Sign, Attend and Tell: Spatial Attention for Sign Language Recognition

Noha Sarhan and Simone Frintrop
Department of Mathematics, Informatics and Natural Sciences, Universität Hamburg, Germany

***Abstract*— Sign Language Recognition (SLR) has witnessed a boost in recent years, particularly with the surge of deep learning techniques. However, most existing methods do not exploit the concept of attention mechanisms, despite their success in several computer vision tasks. In this paper, we propose a novel method for isolated SLR which utilizes spatial attention to focus the processing on the informative, discriminating parts of the input. This is particularly important for SLR, since the RGB image contains several distracting information such as background and signer's clothes, which are irrelevant for the task. We investigate three ways for incorporating spatial attention: a) pre-focused attention, which uses optical-flow-based motion as a prior b) learned attention, where the network learns where to focus during training, and c) hybrid attention, which combines both approaches by initializing the attention layer in the learned attention with the motion-based attention masks used in the pre-focused attention. We show, first, that all three approaches outperform state-of-the-art methods on one of the largest isolated SLR datasets, validating the effectiveness of attention mechanisms on the SLR task, and second, that the best performing approach is the hybrid attention, combining both ideas.**

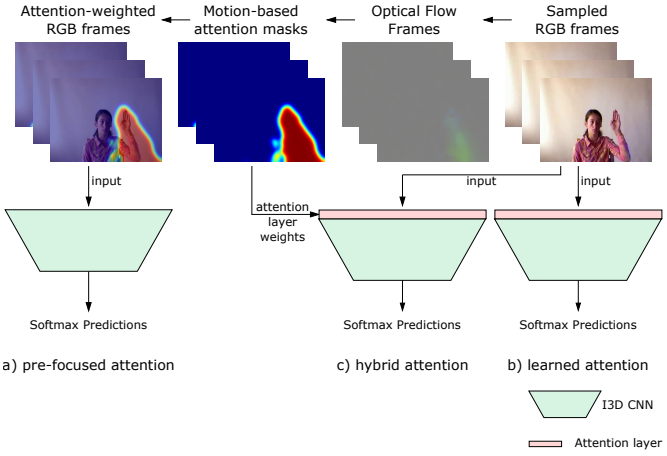

Fig. 1. **Overview of the proposed methods.** The top part shows the steps for generating the attention-weighted RGB frames. The inputs used for the three proposed methods are then depicted.

I. Introduction

Sign language is an important means of communication used by millions of people amongst the deaf and hard-of-hearing. Unfortunately, sign language is not common knowledge to everyone, creating a huge communication barrier. Sign languages provide a complete, well-defined set of hand gestures, governed by grammatical rules that differs from country to country, even from region to region with more than 144 official sign languages [11] existing worldwide. This is represented via hand movement, shape, orientation and place of articulation. Therefore, creating an effective Sign Language Recognition (SLR) system would not only have a great social impact, but would also be impactful to research particularly neighboring computer vision tasks, such as hand gesture and action recognition.

Research in SLR has surged in recent decades [40], [26], [35], however, developing an automated SLR system remains a challenging, open research problem. SLR can be viewed as a hand gesture recognition problem with quick, fine-grained motion. Isolated SLR involves the recognition of a gesture in one video, often representing one word or a compound word. Unlike action recognition, objects in the background should not influence the recognition of the gesture being signed in SLR. The background, the signer's clothes and skin color should have no effect in recognizing the sign being gestured. Therefore, ensuring that the processing is focused on the correct, discriminating features in the image is crucial.

A recently popular and successful approach to focus the processing of a neural network to relevant aspects is utilizing attention mechanisms, not only in computer vision [49], [2], [48], but also in other fields such as natural language processing [12] and speech analysis [5], showing that guiding the network on where to focus the processing positively impacts learning. However, little work has been done in terms of attention in the scope of SLR [17], [42].

In this paper, we build on top of the work by Sarhan et al. [40] and enhance it by including attention (see Fig. 1). [40] uses Inflated 3D (I3D) networks [4] in a two-stream architecture, where the first stream takes RGB sequence as input, and the second one is fed an optical flow sequence. In this work, we incorporate spatial attention to the input RGB stream by generating motion attention maps that are based on optical flow data. We experimented with three different ways of integrating attention: a) *pre-focused attention:* an attention map is pre-computed based on thresholding the optical flow data, b) *learned attention:* a newly added attention layer is incorporated into the network, which learns without priors where to focus more on the RGB image and c) *hybrid attention:* a combination of learned and pre-focused attention, which initializes the weights of the learned attention layer with the pre-focused, motion-based attention map and is finally fine-tuned. Although the computation of the pre-focused attention masks is necessary for the hybrid approach, this overhead is only necessary during training, and is alleviated during testing.

We show that all three approaches outperform the state-of-the-art methods, indicating that attention is beneficial for SLR, and that using motion as a prior is fitting to the task,

978-1-6654-3176-7/21/$31.00 ©2021 European Union

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9666939

Authorized licensed use limited to: Bibliothekssystem Universitaet Hamburg. Downloaded on February 03,2023 at 15:09:58 UTC from IEEE Xplore. Restrictions apply.

since it is an important cue in characterizing sign language gestures. Our best performing approach is the hybrid approach that combines both ideas for attention integration. We will make our code publicly available once the paper has been published.

