The Impact of Temporal Regularisation in Egocentric Saliency Prediction

Panagiotis Linardos[1], Eva Mohedano[2], Monica Cherto[1], Cathal Gurrin[1], and Xavier Giro-i-Nieto[1]

[1] Universitat Politecnica de Catalunya, 08034 Barcelona, Catalonia/Spain
[2] Dublin City University, Glasnevin, Whitehall, Dublin 9, Ireland
linardos.akis@gmail.com, xavier.giro@upc.edu

Abstract. This work explores how temporal regularization in egocentric videos may have a positive or negative impact in saliency prediction depending on the viewer behavior.Our study is based on the new EgoMon dataset, which consists of seven videos recorded by three subjects in both free-viewing and task-driven set ups. We predict a frame-based saliency prediction over the frames of each video clip, as well as a temporally regularized version based on deep neural networks. Our results indicate that the NSS saliency metric improves during task-driven activities, but that it clearly drops during free-viewing. Encouraged by the good results in task-driven activities, we also computed and publish the saliency maps for the EPIC Kitchens dataset.

Keywords: Computer Vision · Egocentric Vision· Saliency · Visual Attention

1 Motivation

Saliency prediction refers to the task of estimating which regions of an image have a higher probability of being observed by a viewer. The result of such predictions is expressed under the form of a saliency map (heat maps), in which lighter values are aligned with the pixel locations with higher chances to attract the viewer's attention. This information can be used for multiple applications, such as a higher quality coding of the salient regions [22], spatial-aware feature weighting [15], or image retargeting [19]. This task has been extensively explored in set ups where the viewer is asked to observe an image [10, 7, 12, 2] or video [20] depicting a scene.

Our work focuses in the case of egocentric vision, which presents the particularity of having the viewer immersed in the scene. In this case the user is not only free to fixate the gaze over any region, but also to change the framing of the scene with the head motion. When collecting datasets, this set up also differs from others in which the same image or video is shown to many viewers, as in this case each recording and scene is unique for each user. Egocentric saliency prediction has received the attention of other researchers in the past [5, 18], which we extend by assessing a state of the art model in video saliency prediction in an egocentric set up. In particular, we observe that including a temporal regularization over frame-based prediction results into a gain or loss of performance depending on whether the viewer is engaged in an activity or is just free-vieweing the scene.

We developed our study on a new egocentric video dataset, named *EgoMon*, and added a temporal regularization on the SalGAN model [14] for image saliency prediction. Both the dataset and trained models are publicly available [3].

2 Related Work

2.1 Egocentric Video Datasets with Eye Tracking

The recording of an egocentric video dataset requires a wearable camera, but also a wearable eye tracker. This specificity in the hardware, together with the privacy constraints, limits the amount of public datasets for this task

The GTEA Gaze dataset was collected using Tobii eye-tracker glasses [5]. The more updated version of the dataset (EGTEA+) contains 28 hours of cooking activities from 86 unique sessions of 32 subjects. These videos come with audios and gaze tracking (30Hz) and provided with human annotations of actions (human-object interactions) and hand masks. In this work [5], saliency prediction models based on SVMs are trained separately for each activity, while in our case we train a single model and apply it to any activity.

The UT Ego Dataset [18] was collected using the Looxcie wearable (head-mounted) camera and contains four videos. Each video is 3-5 hours long, captured in a natural, uncontrolled setting. The videos capture a variety of activities such as eating, shopping, attending a lecture, driving, and cooking.

2.2 Deep Neural Models for Video Saliency Prediction

The recent success of deep neural networks for solving computer vision tasks has been recently explored in the context of video saliency prediction. These works have basically applied to this domain the architectures developed in the field of action recognition.

