Increasing Video Saliency Model Generalizability by Training for Smooth Pursuit Prediction

Mikhail Startsev, Michael Dorr
Technical University of Munich
{mikhail.startsev, michael.dorr}@tum.de

Abstract

Saliency prediction even for videos is traditionally associated with fixation prediction. Unlike images, however, videos also induce smooth pursuit eye movements, for example when a salient object is moving and is tracked across the video surface. Nevertheless, current saliency data sets and models mostly ignore pursuit, either by combining it with fixations, or discarding the respective samples. In this work, we utilize a state-of-the-art smooth pursuit detector and a Slicing Convolutional Neural Network (S-CNN) to train two saliency models, one targeting fixation prediction and the other targeting smooth pursuit. We hypothesize that pursuit-salient video parts would generalize better, since the motion patterns should be relatively similar across data sets. To test this, we consider an independent video saliency data set, where no pursuit-fixation differentiation is performed. In our experiments, the pursuit-targeting model outperforms several state-of-the-art saliency algorithms on both the test part of our main data set and the additionally considered data set.

1. Introduction

Saliency modelling can be beneficial for various computer vision and engineering applications [4]. In the case of dynamic stimuli, further consideration has to be given to the means of our perception: eye movements.

In general, humans shift their eyes in order to maintain a sufficient understanding of their constantly changing environment, since our eyes can see fine detail only where the image falls on a small part of the retina (ca. 1%) – the fovea. While image viewing mostly consists of fixations and saccades, video stimuli introduce smooth pursuit (SP) as well. This is a relatively slow (compared to saccades) motion of the eye, while the pursuit target is kept foveated [11]. Generally, SP cannot be performed without a target, setting it apart from fixations, which will be numerous even if the observer is presented with a blank screen. Additionally, several aspects of human perception are improved during SP [15, 18].

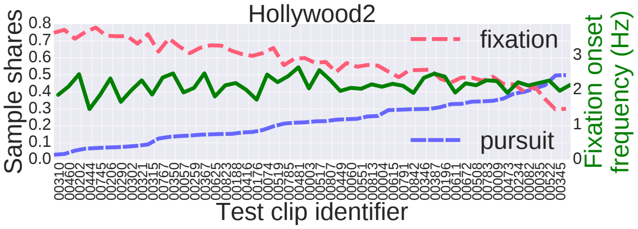

Figure 1. Saliency metrics typically evaluate against fixation onsets, detected by a traditional approach [6] (green line). These have an approximately equal frequency across videos. When a more principled approach to separating smooth pursuit from fixations [1] is applied, a great variation in the proportion of fixation (red line) and pursuit (blue line) emerges. The clips (on the x-axis) are sorted according to their pursuit sample shares.

SP has been largely neglected in automatic eye movement analysis, especially for natural scenes, where the stimulus and its potential moving SP targets are not known a priori. A recently introduced offline algorithm for SP detection [1] substantially improved performance for this challenging scenario. However, it requires the gaze traces of multiple observers for its operation, since the core idea is finding similarities of gaze trajectories of several observers at once.

The selectivity of pursuit can be visualized by examining the shares of all recorded gaze samples that are labelled as SP and as fixations, respectively. For a randomly selected subset of 50 Hollywood2 test videos, Figure 1 displays the fixation-SP balance, which varies greatly between different clips. We can hypothesize that SP is more stimulus-driven (since it requires a target), and the salient events that induce pursuit should “stand out” more. All this points to the need of systematically separating the two eye movement types in the context of saliency prediction and analysis.

In our work, we combined advances in automatic eye movement detection [1] with an existing large-scale data set (Hollywood2 [12]) to learn models that specifically target predicting either fixations or SP. We then tested our models against several recent literature models on a part of the test set and an independent data set (CITIUS-R [10]).

2. Related work

Saliency data sets in the literature ignore the issue of pursuit-fixation separation, either not mentioning SP [7, 12], or relying on eye trackers (which in turn do not consider SP) to detect fixations [2, 10]. DIEM [13] explicitly combines fixation and SP into generic *foveations*. GazeCom [6] uses a relatively simple approach to separate the fixations from SP, which is shown to be insufficient by the recently published manual annotations [19].