The contributions of this paper can be summarized as follows:

- To the best of our knowledge, we are the first to incorporate motion-guided spatial attention to the task of SLR
- We present and evaluate two main ways attention can be incorporated: pre-focused attention and learned attention, and show that the most suitable way to integrate attention is a hybrid approach that utilizes both methods
- While our best approach benefits from the pre-computed attention masks, this overhead is alleviated during testing, saving resources and processing time.
- We surpass current state-of-the-art on one of the largest isolated SL dataset, ChaLearn249 IsoGD [45].

II. Related Work

In this section, we give a brief introduction to SLR and then focus on recent research using deep learning methods for the recognition of isolated gesture. Afterwards, we give a brief overview of spatial attention mechanisms in other video understanding tasks, followed by highlighting the latest research on SLR involving attention mechanisms.

A. Sign Language Recognition

SLR can be classified into three subtopics: a) alphabet SLR, b) isolated SLR and c) continuous SLR. Alphabet SLR (aka finger-spelling), as the name suggests involves gestures representing a single alphabet in that particular language. It is usually a static gesture and it suffices to represent the corresponding gesture in a single static image [37], [33]. Isolated SLR, recognizes word-by-word [40], [17], while continuous SLR involves recognition [23], [18], [36], or translation [14], [47], [3] of entire sentences. Both subtopics involve motion in the gesture, and are therefore represented in videos. Since this paper works on isolated SLR, in this subsection, we will focus on the latest research in this area.

Since motion plays a key factor in characterizing sign language, lots of researchers have integrated motion in various ways. This includes motion trajectories [30], optical flow video sequences generated from RGB videos [40], [34], HMM [41], [44] or DTW [28]. Even in deep learning methods, to model temporal dependencies in deep models, RNNs [25] and 3D CNNs are commonly used [40], [34].

It is not surprising that current research in SLR heavily relies on deep learning techniques [35], [26], [40]. Conventional techniques for SLR that used to rely on hand-crafted features, have proven to have very limited success in comparison [53], [52], [41]. Even methods that rely on external equipment such as data gloves that capture hands position, orientation and velocity still do not deliver satisfactory performance [1]. Therefore, in this section, we only focus on current state-of-the-art progress in the field of SLR that uses deep learning methods.

Prominent work in isolated SLR includes [35] by Pigou *et al*. They explore a deep end-to-end neural network incorporating temporal convolution and bidirectional recurrence, showing a significant improvement in frame-wise gesture recognition in videos. More recently, [29] proposed a two-phase recognition system: hand tracking, and hand representation. In the first phase the hand is tracked with the help of a particle filter that combines hand motion and a pre-trained CNN hand model to predict hand position. In the second phase, a compact hand representation is computed by averaging the segmented hand regions.

Transferring domain knowledge has recently been proven beneficial to the task of SLR. Li et al. [26] has transferred the knowledge form web news signs to common everyday words in SLR by learning domain-invariant features. On a larger scale, Sarhan et al. [40] transferred the knowledge learned from the task of action recognition to SLR. They applied a two-stream inflated 3D network, one taking as input RGB frames, and the second optical flow frames. Both streams were pre-trained on RGB and optical flow data from an action recognition dataset Kinetics dataset [4]

B. Attention in Video Understanding Tasks

Since spatial attention mechanisms have shown great success in many image-level tasks, including image captioning [20], [46] and visual question answering [7], [16], it was natural to also incorporate them to extract features for video inputs. This includes tasks such as video captioning [15], [50] and human action recognition [6], [13], [43], [31], [8].

Unlike with still images, attention in for video inputs can be classified into two forms: spatial attention and temporal attention, where the former is only done spatially on a frame-level and the latter involves temporal information. In spatial attention, each frame is processed individually and independently of neighboring frames in the sequence [13], [51]. The authors in [13] proposed an attentional pooling method for action recognition, where they consider both saliency-based and class-specific attention, without utilizing any temporal information.

In the second class of attention, temporal information is also considered, usually along with spatial attention as well (spatio-temporal attention) [9], [27], [32]. For instance, the work by Das *et al.* [8] propose a pose-based spatio-temporal attention mechanism to weigh different body parts for the task of action classification. They utilize RGB videos along with 2D and 3D skeleton points as input. They propose an RNN attention mechanism that provides appropriate weights to the relevant human body parts involved in the action, which improves the action classification.

C. Attention in SLR

In comparison to other computer vision fields, integrating attention mechanisms into the task of SLR is in its infancy. In this section, we present recent work where attention was

Authorized licensed use limited to: Bibliothekssystem Universitaet Hamburg. Downloaded on February 03,2023 at 15:09:58 UTC from IEEE Xplore. Restrictions apply.

