseline [1]	49.22%	x		

Our results are less than 0.5% lower than the top-performing methods [16, 25]. They propose complex, ensemble methods that rely on additional skeleton and face information, while we rely on fewer input modalities offering a simpler approach, making training more feasible.

3.3. Ablation Studies

We perform ablation studies on the ChaLearn249 dataset and evaluate different adjustments to the proposed architecture to assess the influence of different settings.

Hand Crops vs. Pixel-Precise Segmentation. To assess the value of having pixel-precise attention maps, we experimented with hand crops. We had two separate crops, one for the left hand and one for the right hand. Accordingly, two streams, one for each hand, were used. The overall performance of the network decreased by 1.2% as opposed to using pixel-precise segmentation of the hands. We hypothesize that the main reason would be due to the loss of the positional information of the hands with respect to each other and over the sequence of frames.

Attention Mechanisms. In order to verify our hypothesis that TD attention (using Hand-CNN) is the optimal choice for SLR, we evaluate the use of two different BU attention techniques: a) a CNN-based BU attention network for eye fixation prediction, *ACLNet* [26], and b) a biologically-inspired, traditional saliency system, VOCUS2 [27]. VOCUS2 is a handcrafted feature model, which is fast, simple, and does not require training data.

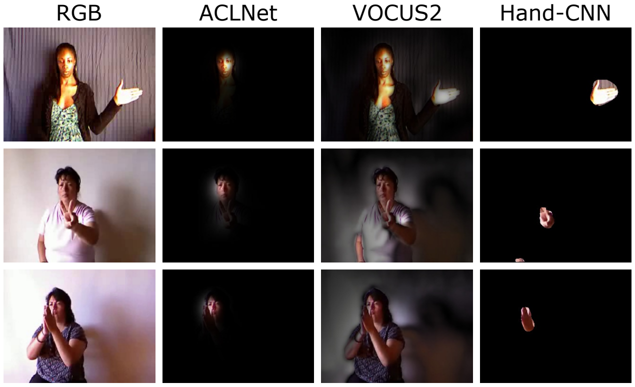

Fig. 3: Sample RGB frames after applying different approaches for attention. From left to right: Original RGB frame, ACLNet [26], VOCUS2 [27], and Hand-CNN [17].

Table 3: Accuracy results of our proposed system using different attention mechanisms in the attention stream.

Method	**Attention Mechanism**	**Accuracy**	
		Alone	**Overall System**
Hand-CNN	TD	52.13 %	67.13 %
ACL Net	BU	50.09 %	65.56 %
VOCUS2	BU	24.12 %	64.33 %

In Table 3, we show the accuracy of the attention stream using the various attention mechanisms as a stand-alone stream, and in the overall architecture. As a stand-alone stream, the TD attention approach using Hand-CNN performed best, outperforming ACLNet by a vast 28% and VOCUS2 by more than 2%. When considering the performance of the overall architecture using the different attention mechanism, Hand-CNN still performed best, surpassing ACLNet and VOCUS2 by 2.8% and 1.57% respectively. To further understand these results, we visualize what each attention mechanism attends to in Figure 3. We observe that ACLNet has the tendency to focus on the signer's face, due to the BU nature of the approach. This explains its poor performance as a stand-alone stream. Facial features alone are not enough to recognize the gesture being signed. VOCUS2 has a much larger region-of-interest, more or less focusing on the signer. While alone this represents valuable information, in the overall architecture it is not impactful, and redundant together with a full-frame RGB.

4. CONCLUSION

An issue frequently encountered in SLR is determining which modalities to incorporate in order to effectively recognize gestures. Consequently, many SLR methods address this challenge by including multiple modalities and proposing intricate ensemble techniques that may be difficult to train.We proposed a TD attention approach to ISLR that relies solely on RGB and utilizes domain knowledge to focus on the hands. By utilizing pixel-precise segmentation of both hands as a TD attention cue, we were able to keep our model simple. Our proposed approach beats state-of-the-art methods on one of the largest, challenging SLR datasets (ChaLearn249), and achieves results that are on par with more complex systems on the relatively new ISLR dataset (AUTSL).