The saliency models, being developed in connection with certain data sets, similarly disregard SP: Not one of the literature models we encountered even mentions pursuit. The models themselves are usually separated into two groups: The bottom-up approach to video saliency is often explored via compression-domain algorithms (*e.g.* [9]), or more traditional pixel-domain ones (*e.g.* [10]). The top-down approach is represented by models that incorporate high-level object concepts explicitly [12], or rely on deep learning to implicitly learn those (*e.g.* [3]).

In this work, we train a recent deep learning architecture [16] to predict either fixations or pursuit and demonstrate that SP-oriented training has the potential to make resulting models more generalizable. Compared to the state of the art, our models show improved performance on both fixation- and SP-saliency prediction on Hollywood2, and a traditional saliency data set CITIUS-R.

3. Our approach

3.1. Data sets

We used the **Hollywood2** data set [12], since it is one of the largest video saliency sets that are publicly available, and it would be suitable as the source of training data for our deep model. This set contains ca. 5.5 hours of video data (training and test sets combined), viewed by 16 observers. The diverse clips contain camera motions, zoom level changes, and scene cuts. We used the full 823-clip training set for training (90%) and validation (10%). A random test subset of 50 clips (same as in Figure 1) was used for testing all the compared models. We resized all the clips to 640×360 for consistency.

We processed all the eye tracking recordings to detect pursuit and fixation samples with the toolbox in [19] (the implementation of [1]). The entire Hollywood2 training set contains a total of 4.5 million unique detected SP samples (*i.e.* coordinates within videos: frame number, x and y pixel coordinates) and over 10 million fixation samples.

Figure 2 displays an example scene from one of the data set clips, together with its empirical saliency maps for both fixations and smooth pursuit, and the same frames in saliency maps predicted by different models.

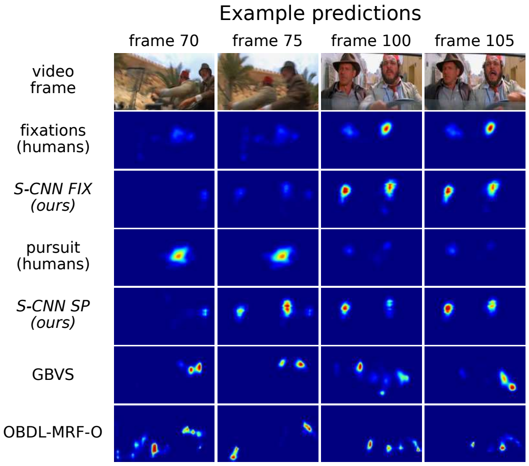

Figure 2. Frame examples from “actioncliptest00416” (1^{st} row), with respective empirical ground truth fixation-based saliency (2^{nd} row) and SP-based saliency (4^{th} row) frames. Dynamic frames (first two columns) are dominated by SP; relatively static frames (last two columns) mostly contain fixations. Predictions by our models are in the 3^{rd} and the 5^{th} rows, by GBVS and OBDL-MRF-O – in 6^{th} and 7^{th} rows. All predicted frame sequences are identically histogram-equalized for visual comparison.

CITIUS [10] contains both real-life stimuli and synthetically generated clips. In our evaluation, we only consider the real part (**CITIUS-R**; ca. 7 minutes, 22 observers), since we train for the data of the same domain. The eye tracking data contains only fixation data, but the stimuli are dynamic, with ample potential for SP targets.

3.2. Slicing CNN saliency model

We adopted the slicing convolutional neural network (S-CNN) architecture from [16]. In order to achieve temporal integration during video processing, this architecture rotates the feature tensors after initial individual frame-based feature extraction (see Figure 3). This way, time (frame index) is one of the axes of the network's subsequent convolution operations. This approach to sequence processing is an alternative to using handcrafted motion descriptors, 3D CNNs, or recurrent architectures.

We used one of the three branches of the whole network in [16], as memory constraints only allow training one branch at a time. We chose the branch where the temporal integration stage contains convolutions in the xt plane of the feature tensors. This xt-branch demonstrated the best individual results in [16], and the horizontal axis seems to be more important for human vision [14].