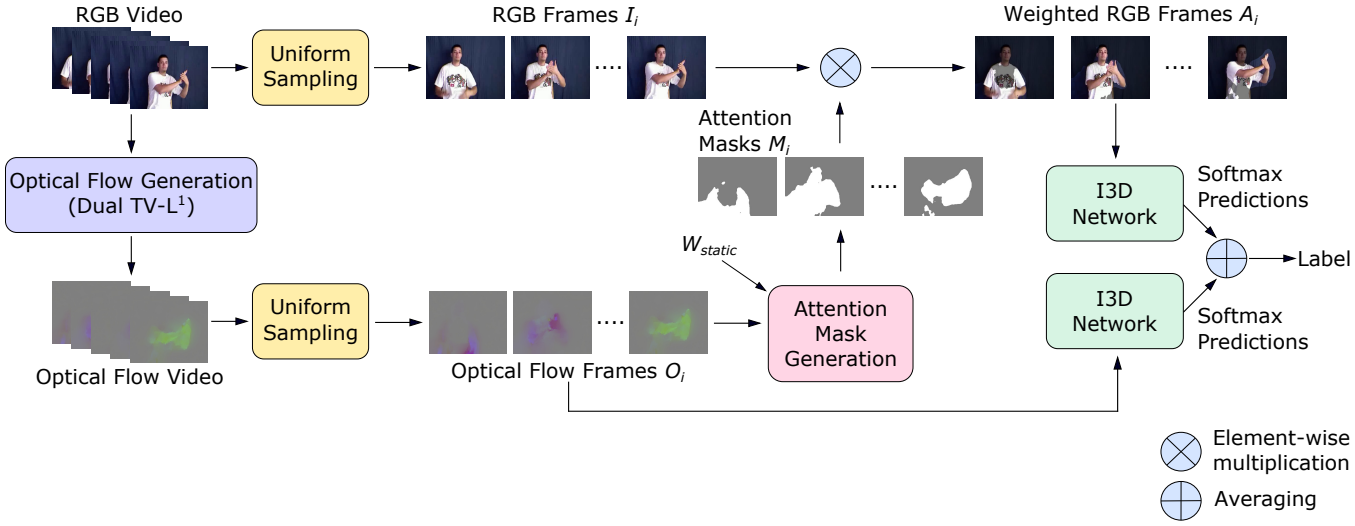

Fig. 2. **Pre-focused Attention Architecture.** A motion prior focuses attention on regions of interest via Attention Masks. The general network architecture consists of two streams of I3D networks [4]. The first (upper) one is fed weighted RGB frames, and the second (lower) one takes optical flow frames as input, which have been computed using the Dual-TVL[1] algorithm [54]. The weighted RGB frames are computed by generating attention masks from the optical flow frames and multiplying them element-wise with the RGB frames. The two I3D Networks generate predictions of the presented gesture and are finally averaged to yield a final label (details of the I3D network in Fig. 3).

involved in SLR. The authors in [17] propose an attention-based 3D CNN for isolated SLR. They rely on skeletal joints captured by Kinect camera and highlight the areas representing the hand and arm joints. They also apply temporal attention to select the significant motions for classification.

While Shi et al. [42] focus on recognition of fingerspelling in American sign language, they proposed a model based on an iterative attention mechanism to obtain the regions of interest of high resolution. It is based on a convolutional RNN to extract a feature map from which the attention map is computed.

The paper by Rodriguez et al. [38] introduce attention in continuous sign language translation. They propose an attention-based encoder-decoder architecture for sequential motion learning. The attention model is included to highlight local temporal patterns that mainly contribute to word translation. They show how important motion is for the task of sign language translation.

III. Proposed Method

The main idea of our proposed method is to integrate attention to the input RGB images with motion as a prior. We categorize the way we apply attention into three categories: *pre-focused attention*, *learned attention*, and *hybrid attention*. In the first method, an attention mask is used to weigh the input RGB image beforehand, while in the second, attention is infused in the network and learned during training. The hybrid approach combines both methods by utilizing the motion-based attention masks from the pre-focused attention approach to initialize the attention layer used in the learned attention approach, while still allowing it fine-tune during training. Applying spatial attention based on optical flow implicitly incorporates temporal information well due to the inherent recurrence in optical flow.

In this section, we first explain the general network architecture, followed by a detailed explanation on how each approach for attention is applied.

A. Network Architecture

The core of our SLR architecture are the Inflated 3D (I3D) networks [4], which have recently shown success in SLR [40], [21]. They are an inflated version of Inception-V1 architecture [19], were 2D $k \times k$ kernels are "inflated" to 3D $t \times k \times k$ kernels that span over t frames. The benefit behind this inflation is to maintain the kernels that were initialized with the pre-trained ImageNet [39] weights via a single-frame static video. I3D models have proven to outperform equivalent CNN+LSTM architectures [4].

Accordingly, we opt for the two-stream I3D architecture proposed by Sarhan et al. [40] to build upon. The first stream takes RGB frames as input, while the second is fed optical flow frames, which have been generated from the RGB video using the Dual-TVL[1] algorithm [54]. The input videos are sampled uniformly in order to have a fixed number of frames for all videos, to be fed to a 3D CNN network. The weights of both streams are initialized with corresponding pre-trained weights from the Kinetics dataset [4], a large-scale action recognition dataset. A new final classification layer is then added to each stream, which starts with randomly-initialized weights. Both streams share no parameters, each network is trained separately, and their output softmax predictions are averaged together to yield a final label, as proposed in [40].

In the following, we present the three different ways how to integrate spatial attention into the architecture to focus the processing on the informative, discriminating parts of the input.

Authorized licensed use limited to: Bibliothekssystem Universitaet Hamburg. Downloaded on February 03,2023 at 15:09:58 UTC from IEEE Xplore. Restrictions apply.

