Temporal Saliency Adaptation in Egocentric Videos

Panagiotis Linardos[1], Eva Mohedano[2], Monica Cherto[2],
Cathal Gurrin[2], 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 adapts a deep neural model for image saliency prediction to the temporal domain of egocentric video. We compute the saliency map for each video frame, firstly with an off-the-shelf model trained from static images, secondly by adding a a convolutional or conv-LSTM layers trained with a dataset for video saliency prediction. We study each configuration on EgoMon, a new dataset made of seven egocentric videos recorded by three subjects in both free-viewing and task-driven set ups. Our results indicate that the temporal adaptation is beneficial when the viewer is not moving and observing the scene from a narrow field of view. Encouraged by this observation, we compute and publish the saliency maps for the EPIC Kitchens dataset, in which viewers are cooking.

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 higher values are aligned with those pixel locations with higher probabilities of attracting the viewer's attention. This information can be used for multiple applications, such as a higher quality coding of the salient regions [24], spatial-aware feature weighting [17], or image retargeting [21]. This task has been extensively explored in set ups where the viewer is asked to observe an image [10,7,13,2] or video [22] depicting a scene.

Our work focuses on 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 his 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 been studied in the past [5,20], a research line that we extend by assessing a state of the art model in image saliency prediction to this egocentric video set up. We developed our study on a new egocentric video dataset, named *EgoMon*, and added a temporal adaptation on the SalGAN model [16] for image

saliency prediction. We observe that the temporal saliency adaptation improves performance when the viewer is engaged in a task and with a narrow field of view, but, on the other hand, losses are measured when the viewer is simply free-viewing an open scene. Encouraged by these results, we have computed 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 [17] or video summarization [23]. Both the EgoMon dataset, Epic Kitchens saliency maps and trained models are publicly available [3].

2   The EgoMon Gaze and Video Dataset

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 availability of public datasets in this domain. 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. Similarly, the University of Texas at Austin Egocentric (UT Ego) Dataset [20] was collected using the Looxcie wearable (head-mounted) camera. It contains four videos, each video 3-5 hours long and captured in a natural, uncontrolled setting. The videos depict a variety of activities such as eating, shopping, attending a lecture, driving, or cooking.

In this work we introduce *EgoMon*, a new egocentric gaze and video dataset. Data was recorded in Dublin (Ireland) 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.

3   Deep Neural Models for Temporal Saliency Adaptation

Video saliency prediction with deep neural networks has basically adapted to this task the architectures proposed for video action recognition. Two-stream networks [19] combining video frames and optical flow were applied in [1] for saliency prediction, while temporal sequences modeled with RNN [4] were adopted in [11]. The authors of the largest dataset for video saliency prediction, the DHF1K (Dynamic Human Fixation 1K) dataset[22], 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 saliency prediction with

---

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

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 module, which explicitly models the temporal shift of gaze fixations during different manipulation tasks.

Our proposed architecture starts by processing each video frame separately with SalGAN [16], 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 two types of adaptation layers: 128 convolutional filters [14] of kernel size 3x3 and padding of 1, and its temporal-aware counterpart as ConvLSTM [18] with the same convolutional parameters. Their parameters were estimated from 700 training videos from the DHF1K dataset [22]. An SGD optimizer with 0.9 momentum was used, and the learning rate started at 0.00001 and decayed with a 0.1. There was also a weight decay of 0.0001.
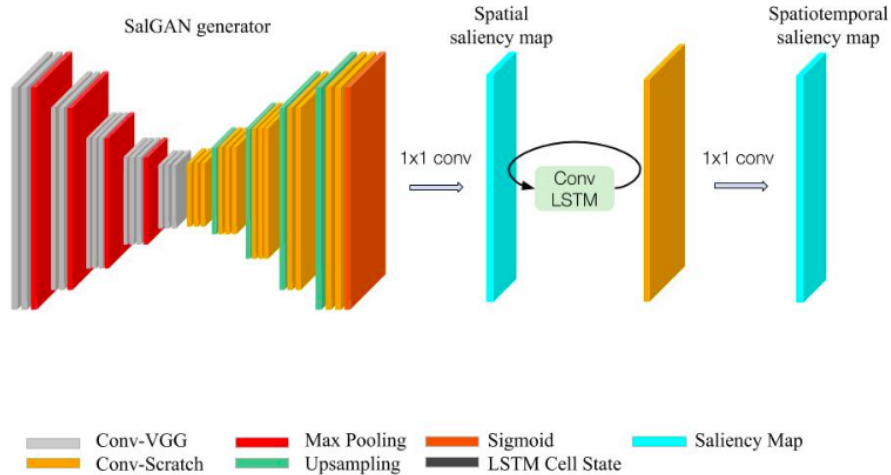