Two-stream networks [17] combining video frames and optical flow were applied in [1] for saliency prediction, while temporal sequences modeled with RNN [4] were explored for saliency in [11]. Our model is pre-trained on the DHF1K dataset[20], which contained 700 annotated videos for video saliency prediction. Its authors also trained a deep neural model based on ConvLSTM layers to predict the saliency maps. Similarly, the authors of [6] propose a complex convolutional architecture with four branches fused with a temporal-aware ConvLSTM layer. Regarding egocentric salienyc prediction with deep models, Huang *et al.* [9] propose to model the bottom-up and top-down attention mechanisms on the GTEA Gaze dataset. Their approach combines a saliency prediction with a task-dependent attention that explicitly models the temporal shift of gaze fixations during different manipulation tasks.

[3] https://imatge-upc.github.io/saliency-2018-videosalgan/

3 The EgoMon Gaze and Video Dataset

We propose *EgoMon*, a new egocentric gaze and video dataset. Data was recorded in Dublin by three different individuals wearing a pair of Tobii glasses equipped with a monocular eye tracker. The dataset is delivered as a collection of seven videos of an average length of 30 minutes.

EgoMon includes both *free-viewing activities* (a walk in a park, walking to the office, a walk in the botanic gardens, a bus ride), as well as *task-oriented activities* (cooking an omelette, listening to an oral presentation and playing cards). In the case of the botanic gardens, an additional a sequence of images captured every 30 seconds with a Narrative clip camera is also provided.

4 Model Architecture

The proposed architecture processes each video frame separately with SalGAN [14], an image-based saliency prediction pre-trained trained on the SALICON dataset [8]. SalGAN outputs a sequence of static saliency maps which were fed into a ConvLSTM [16] layer that we trained with the DHF1K dataset [20].

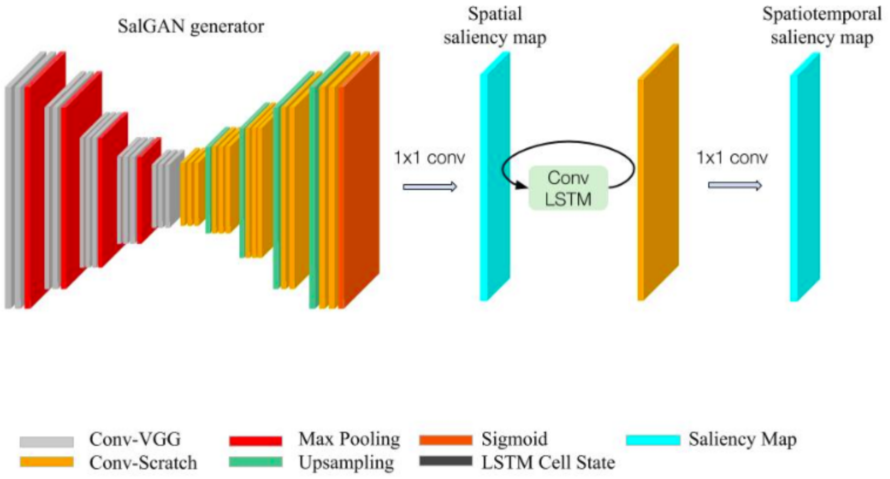

Fig. 1. Overall architecture of the proposed video saliency map estimation pipeline.

5 Experimentation

The proposed model was assessed firstly on the DHF1K dataset to evaluate its quality with respect to the state of the art in non-egocentric datasets, and later

Table 1. Performance on DHF1K.

	AUC-J ↑	sAUC ↑	NSS ↑	CC ↑	SIM ↑
static	**0.930**	**0.834**	**2.468**	**0.372**	**0.264**
dynamic	0.744	0.722	2.246	0.302	0.260
SoA [20]	0.885	0.553	2.259	0.415	0.311

Table 2. Performance on all datasets (NSS metric).

	DHF1K	EgoMon
static	**2.468**	**2.079**
dynamic	2.246	1.2472

on the proposed EgoMon dataset to draw our conclusions in the egocentric domain. We adopt standard saliency metrics when assessing our results on DHF1K dataset, and focus in NSS in the case of EgoMon.