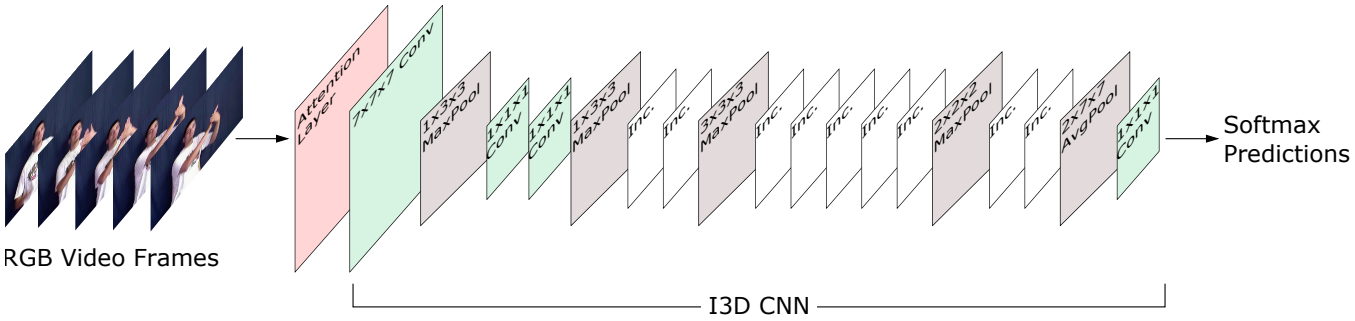

Fig. 3. **Learned Attention RGB Stream.** The detailed architecture of the RGB stream when applying learned attention. A new attention layer (red) is appended to the layers of the I3D CNN at the beginning. Here, the RGB stream is directly fed the original RGB video frames. The weights of the attention layer are initialized to ones and fine-tuned during training.

B. Pre-focused Attention

In this approach, a motion-based attention mask is pre-computed before training, which highlights regions in the image where motion occurs. This is visualized in Figure 2. The attention mask M_i for the i^{th} frame is generated as a binary mask, where pixels that indicate motion have a value of 1, and pixels with no motion have the value W_s, where $0 < W_s <= 1$. Keeping $W_s > 0$ ensures putting more focus on motion areas without losing information from the other parts of the image completely. A pixel at location (x, y) is defined to show motion if the optical flow output from the Dual-TVL[1] algorithm [54] is larger than 0.

The attention-mask, M_i, is then generated, and is used to weigh the RGB image, I_i; the weighted RGB image A_i is obtained by element-wise multiplication of the original RGB image I_i with the corresponding attention mask, such that $A_i = I_i \otimes M_i$. Equation 1 summarizes how the attention-weighted RGB images are generated.

$$A_i(x,y) = I_i(x,y) \otimes M_i(x,y),$$
$$\text{where } M_i(x,y) = \begin{cases} W_s & \text{if } O_i(x,y) = 0 \\ 1 & \text{otherwise,} \end{cases} \quad (1)$$

where O_i is the optical flow output, and $\otimes$ denotes element-wise multiplication. The weighted RGB sequence is then passed through the I3D network of the RGB stream for training, completely replacing the original RGB images.

Rather than having a binary attention mask, where the weights can either be 1 or W_s as shown in Equation 1, we also experimented with blurring the binary attention map beforehand with a Gaussian filter. This results in an attention mask, which has a spatially smoother transition between areas of attention and the surrounding areas and resembles more a typical saliency map. In Table II we show a slight advantage of the blurred attention maps over the binary maps.

C. Learned Attention

In the learned attention approach, we investigate integrating an attention layer to the I3D network rather than having fixed, preset weights as in the previous section. This attention layer is placed as the first layer in the I3D network as shown in Figure 3, directly after the input. The attention layer is time-distributed over the video sequence. The attention layer's weights are initialized to ones, implying that at the first iteration all the pixels in the RGB image get the same weight. As training progresses, the weights are fine-tuned, allowing the network to learn to focus on what is more valuable in the image, without providing any priors beforehand.

D. Hybrid Attention

To reap the benefits of both attention approaches, we propose a hybrid approach. It is basically the same approach as the learned attention, but we initialize the weights of the attention layer with the pre-focused attention mask, M_i (see Equation 1). Similar to the learned attention approach, the weights of the attention layer are also fine-tuned as the network is trained.

IV. Experimental Details

In the following, we describe the dataset and the evaluation framework used to evaluate our proposed methods. In addition, we explain the necessary training details of our experimental setup to make this work reproducible.

A. Dataset & Evaluation

We evaluate our proposed methods on the ChaLearn249 IsoGD dataset [45], a large, isolated SLR dataset. The reason behind choosing this dataset is two-fold: a) it is a signer-independent dataset, which allows to investigate the generalization to unseen individuals in a realistic setting; b) It is one of the most commonly used datasets for isolated sign language gestures [34], [24], [10], [55], [40], which allows us to compare and properly evaluate our proposed method with the current state-of-the-art.

The dataset is made up of 47,933 isolated sign language videos belonging to 249 different classes performed by 21 signers. The videos have been recorded by a Microsoft Kinect camera. While the dataset provides both RGB and depth images, in this work we *only* make use of the RGB data.

For evaluation, we follow the protocol that is already provided by the dataset. It is split into training (35,878 videos), validation (5,784 videos), and test (6,271 videos)

Authorized licensed use limited to: Bibliothekssystem Universitaet Hamburg. Downloaded on February 03,2023 at 15:09:58 UTC from IEEE Xplore. Restrictions apply.

