GASP: Gated Attention for Saliency Prediction

Fares Abawi, Tom Weber and Stefan Wermter

University of Hamburg

Objectives

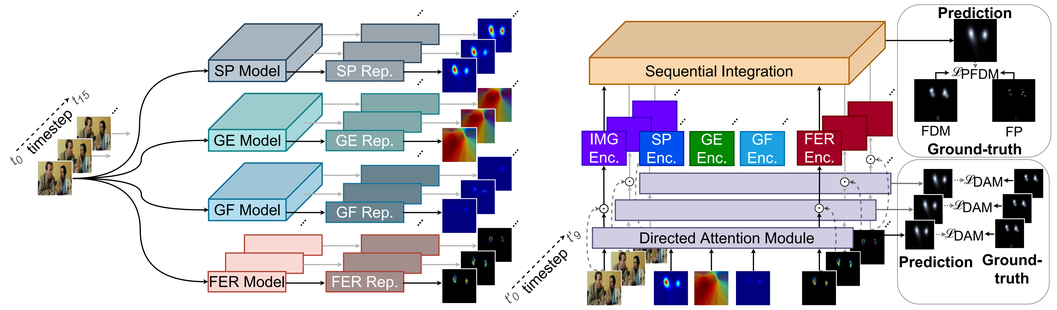
Top-down and bottom-up attention is influenced by social cues. We design a saliency prediction model to integrate these cues in dynamic social settings. Our approach is motivated by the following findings:

- Task-driven strategies are pertinent to predicting saliency. [1]
- Changes in motion contribute to the relevance of an object, underlining the importance of spatiotemporal features for predicting saliency. [2]
- Psychological studies indicate attention is driven by social stimuli. [3]

Overview

Saliency prediction (SP) refers to the computational task of modeling overt attention. Social cues greatly influence our attention, consequently altering our eye movements and behavior. To emphasize the efficacy of such features, we present a neural model for integrating social cues and weighting their influences. Our model consists of two stages. During the first stage, we detect two social cues by **following gaze** (GF), **estimating gaze direction** (GE), and **recognizing affect** (FER). These features are then transformed into spatiotemporal maps and propagated to the second stage (*GASP*), where we explore late fusion techniques for integrating social cues and introduce two sub-networks in the **Directed Attention Module** (DAM) for directing attention to relevant stimuli.

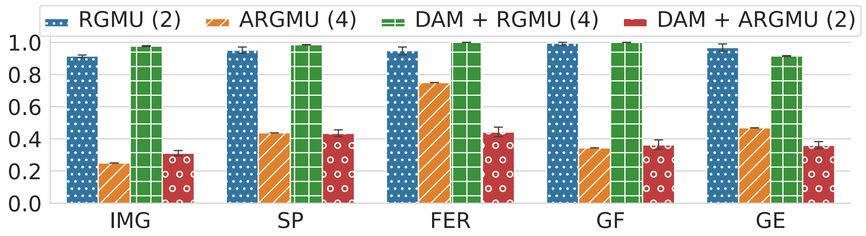
Model Architecture



In the first stage *(left)* we extract and transform social cue features to spatiotemporal representations (Rep.: Representation). GASP *(right)* acquires the representations and integrates encoded (Enc.: Encoder) features from the different modalities.

IMG: RGB Image; FDM: Fixation Density Map; PFDM: Predicted FDM; FP: Fixation Points.

Sequential Fusion



Gating weights of different sequential model variants showing the context size in parentheses. Introducing DAM allows modalities to have a uniform contribution. RGMU: Recurrent Gated Multimodal Unit; ARGMU: Attentive RGMU.

Social Cue Ablation

GE	GF	FER	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑
-	-	-	0.8767	0.6338	0.6542	2.72	0.5228
-	-	✓	0.7535	0.5951	0.4466	2.17	0.3578
-	✓	-	0.6893	0.5679	0.3222	1.84	0.2539
-	✓	✓	0.8778	0.6442	0.6652	2.76	0.5350
✓	-	-	0.8769	0.6272	0.6493	2.70	0.4798
✓	-	✓	0.8859	0.6505	0.6840	2.86	0.5381
✓	✓	-	0.8776	0.6367	0.6543	2.74	0.5216

Social cue ablation applied to our best GASP model (DAM + LARGMU; Context Size = 10).

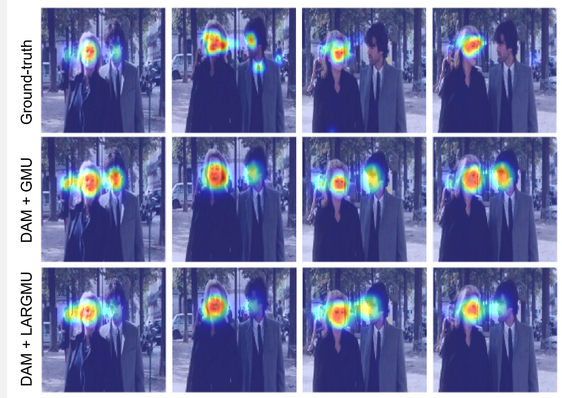
References

[1] Tom Foulsham, Jason JS Barton, Alan Kingstone, Richard Dewhurst, and Geoffrey Underwood.
Modeling eye movements in visual agnosia with a saliency map approach: Bottom–up guidance or top–down strategy?
Neural Networks, 24(6):665–677, 2011.

[2] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan.
A multimodal saliency model for videos with high audio-visual correspondence.
IEEE Transactions on Image Processing, 29:3805–3819, 2020.

[3] Brenda Salley and John Colombo.
Conceptualizing social attention in developmental research.
Social Development, 25(4):687–703, 2016.

Static & Sequential Models



DAM + GMU: Directed Attention Module followed by the Gated Multimodal Unit for static integration *(middle)*;

DAM + LARGMU (Context Size = 10): Directed Attention Module followed by the Late Attentive Recurrent GMU for sequential integration *(bottom)*.

Model	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑
DAM + GMU	0.8845	0.6397	0.6620	2.77	0.5233
DAM + LARGMU	0.8830	0.6527	0.6980	2.87	0.5566

Conclusion

We show that gaze direction and facial expression representations have a positive effect when integrated with saliency models, improving their prediction performances on multiple metrics: supports the importance of considering affect-biased attention.

Acknowledgements & Code

The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169).