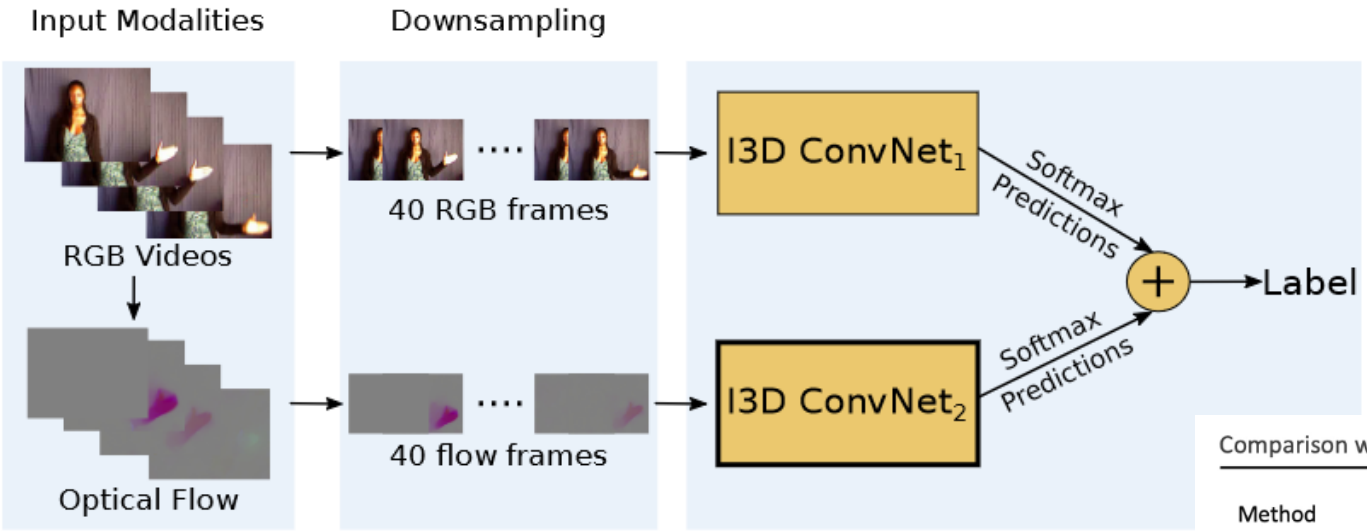
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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

Noha Sarhan and Simone Frintrop
University of Hamburg, Department of Informatics, Computer Vision Group,
Hamburg, Germany

Optical flow-based sign language recognition

Sarhan/Frintrop, ICIP 2020: a two-stream architecture (RGB & optical flow) using Inflated 3D networks (I3D)



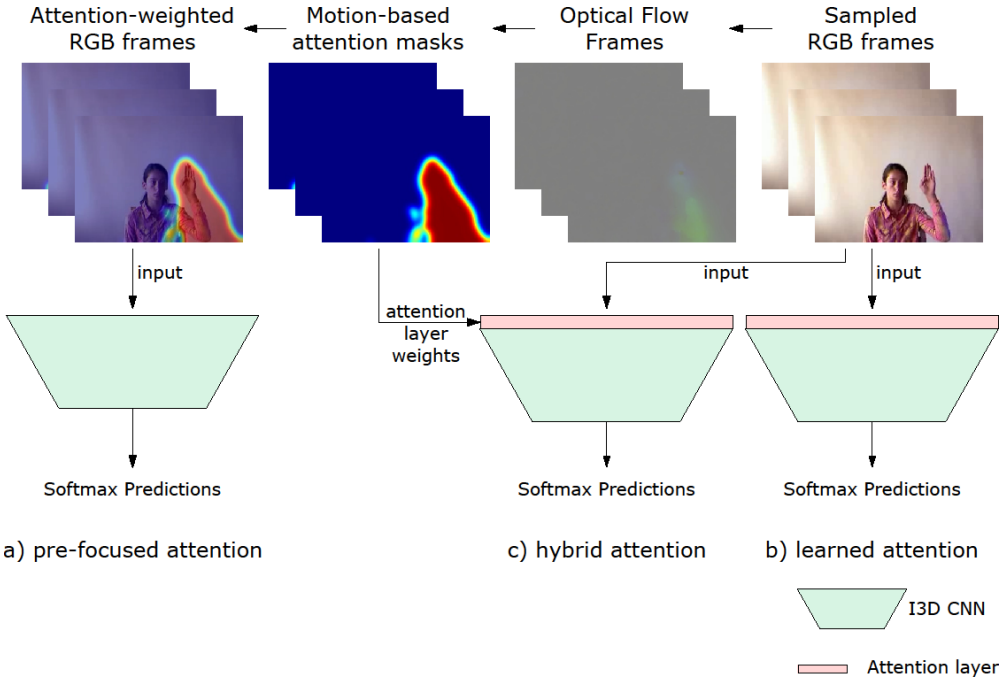
Comparison with state-of-the-art

Method	Accuracy		
	RGB	Flow	RGB + Flow
XDETVP[4]	51.31%	45.30%	
SYSU_ISEE[5]	47.29%	50.02%	
ASU[6]	45.07%	44.45%	
I3D-SLR (ours)	**54.63%**	**54.84%**	**62.09%**