Our model performs binary classification of video RGB subvolumes $128\,\mathtt{px} \times 128\,\mathtt{px} \times 15\,\mathtt{frames}$, outputting the probabilities of the central pixel of the subvolume belonging

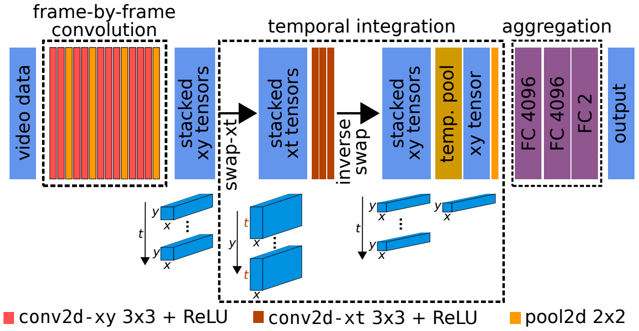

Figure 3. The xt branch of the S-CNN architecture for binary salient vs not salient video subvolume classification. The middle block is responsible for temporal integration: The three convolutional layers operate in the xt plane.

to the positive (salient) or the negative (not salient) classes.

The described model was trained twice: Either fixation or SP locations were considered salient. We sampled $50{,}000$ salient locations each time. The negative samples (also $50{,}000$) were drawn uniformly from the whole set of pixels of the training set (except for the locations used as salient class examples). This way, our pursuit-oriented model *S-CNN SP* and our fixation-oriented *S-CNN FIX* are trained under similar conditions to predict two different concepts. For validation, we used $10{,}000$ subvolumes: 5000 salient subvolumes and 5000 not salient.

The convolutional layers of our model were initialized with pre-trained VGG-16 [17] weights. We trained the models with a batch size of 5 for $50{,}000$ iterations with stochastic gradient descent (momentum of 0.9, learning rate 10^{-4}, divided by 10 every $20{,}000$ iterations).

The saliency maps were generated by taking the positive class probability of every 10^{th} pixel (along x and y axes), and up-scaling these low-resolution maps to 640×360.

4. Evaluation

As reference video saliency models, we use a pixel-domain GBVS [8], a range of compression-domain OBDL-models [9] (both through the framework provided by [9]), and AWS-D [10]. We tested ten OBDL model variations, but present the results only for the one performing best on both data sets (OBDL-MRF-O). On Hollywood2, we also evaluated the Mathe [12] model, which combines static (low-, mid- and high-level) and motion features.

We used three baselines, where possible – *Centre Baseline* (a square Gaussian reshaped to fit the aspect ratio of each video), *Permutation Baseline* (the "true" saliency map of another random video against the ground truth of the evaluated clip), and *One Human Baseline* (the empirical saliency map of one random observer vs. the overall saliency map of the same clip). In all cases, random selection was repeated five times. As CITIUS-R provides a set of fixation locations for all observers together, the latter baseline could not be tested there.

All empirical ground truth saliency maps were obtained by counting the amount of positive (*i.e.* either SP or fixation) samples in each pixel, and applying a Gaussian filter with the spatial σ equivalent to 1° of visual angle – the approximate size of the fovea. The temporal σ was set to a frame count equivalent of $1/3\,$s, so that the effect would be mostly contained within $3\sigma = 1\,s$ from the gaze sample.

We employ several typical metrics [5] that treat the saliency distribution either in a location-based fashion – AUC Borji, normalized scanpath saliency (NSS) – or as distributions – Kullback-Leibler divergence (KLD) and correlation coefficient (CC).

5. Results and discussion

First, we evaluate both *S-CNN FIX* and *S-CNN SP* on the 50-clip test subset of Hollywood2 (see Table 1). Our models outperform all the literature reference models, achieving highest results for all the reported metrics. Unsurprisingly, the models that were trained for fixation and SP prediction fare better in their respective domains. When we examine the models' performance on an independent CITIUS-R (see Table 2), however, our *S-CNN SP* consistently shows performance superior to that of the fixation-oriented *S-CNN FIX*, and both compare favourably to the state of the art.

We hypothesize that when a model is trained to predict pursuits, the learnt dynamic input signal properties can be more stably transferred to other data sets, where motion is important as well. Movement should augment both high- and low-level saliency, potentially making the models that explicitly learn to detect moving saliency areas more robust.

6. Conclusion

We have explored the possibility of taking eye movement class into account when dealing with saliency prediction. Our experiments and analysis show that (i) videos can have highly varying SP-fixation balances, so it is a factor that should be taken into consideration, and (ii) learning to predict a more selective and stimulus-driven eye movement – pursuit – can help model generalization.

Acknowledgements

Supported by the Elite Network Bavaria, funded by the Bavarian State Ministry for Research and Education.