5. REFERENCES

[1] O. M. Sincan and H. Y. Keles, "AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181340–181355, 2020.

[2] N. Sarhan and S. Frintrop, "Transfer learning of videos: From action recognition to sign language recognition," *ICIP*, pp. 1811–1815, 2020.

[3] N. Sarhan and S. Frintrop, "Sign, Attend and Tell: Spatial attention for sign language recognition," in *FG*, 2021, pp. 1–8.

[4] B. Li, W. Li, Y. Tang, J. Hu, and W. Zheng, "GL-PAM RGB-D gesture recognition," in *ICIP*, 2018, pp. 3109–3113.

[5] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep Sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs," *IJCV*, vol. 126, no. 12, pp. 1311–1325, 2018.

[6] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Trans. Multimed.*, vol. 21, no. 7, pp. 1880–1891, 2019.

[7] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-end hand shape and continuous sign language recognition," in *ICCV*, 2017, pp. 3075–3084.

[8] N. Sarhan, Y. El-Sonbaty, and S. Youssef, "HMM-based Arabic sign language recognition using Kinect," in *ICDIM*, 2015, pp. 169–174.

[9] M. Boháček and M. Hrúz, "Sign pose-based transformer for word-level sign language recognition," in *WACV*, 2022, pp. 182–191.

[10] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li, "ChaLearn looking at people RGB-D isolated and continuous datasets for gesture recognition," in *CVPRW*, 2016, pp. 56–64.

[11] H. E. Pashler, *The psychology of attention*, MIT press, 1999.

[12] O. M. Sincan and H. Y. Keles, "Using motion history images with 3D convolutional networks in isolated sign language recognition," *IEEE Access*, vol. 10, pp. 18608–18618, 2022.

[13] I. Gruber, Z. Krnoul, M. Hrúz, J. Kanis, and M. Bohacek, "Mutual support of data modalities in the task of sign language recognition," in *CVPRW*, 2021, pp. 3424–3433.

[14] M. De Coster, M. Van Herreweghe, and J. Dambre, "Isolated sign recognition from RGB video using pose flow and self-attention," in *CVPRW*, 2021, pp. 3441–3450.

[15] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[16] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *CVPRW*, 2021, pp. 3413–3423.

[17] S. Narasimhaswamy, Z. Wei, Y. Wang, J. Zhang, and M. Hoai, "Contextual attention for hand detection in the wild," in *ICCV*, 2019, pp. 9567–9576.

[18] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L1 optical flow," in *DAGM*, 2007, pp. 214–223.

[19] Q. Miao, Y. Li, W. Ouyang, Z. Ma, X. Xu, W. Shi, and X. Cao, "Multimodal gesture recognition based on the RESC3D network," in *ICCVW*, 2017, pp. 3047–3055.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017, pp. 2961–2969.

[21] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *arXiv preprint arxiv:1705.06950*, 2017.

[22] J. Duan, J. Wan, S. Zhou, X. Guo, and S. Z. Li, "A unified framework for multi-modal isolated gesture recognition," *TOMM*, vol. 14, no. 1s, pp. 1–16, 2018.

[23] O. Kopuklu, N. Kose, and G. Rigoll, "Motion fused frames: Data level fusion strategy for hand gesture recognition," in *CVPRW*, 2018, pp. 2103–2111.

[24] L. Zhang, G. Zhu, P. Shen, J. Song, S. Afaq Shah, and M. Bennamoun, "Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition," in *ICCVW*, 2017, pp. 3120–3128.

[25] M. Vazquez-Enriquez, J. L. Alba-Castro, L. Docío-Fernández, and E. Rodriguez-Banga, "Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks," in *CVPRW*, 2021, pp. 3462–3471.

[26] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *TPAMI*, vol. 43, no. 1, pp. 220–237, 2019.

[27] S. Frintrop, T. Werner, and G. Martin Garcia, "Traditional saliency reloaded: A good old model in new shape," in *CVPR*, 2015, pp. 82–90.