[Carreira/Zisserman, "Quo vadis, action recognition? a new model and the Kinetics dataset," CVPR, 2017]
[Sarhan/Frintrop: „Transfer learning for videos: from action recognition to sign language recognition", ICIP 2020]

Attention-based sign language recognition

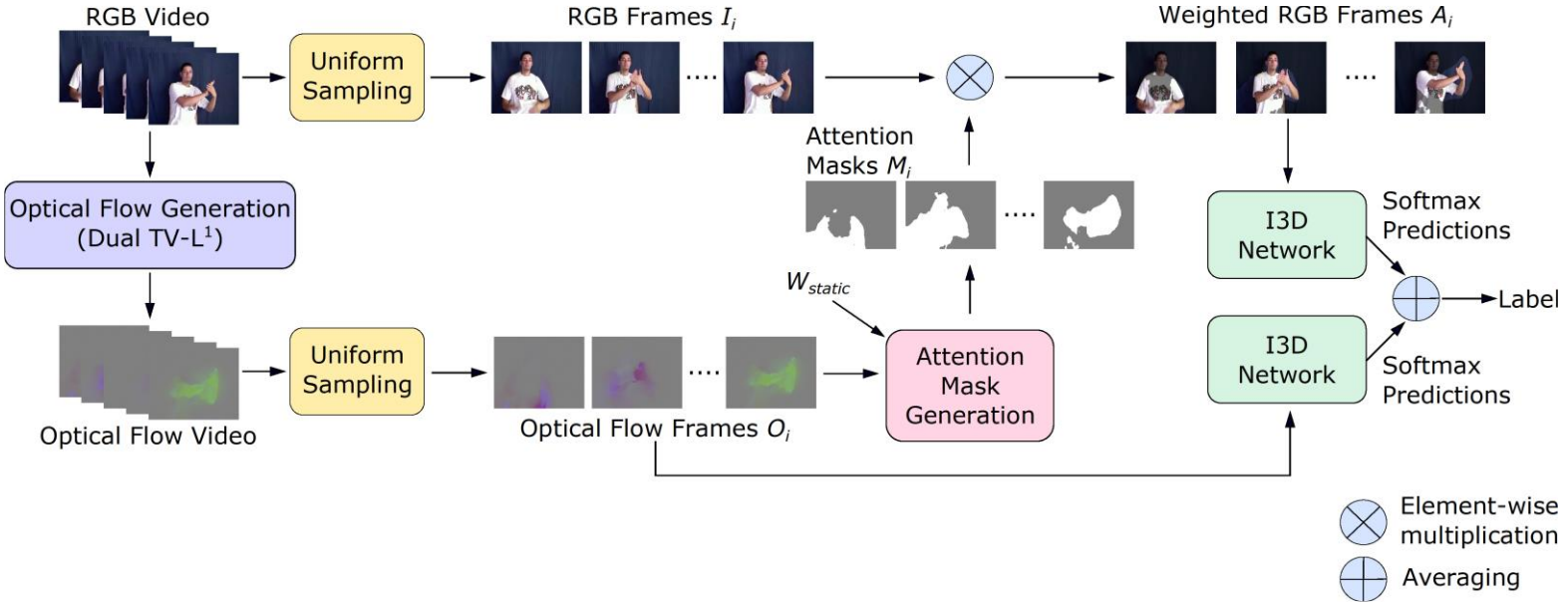
Sarhan/Frintrop: FG 2021:
3 approaches for integrating attention:
a) pre-focused, b) learned, c) hybrid



Noha Sarhan, Simone Frintrop: **Sign, Attend and Tell: Spatial Attention for Sign Language Recognition**, IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2021

Pre-focused attention

- We extend the baseline by adding a motion prior to focus attention of network on motion cues:
- an optical flow based motion mask is precomputed and used as attention prior