References

[1] I. Agtzidis, M. Startsev, and M. Dorr. Smooth pursuit detection based on multiple observers. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research*

Model	Pursuit-saliency				Fixation-saliency				average rank
	AUC-Borji	NSS	KLD	CC	AUC-Borji	NSS	KLD	CC	
S-CNN SP	**0.91**	**1.97**	**2.38**	**0.28**	0.87	1.72	1.89	0.34	**2.38**
S-CNN FIX	0.90	1.95	2.47	0.27	**0.88**	**1.84**	**1.88**	**0.36**	**2.38**
Centre Baseline	0.87	1.72	2.24	0.27	0.85	1.64	1.72	0.34	3.00
One Human Baseline	0.82	3.66	8.32	0.41	0.87	3.21	5.45	0.48	3.25
GBVS	0.84	1.44	2.47	0.22	0.81	1.26	1.96	0.26	4.75
OBDL-MRF-O	0.81	1.40	2.70	0.19	0.77	1.22	2.22	0.24	6.12
Mathe	0.75	1.47	9.29	0.18	0.73	1.27	9.11	0.22	7.12
AWSD	0.73	0.89	2.91	0.12	0.75	1.08	2.21	0.21	7.50
Permutation Baseline	0.73	0.64	10.56	0.10	0.76	1.20	7.00	0.21	8.50

Table 1. Saliency prediction results on Hollywood2, depending on the considered eye movement type. The rows are sorted by the average rank (last column). Best non-baseline performance for each metric is boldified. Baselines are represented by rows with grey background.

Model	AUC-Borji	NSS	KLD	CC	avg. rank
S-CNN SP	**0.85**	1.56	**1.24**	**0.45**	**1.50**
S-CNN FIX	0.84	1.54	1.29	0.43	2.50
GBVS	0.83	1.58	1.33	0.40	3.25
AWSD	0.79	**1.72**	1.46	0.41	3.50
Centre	0.81	1.30	1.31	0.40	4.50
OBDL-MRF-O	0.79	1.48	2.28	0.37	5.75
Permutation	0.73	0.67	5.10	0.20	7.00

Table 2. Saliency prediction results on CITIUS-R.

& Applications, ETRA '16, pages 303–306, New York, NY, USA, 2016. ACM.

[2] H. Alers, J. A. Redi, and I. Heynderickx. Examining the effect of task on viewing behavior in videos using saliency maps. In *Human Vision and Electronic Imaging*, page 82910X, 2012.

[3] L. Bazzani, H. Larochelle, and L. Torresani. Recurrent mixture density network for spatiotemporal visual attention. *CoRR*, abs/1603.08199, 2016.

[4] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, Jan 2013.

[5] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016.

[6] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10):28, 2010.

[7] H. Hadizadeh, M. J. Enriquez, and I. V. Bajic. Eye-tracking database for a set of standard video sequences. *IEEE Transactions on Image Processing*, 21(2):898–903, Feb 2012.

[8] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, pages 545–552, 2007.

[9] S. Hossein Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan. How many bits does it take for a stimulus to be salient? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[10] V. Leborán, A. García-Díaz, X. R. Fdez-Vidal, and X. M. Pardo. Dynamic whitening saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):893–907, May 2017.

[11] R. J. Leigh and D. S. Zee. Smooth pursuit and visual fixation. In *The neurology of eye movements*, volume 90, pages 188–240. Oxford University Press, USA, 2015.

[12] S. Mathe and C. Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1408–1424, July 2015.

[13] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, Mar 2011.

[14] K. G. Rottach, A. Z. Zivotofsky, V. E. Das, L. Averbuch-Heller, A. O. Discenna, A. Poonyathalang, and R. Leigh. Comparison of horizontal, vertical and diagonal smooth pursuit eye movements in normal human subjects. *Vision Research*, 36(14):2189 – 2195, 1996.

[15] A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner. Eye movements and perception: A selective review. *Journal of Vision*, 11(5):9, 2011.

[16] J. Shao, C.-C. Loy, K. Kang, and X. Wang. Slicing convolutional neural network for crowd video understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[18] M. Spering, A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner. Keep your eyes on the ball: Smooth pursuit eye movements enhance prediction of visual motion. *Journal of Neurophysiology*, 105(4):1756–1767, 2011.

[19] M. Startsev, I. Agtzidis, and M. Dorr. Smooth pursuit. `http://michaeldorr.de/smoothpursuit/`, 2016.