TABLE I

PRE-FOCUSED ATTENTION RESULTS. THIS TABLE SHOWS THE RESULTS OF APPLYING THE ATTENTION MASK IN A PRE-FOCUSED MANNER TO THE INPUT RGB STREAM. NOTE THAT AN ATTENTION WEIGHT OF 1.0 IS EQUIVALENT TO PASSING THE ORIGINAL RGB IMAGE, WE USE THIS AS OUR BASELINE, AS IN [40].

Attention Weight (W_s)	**Validation Accuracy**		**Test Accuracy**	
	RGB	RGB + Flow	RGB	RGB + Flow
0.5	53.08%	61.15%	56.82%	64.0%
0.6	54.0%	61.91%	57.03%	64.23%
0.7	56.1%	63.0%	59.04%	65.78%
0.8	**57.45%**	**63.95%**	**59.96%**	**66.26%**
0.9	55.9%	62.53%	58.23%	65.0%
1.0 (*baseline [40]*)	54.63 %	62.09 %	57.73 %	64.44 %

subsets. For all our experiments, we report and compare the accuracy on both the validation and test sets.

B. Training Details

a) Preprocessing.: The video sequences are uniformly sampled to a fixed number of frames. The sampled frames are then cropped around the center to a spatial size of 224×224. In order to reduce training time, optical flow frames have been generated beforehand as a pre-processing step.

b) Training and fine-tuning.: Both I3D networks have been initialized with pre-trained weights from Kinetics dataset [4]. Training is monitored by calculating the categorical cross-entropy loss, and using Adam as an optimizer [22]. Throughout our methods, we consistently start by training only the randomly initialized classifier layers added to the top of the network, while freezing the pre-trained weights. Afterwards, the pre-trained weights are fine-tuned along with the rest of the network at a lower learning rate. Here we employ early-stopping to automatically halt training once validation loss has not improved for 3 consecutive epochs.

c) Hyperparameters.: When training the randomly initialized top layers, the learning rate was set to 10^{-3}, and the top layers were trained for 3 epochs. The learning rate was then dropped to 10^{-4} to fine-tune the entire network. A mini-batch size of 4 is used.

d) Data augmentation.: While data augmentation is crucial for small- and medium-sized datasets, it remains quite tricky to apply to SLR. Basic data augmentation methods, like flipping the image or slightly rotating it, can directly affect the sign that is being gestured. Taking this into account, we only perform data augmentation through image shifts along both the x- and y-axes and changes in brightness.

V. RESULTS & ANALYSIS

In this section, we present and analyze the results using pre-focused, learned, and hybrid attention methods and compare to state-of-the-art methods on ChaLearn 249. For all experiments, we build on the setup from [40] and change only the RGB stream; the optical flow stream is kept as in [40]. We report both the performance of the RGB stream alone and that of the combined RGB + optical flow streams.

A. Results of Pre-focused Attention

In Table I we present the results of applying the pre-focused attention mask to the RGB input. We experimented with different attention weights W_s for the static areas, while maintaining the focus on the areas where motion is detected at 1. The attention mask has been applied both during training and testing the network.

We observe from the results that giving lower weights to areas with no motion while emphasizing areas with motion achieves better results than feeding the original RGB frames only (corresponding to $W_s = 1$). This is however limited to the point where $0.7 \leq W_s \leq 0.9$. We observe that below 0.7, other important information also gets lost. This may include facial expressions, which are also important for SLR, and the position of the hands with respect to the body. For weights higher than 0.8, we observe that they no longer perform as well, implying that motion areas still need more focus than other static areas.

In order to decrease the processing time, we also experimented by training the network only on the weighted RGB frames, but testing on the original RGB frames. This alleviates heavy computation for the attention mask generation during testing time. The results using the highest performing attention weight of 0.8 were slightly better. The test set achieved an accuracy of $60.39\,\%$ ($+0.43\,\%$ improvement) for the RGB stream alone, and $66.59\,\%$ ($+0.33\,\%$ improvement) when using both the RGB and optical flow streams. This result is interesting, since it shows that it is possible to use the motion prior only during training to guide the training of the network to relevant regions of the input data, and the final trained network does not require the motion prior any more.

We also experimented with blurring the attention map beforehand. We worked on our best-performing results from the binary attention map, with $W_s = 0.8$. We applied a Gaussian filter for smoothing, and achieved best results with a standard deviation of 7. Blurring the mask resulted in slightly better results than a binary mask. For the validation accuracy, we achieved $57.8\,\%$ for the RGB stream alone, and $64.21\,\%$ for both streams. For the test accuracy, we report $60.3\,\%$ for the RGB stream, and $67.11\,\%$ for both the RGB and optical flow streams. A comparison between the binary and blurred attention masks, with $W_s = 0.8$, is shown in Table II.

In Figure 4, we visualize example frames from the ChaLearn249 dataset after applying the motion-based pre-focused attention masks. We observe how optical flow is able

Authorized licensed use limited to: Bibliothekssystem Universitaet Hamburg. Downloaded on February 03,2023 at 15:09:58 UTC from IEEE Xplore. Restrictions apply.