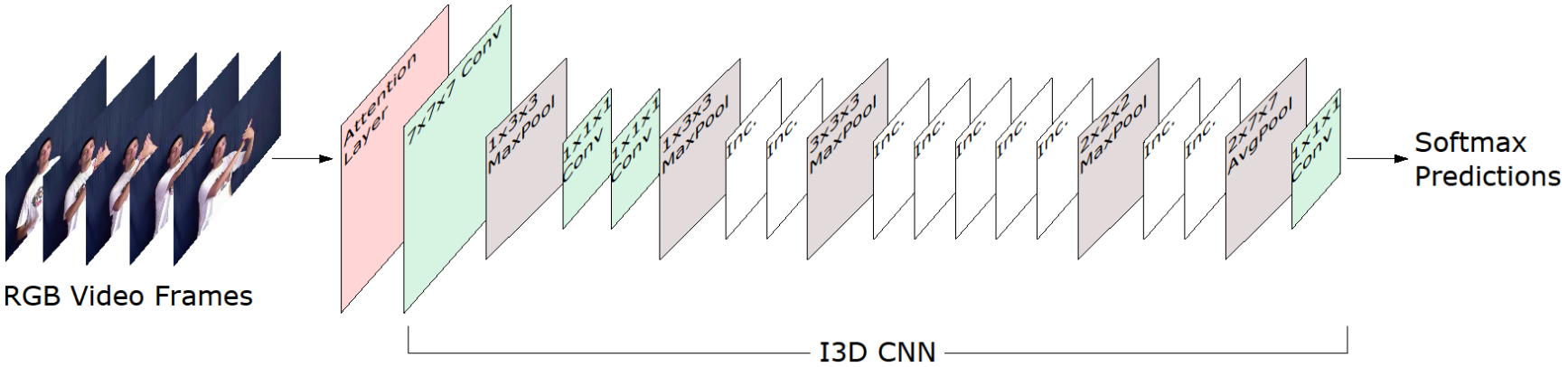


C. Zach, T. Pock, and H. Bischof, “A duality based approach for realtime TV-L1 optical flow,” in Proceedings of the 29th DAGM Conference on Pattern Recognition.Springer, 2007