Fig. 1. Architecture of the dynamic model. The static model uses plain convolutions without the LSTM temporal recurrence.

4   Experimentation

The proposed model was assessed firstly on the same DHF1K dataset [22] the same from which the conv and convLSTM layers were trained. Afterwards, the model was assessed on the proposed EgoMon dataset to draw our conclusions in the egocentric domain.

Table 1 indicates that, surprisingly, the off-the-shelf (frame-based) SalGAN model [16] outperformed the state of the art model on the DHF1K [22] dataset. On the other hand, the quality of the prediction decreases when the conv or

Table 1. Performance on the DHF1K dataset.

	AUC-J ↑	sAUC ↑	NSS ↑	CC ↑	SIM ↑
SoA [22]	0.885	0.553	2.259	**0.415**	**0.311**
SalGAN [16]	**0.930**	**0.834**	**2.468**	0.372	0.264
+ conv	0.743	0.723	2.208	0.303	0.261
+convLSTM	0.744	0.722	2.246	0.302	0.260

Table 2. NSS metric across the DHF1K and EgoMon datasets.

	DHF1K	EgoMon
SalGAN [16]	**2.468**	**2.079**
+conv	2.208	1.2495
+convLSTM	2.246	1.2472

convLSTM layers are trained on top, which indicates that the domain adaptation is damaging the performance of the original SalGAN.

Table 3. Performance on different EgoMon tasks (NSS metric).

	free-viewing recordings (bottom-up saliency)				
	bus ride	botanical gardens	dcu park	walking office	AVERAGE
SalGAN [16]	**1.618**	**1.182**	**4.374**	**3.435**	**2.652**
+ conv	0.947	0.846	0.683	0.745	0.805
+ convLSTM	0.827	0.576	1.172	1.040	0.904
	task-driven recordings (top-down saliency)				
	playing cards	presentation	tortilla		AVERAGE
SalGAN [16]	0.967	1.360	1.618		1.315
+ conv	1.114	**1.966**	2.002		1.694
+ convLSTM	**1.141**	1.897	**2.077**		**1.705**

Table 2 indicates an even worse loss of performance when adding this adaptation layers in the EgoMon dataset. Nevertheless, the more detailed look provided in Table 3 that actually the adaptation layers are beneficial in those scenes where the user is engaged in an activity.

We hypothesize that this gap in performance is because the visual content in EgoMon is more similar to DHF1K than to SALICON [12], the dataset used to train SalGAN. SALICON was built from with a subset of images from the *Microsoft COCO: Common objects in context (MS COCO)* [15] dataset, a collection of images where the scene typically captures a wide angle with multiple objects and their surroundings. On the other hand, the task-driven egocentric activitites tend to show a narrow field with little context around the objects. Examples of these scenes could be contained in the DHF1K dataset, which was built by downloading videos from YouTube after a search by keyword.

5   Ackowledgements

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. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1072–1080 (2015)
13. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (ICCV) (2009)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
16. 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)
17. 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)
18. 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`
19. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
20. Su, Y.C., Grauman, K.: Detecting engagement in egocentric video. In: European Conference on Computer Vision. pp. 454–471. Springer (2016)
21. Theis, L., Korshunova, I., Tejani, A., Huszár, F.: Faster gaze prediction with dense networks and fisher pruning. arXiv preprint arXiv:1801.05787 (2018)
22. 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)
23. 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)
24. Zhu, S., Xu, Z.: Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. Neurocomputing **275**, 511–522 (2018)