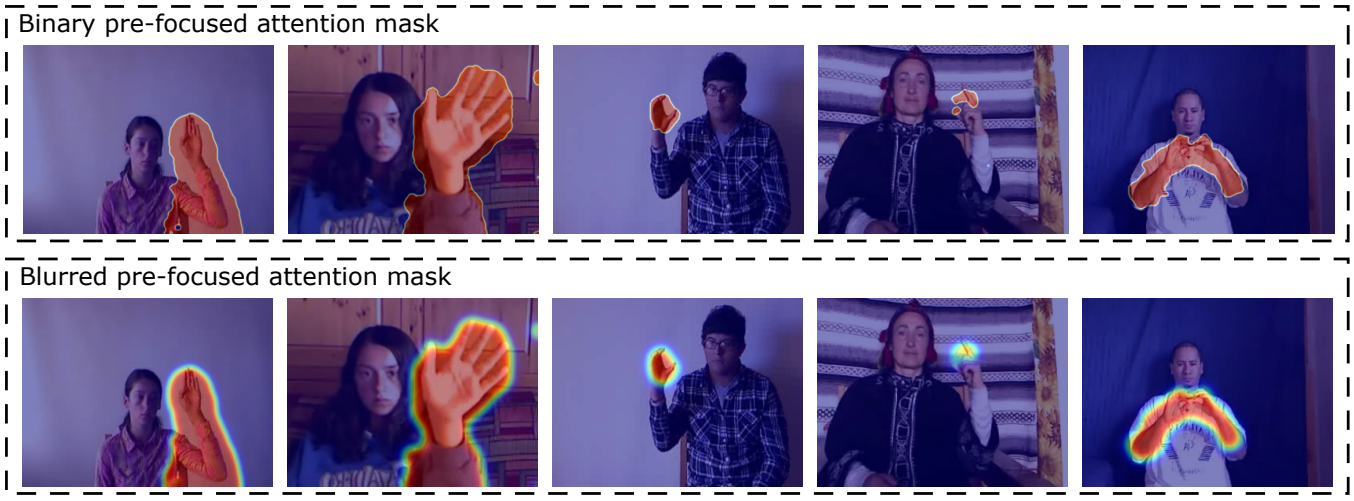

Fig. 4. **Visualization of Pre-focused Attention Maps.** An example of attention-weighted RGB frames from ChaLearn249 dataset [45] after applying pre-focused attention to the RGB frame. Top: binary masks, where the red areas represent motion-based attention areas and the blue areas show no motion, therefore less attention. Bottom: Blurring out the mask to allow for a smoother transition between the focus areas and the surrounding areas.

TABLE II

BINARY VS. BLURRED ATTENTION MASKS. COMPARISON BETWEEN USING A BINARY AND BLURRED MOTION ATTENTION MASK FOR THE PRE-FOCUSED ATTENTION APPROACH.

Method	**Validation accuracy**		**Test accuracy**	
	RGB	RGB + Flow	RGB	RGB + Flow
0.8 (binary)	57.54%	63.95%	59.96%	66.26%
0.8 (blurred)	**57.8%**	**64.21%**	**60.3%**	**67.11%**

to emphasize the hand and arm movement in the RGB image. Even in the fourth column, the minor movements of the index and thumb fingers are emphasized. The fifth column shows its success when the gesture also involves both hands.

B. Results of Learned Attention

In this section, we present the results of adding an attention layer to the network. The weights of the attention layer are initialized with ones. This means that, initially, the entire RGB image has the same weights, i.e. no motion-based attention prior. This weights of the newly added attention layer are first frozen, while the top, randomly-initialized layers are being trained. Afterwards, the attention layer's weights are unfrozen, and are fine-tuned along with the entire network. In this set up, the network decides end-to-end what to focus on in the RGB image, without being given any priors.

We report the results of these experiments in Table III. We observe that this performs significantly better than pre-focused attention, where the focus of the image is directed only where there is motion. We believe that these results make sense because even though throughout a sign language gesture motion is key in directing where to look, or focus the attention, once the desired hand position is reached and the desired hand shape is made, the signer holds that position and shape without any motion. In other words, during the peak of the gesture, there is no motion. Hence, in the aforementioned approach, these "key frames" where the hand shape and position are the most clear, lower weight is given due to the lack of motion. However, the learned attention approach took almost twice the number of training epochs as the pre-focused attention to converge.

C. Results of Hybrid Attention

To combine both approaches, we employ the learned attention approach, but in this case, we initialize the weights of the attention layer with the best-performing optical flow attention masks from the pre-focused attention experiment. Similarly, the randomly-initialized top layers are first trained while the attention layer's weights are frozen. Afterwards, the entire network, including the attention layer, is fine-tuned. From the results we observe that combining both approaches in the hybrid attention performs better than letting the network completely learn which areas to attend more to. In addition, training when using the hybrid approach converged faster than the purely learned approach, without priors.

D. Comparison with State-of-The-Art

In Table IV, we compare our best results from both attention methods to current state-of-the-art results on SLR using ChaLearn249 Dataset. For most methods, only results on the validation set are available, as the ChaLearn249 dataset was part of a competition, during which the test set was not yet available. We consider the work by Sarhan et al. [40] to be the baseline to which we compare our results, since we base our model on their work. It is clear that both methods for integrating attention already outperform state-of-the-art methods. However, integrating the attention layer into the network and allowing the weights to be learned as in the hybrid approach shows superior performance. In comparison to [40], the RGB stream performs better by $4.4\,\%$, leading to an overall increase in performance from $62.09\,\%$ to $65.02\,\%$.