Attention Maps

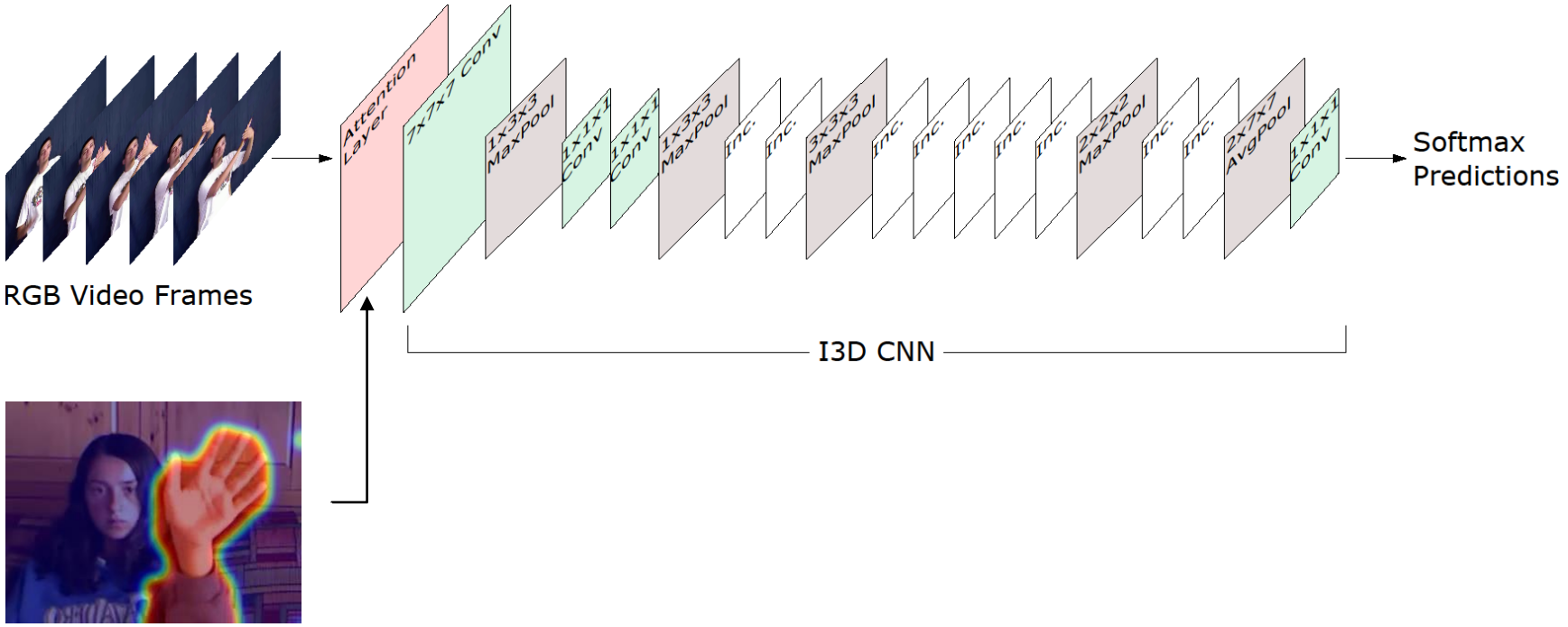
Learned Attention

An attention layer is
integrated into the I3D CNN to learn how to weigh the input



Hybrid Attention

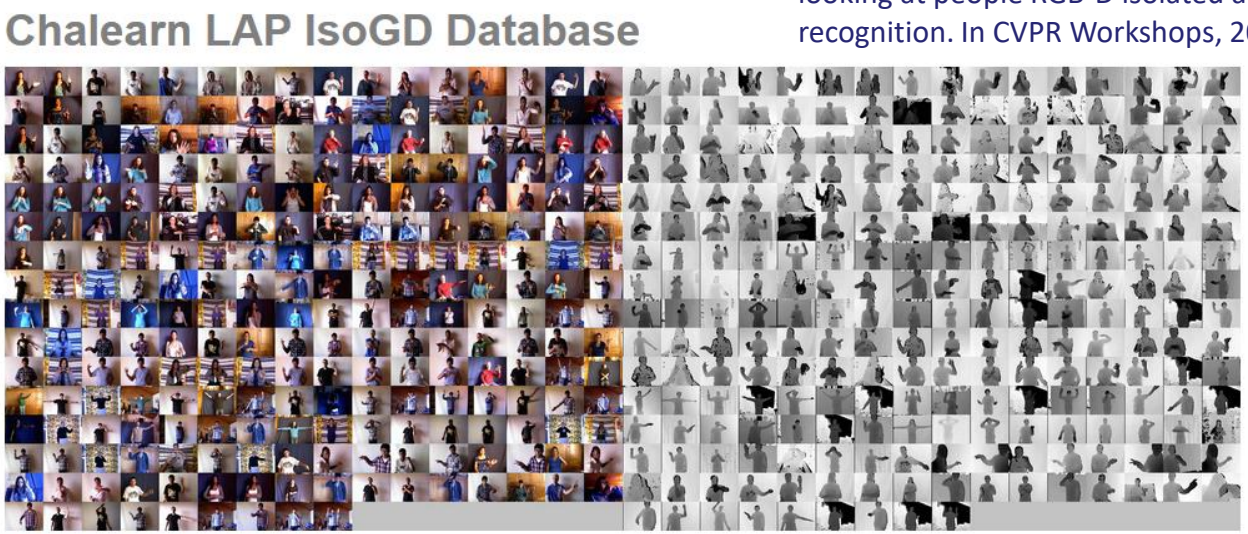
Same architecture as learned attention, but the attention map is initialized with the optical flow motion map from the pre-focused attention approach

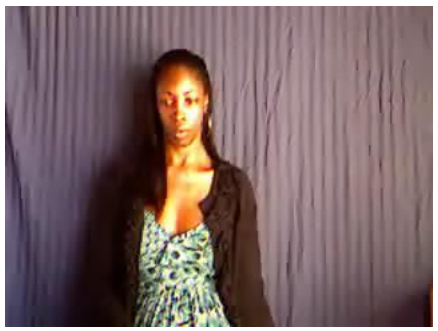


Dataset & Training

- **Dataset**: ChaLearn249 IsoGD, 47,933 videos of isolated sign language, 249 classes, 21 signers

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





- I3D networks were initialized with weights pre-trained on Kinetics dataset

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

Results

1. Hybrid attention clearly outperforms pre-focused & learned attention:

Method	Validation accuracy		Test accuracy	
	RGB	RGB + Flow	RGB	RGB + Flow
I3D-SLR [40] (*baseline*)	54.63 %	62.09%	57.73%	64.44%
Pre-focused attention	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%**

Results

2. All attention-based models outperform baseline and state of the art

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%**

Conclusion

- Sign language recognition models can strongly profit from motion cues and attention mechanisms
- Best approach: hybrid model which learns attention weights and is initialized with motion prior
- Future work: continuous sign language recognition

“Denmark, British, American, and Germany Sign Language by Deaf Furs” by WakeWolf