To our surprise, the static (frame-based) SalGAN model outperformed the state of the art for on DHF1K [20] (Table 1). On the other hand, when considering our dynamic-aware model, where a ConvLSTM is trained on top of the frame-based saliency maps, the performance decreases.

This drop of performance of the dynamic model is also observed in some of the videos of the EgoMon dataset (Table 2). In particular, the static model performs better during free-viewing recordings, where saliency arises according to the intrinsic visual characteristics of a scene (bottom-up). Notice that DHF1K fixations were mostly recorded during free-viewing as well. On the other hand, our dynamic-aware model obtains better results than the static ones in those scenes where the used is engaged in an activity, in which task-driven saliency (top-down) [13] dominates.

Encouraged by these results, we have inferred the saliency maps pertaining to the Epic Kitchens object detection challenge [3]. We believe that these data can be valuable for third-party research focusing on other task such as object detection [15] or video summarization [21].

Table 3. Performance on different EgoMon tasks (NSS metric). Static refers to vanilla SalGAN, while Temporal refers to the augmented version.

	free-viewing recordings (bottom-up saliency)				
	bus ride	botanical gardens	dcu park	walking office	average
Static	**1.618**	**1.182**	**4.374**	**3.435**	**2.652**
Temporal	0.827	0.576	1.172	1.040	0.904
	task-driven recordings (top-down saliency)				
	playing cards	presentation	tortilla		average
Static	0.967	1.360	1.618		1.315
Temporal	**1.141**	**1.897**	**2.077**		**1.705**

References

1. Bak, C., Kocak, A., Erdem, E., Erdem, A.: Spatio-temporal saliency networks for dynamic saliency prediction. IEEE Transactions on Multimedia (2017)
2. Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
3. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: European Conference on Computer Vision (ECCV) (2018)
4. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
5. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **7572 LNCS**(PART 1), 314–327 (2012)
6. Gorji, S., Clark, J.J.: Going from image to video saliency: Augmenting image salience with dynamic attentional push. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7501–7511 (2018)
7. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Neural Information Processing Systems (NIPS) (2006)
8. Huang, X., Shen, C., Boix, X., Zhao, Q.: Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: IEEE International Conference on Computer Vision (ICCV) (2015)
9. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. arXiv preprint arXiv:1803.09125 (2018)
10. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (20), 1254–1259 (1998)
11. Jiang, L., Xu, M., Wang, Z.: Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm. arXiv preprint arXiv:1709.06316 (2017)
12. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (ICCV) (2009)
13. Murabito, F., Spampinato, C., Palazzo, S., Pogorelov, K., Riegler, M.: Top-Down Saliency Detection Driven by Visual Classification (2017)
14. Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X.: SalGAN: Visual Saliency Prediction with Generative Adversarial Networks (2017)
15. Reyes, C., Mohedano, E., McGuinness, K., O'Connor, N.E., Giro-i Nieto, X.: Where is my phone?: Personal object retrieval from egocentric images. In: Proceedings of the first Workshop on Lifelogging Tools and Applications. pp. 55–62. ACM (2016)
16. Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., Woo, W.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. CoRR **abs/1506.04214** (2015), http://arxiv.org/abs/1506.04214
17. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)

18. Su, Y.C., Grauman, K.: Detecting engagement in egocentric video. In: European Conference on Computer Vision. pp. 454–471. Springer (2016)
19. Theis, L., Korshunova, I., Tejani, A., Huszár, F.: Faster gaze prediction with dense networks and fisher pruning. arXiv preprint arXiv:1801.05787 (2018)
20. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting video saliency: A large-scale benchmark and a new model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4894–4903 (2018)
21. Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J.M., Singh, V.: Gaze-enabled egocentric video summarization via constrained submodular maximization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2235–2244 (2015)
22. Zhu, S., Xu, Z.: Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. Neurocomputing **275**, 511–522 (2018)