Authorized licensed use limited to: Bibliothekssystem Universitaet Hamburg. Downloaded on February 03,2023 at 15:09:58 UTC from IEEE Xplore. Restrictions apply.

TABLE III

RESULTS OF LEARNED & HYBRID ATTENTION: ACCURACY RESULTS OF ADDING AN ATTENTION LAYER TO THE I3D NETWORK. IN THE *learned attention* APPROACH, THE WEIGHTS OF THE ATTENTION LAYERS ARE INITIALIZED WITH ONES, WHILE IN THE *hybrid attention* THE WEIGHTS ARE INITIALIZED WITH THE PRE-FOCUSED ATTENTION WEIGHTS. IN BOTH CASES THE WEIGHTS OF THE ATTENTION LAYER ARE FINE-TUNED DURING TRAINING.

Method	**Validation accuracy**		**Test accuracy**	
	RGB	RGB + Flow	RGB	RGB + Flow
I3D-SLR [40] (*baseline*)	54.63 %	62.09%	57.73%	64.44%
0.8 (blurred)	57.8%	64.21%	60.3%	67.11%
Learned attention	58.52%	64.7%	61.05%	68.36%
Hybrid attention	**59.2%**	**65.02%**	**61.65%**	**68.89%**

While the overall performance of 2SCVN-Max [10] is higher than that of [40], we still surpass it by a difference of $2.3\,\%$. The results suggest that SLR indeed benefits from the addition of spatial attention using the motion prior.

TABLE IV

COMPARISON WITH STATE-OF-THE-ART METHODS ON CHALEARN249 ISOGD DATASET.

Method	**Validation accuracy**	
	RGB	RGB + Flow
ASU [34]	45.07%	N/A
SYSU_ISEE [24]	47.29%	N/A
3DDSN [10]	46.08%	N/A
XDETVP [55]	51.31%	N/A
2SCVN-Max [10]	45.65%	62.72%
I3D-SLR [40] *(baseline)*	54.63%	62.09%
Attn-I3D-SLR (pre-focused)	**57.8%**	**64.21%**
Attn-I3D-SLR (learned)	**58.52%**	**64.7%**
Attn-I3D-SLR (hybrid)	**59.02%**	**65.02%**

VI. CONCLUSION

In this paper, we proposed a new motion-based attention-based SLR model, Attn-I3D-SLR, for isolated SLR with motion as a prior. The motivation behind that was to direct the network to focus on more discriminating areas than distracting ones. Our two-stream model uses inflated 3D ConvNets, and explicitly integrates spatial attention to the RGB stream. We investigated three methods for incorporating attention, which can serve as baselines and references for subsequent methods. We proved that using motion as a prior via pre-focused attention before training the network enhances performance significantly over passing the pure RGB image stream. We also showed that applying learned attention during training overcame the drawbacks of the pre-focused attention. Finally, we concluded that a hybrid approach, where the weights of an attention are initialized with pre-focused attention maps, and then fine-tuned during training performs best, as opposed to letting the network decide which areas to focus on from pure RGB images. Future work can possibly explore the effect of including further modalities, such as depth, which have proven beneficial to the sign language recognition tasks.

REFERENCES

[1] M. A. Ahmed, B. B. Zaidan, A. A. Zaidan, M. M. Salih, and M. M. b. Lakulu. A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017. *Sensors*, 18(7):2208, 2018.

[2] T. Bluche, J. Louradour, and R. Messina. Scan, attend and read: End-to-end handwritten paragraph recognition with MDLSTM attention. In *IEEE 14th International Conference on Document Analysis and Recognition*, 2017.

[3] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, June 2020.

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[5] W. Chan, N. Jaitly, Q. Le, and O. Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[6] Y. Chen, G. Ma, C. Yuan, B. Li, H. Zhang, F. Wang, and W. Hu. Graph convolutional network with structure pooling and joint-wise channel attention for action recognition. *Pattern Recognition*, 103:107321, 2020.

[7] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

[8] S. Das, A. Chaudhary, F. Bremond, and M. Thonnat. Where to focus on for human action recognition? In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 71–80. IEEE, 2019.

[9] W. Du, Y. Wang, and Y. Qiao. RPAN: an end-to-end recurrent pose-attention network for action recognition in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3725–3734, 2017.

[10] J. Duan, J. Wan, S. Zhou, X. Guo, and S. Z. Li. A unified framework for multi-modal isolated gesture recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):1–16, 2018.

[11] D. M. Eberhard, G. F. Simons, and C. D. Fenning. Ethnologue: Languages of the world. https://www.ethnologue.com/subgroups/sign-language. Accessed: 05-05-2020.

[12] A. Galassi, M. Lippi, and P. Torroni. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[13] R. Girdhar and D. Ramanan. Attentional pooling for action recognition. *arXiv preprint arXiv:1711.01467*, 2017.

[14] D. Guo, W. Zhou, H. Li, and M. Wang. Hierarchical LSTM for sign language translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[15] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10327–10336, 2020.

[16] W. Guo, Y. Zhang, J. Yang, and X. Yuan. Re-attention for visual question answering. *IEEE Transactions on Image Processing*, 2021.

[17] J. Huang, W. Zhou, H. Li, and W. Li. Attention-based 3D-CNNs for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2822–2832, 2018.

[18] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li. Video-based sign

Authorized licensed use limited to: Bibliothekssystem Universitaet Hamburg. Downloaded on February 03,2023 at 15:09:58 UTC from IEEE Xplore. Restrictions apply.

language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
[19] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
[20] J. Ji, C. Xu, X. Zhang, B. Wang, and X. Song. Spatio-temporal memory attention for image captioning. *IEEE Transactions on Image Processing*, 29:7615–7628, 2020.
[21] H. R. V. Joze and O. Koller. MS-ASL: A large-scale data set and benchmark for understanding American sign language. *arXiv preprint arXiv:1812.01053*, 2018.
[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
[23] O. Koller, S. Zargaran, and H. Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In *CVPR*, 2017.
[24] B. Li, W. Li, Y. Tang, J.-F. Hu, and W.-S. Zheng. GL-PAM RGB-D gesture recognition. In *ICIP*, 2018.
[25] D. Li, C. Rodriguez, X. Yu, and H. Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *WACV*, 2020.
[26] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li. Transferring cross-domain knowledge for video sign language recognition. In *CVPR*, June 2020.
[27] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek. VideoLSTM convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018.
[28] J. F. Lichtenauer, E. A. Hendriks, and M. J. Reinders. Sign language recognition by combining statistical DTW and independent classification. *TPAMI*, 30(11):2040–2046, 2008.
[29] K. M. Lim, A. W. C. Tan, C. P. Lee, and S. C. Tan. Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. *Multimedia Tools and Applications*, 78(14):19917–19944, 2019.
[30] Y. Lin, X. Chai, Y. Zhou, and X. Chen. Curve matching from the view of manifold for sign language recognition. In *ACCV*. Springer, 2014.
[31] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot. Global context-aware attention LSTM networks for 3D action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1647–1656, 2017.
[32] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. P. Graf. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2018.
[33] K. Mahdikhanlou and H. Ebrahimnezhad. Multimodal 3D American sign language recognition for static alphabet and numbers using hand joints and shape coding. *Multimedia Tools and Applications*, 79:22235–22259, 2020.
[34] Q. Miao, Y. Li, W. Ouyang, Z. Ma, X. Xu, W. Shi, and X. Cao. Multimodal gesture recognition based on the RESC3D network. In *ICCV Workshops*, 2017.
[35] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *IJCV*, 126(2):430–439, 2018.
[36] J. Pu, W. Zhou, and H. Li. Iterative alignment network for continuous sign language recognition. In *CVPR*, June 2019.
[37] L. Quesada, G. López, and L. Guerrero. Automatic recognition of the american sign language fingerspelling alphabet to assist people living with speech or hearing impairments. *Journal of Ambient Intelligence and Humanized Computing*, 8(4):625–635, 2017.
[38] J. Rodriguez and F. Martínez. How important is motion in sign language translation? *IET Computer Vision*, 15(3):224–234, 2021.
[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.
[40] N. Sarhan and S. Frintrop. Transfer learning for videos: From action recognition to sign language recognition. In *ICIP*. IEEE, 2020.
[41] N. A. Sarhan, Y. El-Sonbaty, and S. M. Youssef. HMM-based Arabic sign language recognition using Kinect. In *IEEE 10th International Conference on Digital Information Management (ICDIM)*, 2015.
[42] B. Shi, A. M. Del Rio, J. Keane, J. Michaux, D. Brentari, G. Shakhnarovich, and K. Livescu. American sign language fingerspelling recognition in the wild. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 145–152. IEEE, 2018.
[43] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1227–1236, 2019.
[44] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE TPAMI*, 20(12):1371–1375, 1998.
[45] 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 *CVPR Workshops*, 2016.
[46] J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan. Learning visual relationship and context-aware attention for image captioning. *Pattern Recognition*, 98:107075, 2020.
[47] S. Wang, D. Guo, W.-g. Zhou, Z.-J. Zha, and M. Wang. Connectionist temporal fusion for sign language translation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1483–1491, 2018.
[48] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*. Springer, 2016.
[49] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*. PMLR, 2015.
[50] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai. STAT: Spatial-temporal attention mechanism for video captioning. *IEEE transactions on multimedia*, 22(1):229–241, 2019.
[51] S. Yan, J. S. Smith, W. Lu, and B. Zhang. Hierarchical multi-scale attention networks for action recognition. *Signal Processing: Image Communication*, 61:73–84, 2018.
[52] Q. Yang. Chinese sign language recognition based on video sequence appearance modeling. In *5th Conference on Industrial Electronics and Applications*. IEEE, 2010.
[53] F. Yasir, P. C. Prasad, A. Alsadoon, and A. Elchouemi. SIFT based approach on bangla sign language recognition. In *8th International Workshop on Computational Intelligence and Applications (IWCIA)*, pages 35–39. IEEE, 2015.
[54] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *Proceedings of the 29th DAGM Conference on Pattern Recognition*, pages 214–223. Springer, 2007.
[55] 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 *ICCV Workshops*, 2017.

Authorized licensed use limited to: Bibliothekssystem Universitaet Hamburg. Downloaded on February 03,2023 at 15:09:58 UTC from IEEE Xplore. Restrictions apply.