How Many Glances? Modeling Multi-duration Saliency

Camilo Fosco*[1], Anelise Newman*[1], Patr Sukhum[2], Yun Bin Zhang[2], Aude Oliva[1], and Zoya Bylinskii[3]

[1]Massachusetts Institute of Technology, [2]Harvard University, [3]Adobe Research
{camilolu, apnewman, oliva}@mit.edu, {psukhum, ybzhang}@g.harvard.edu,
bylinski@adobe.com

Abstract

Traditional models of visual saliency have ignored the temporal aspect of visual attention and have produced prediction maps at fixed viewing durations. As a result, current applications of saliency are rigidly tailored for a fixed viewing duration. To incorporate knowledge of viewing duration into saliency modeling, we collect the CodeCharts1K dataset, which contains viewing data at three durations on 1000 images from diverse computer vision datasets. Our analysis shows distinct differences in gaze locations at these time points and exposes recurring temporal patterns about which objects attract attention. We use these insights to develop a lightweight saliency model that simultaneously trains on data from multiple time points. Our Multi-Duration Saliency Excited Model (MD-SEM) achieves state-of-the-art performance on the LSUN 2017 Challenge with 57% fewer parameters than comparable architectures.

1 Introduction

How long an observer examines an image determines what they notice and what tasks they can complete. Despite this dependency of viewing behavior on time, most models of visual attention predict saliency at a fixed duration, e.g., by training on data collected with a viewing time of 3 or 5 seconds per image [1, 6, 9]. In this paper, we introduce the first saliency model that simultaneously outputs multiple saliency maps for different viewing durations. We propose an efficient crowdsourcing methodology for collecting human attention data at scale at several viewing durations. We use it to assemble CodeCharts1K, a dataset of 1000 images with viewing patterns at three durations (0.5, 3, and 5 seconds). Trained on this data, our Multi-Duration Saliency Excited Model (MD-SEM) achieves state-of-the art performance when evaluated at a single duration and outperforms other baseline models if they are trained to predict multiple durations.

2 Dataset of multi-duration attention

Building off [12], we developed the CodeCharts UI interface to capture human gaze data at precise viewing durations without requiring explicit eye tracking (Fig. 1). Participants first view an image for a fixed viewing duration. This is followed by a quickly-flashed jittered grid pattern of three-character alphanumeric codes (code chart). Participants self-report the last three-character code they saw when the grid vanished. By construction, participants report the region of the image they were looking at. We repeat these steps for dozens of images, shown in sequence in a single experiment.

As an initial experiment, we sampled 50 images from the OSIE dataset [13] and collected the gaze locations of 50 participants per image for each of 6 image durations: 0.5, 1, 2, 3, 4, and 5 seconds. We found that the gaze maps obtained at 0.5, 3, and 5 seconds were the most different from each other. Participants were also consistent in terms of where they looked in images at those durations. The gaze

1st Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM), NeurIPS 2019.

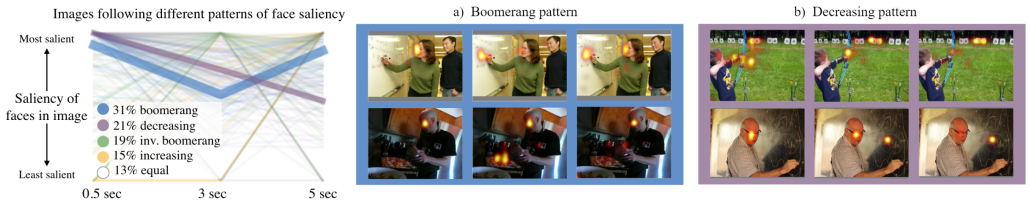

Figure 2: Dominant patterns of human gaze on faces across time: (a) face saliency decreases from 0.5 to 3 sec, and increases again from 3 to 5 sec ("boomerang"), and (b) face saliency decreases from 0.5 to 5 sec. Each line in the graph above represents how the saliency of faces within a single image varies over time, while the thicker lines represent an average over images. We normalize the saliency of faces across time on a per-image basis in this plot.

locations collected with our CodeCharts UI at 3 seconds most closely matched the ground-truth OSIE data, which was originally also collected at 3 seconds. This validates the ability of CodeCharts data to model the natural human gaze. We create our multi-duration dataset, CodeCharts1k, by collecting data at 0.5, 3 and 5 seconds for 1000 images sampled from different computer vision datasets.

3 Data analysis

Data consistency: To determine whether participants look in the same locations, we performed a split-half consistency analysis. We compared the gaze points of one half of the participants to the gaze points of the other half by converting both sets of gaze points to heatmaps and computing a Pearson's Correlation Coefficient between them, per viewing duration. By repeating this analysis *across* viewing durations, we also measured whether the gaze patterns are different for different viewing durations. Our analysis indicated that gaze data collected using our CodeCharts UI (i) contains a consistent signal at each of the viewing durations, (ii) this signal is equally strong at all the viewing durations, and (iii) the signal is significantly different between viewing durations. These findings motivated our computational model.

Face saliency at different times: It is known that gaze is attracted by faces [3, 4]. For a finer-grained analysis over time, we ran a face detection network [7] over all the images in CodeCharts1K. For the 303 images on which faces were detected, we computed face saliency by aggregating gaze counts per face region and then normalizing both by the number of gaze points per image and across all 3 durations. Fig. 2 plots these results as one line per image, where thick lines are averages to visualize the dominant patterns. We find a dominant "boomerang" pattern: people start out by looking at faces at 0.5 sec, their gaze shifts elsewhere at 3 sec, and returns to faces at 5 sec. The second most prevalent pattern is a decrease in gaze on faces over time.

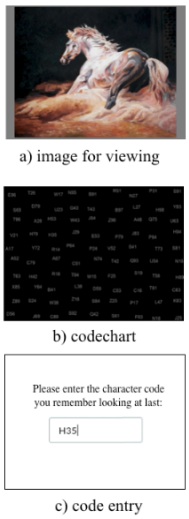

Figure 1: CodeCharts UI task flow: a) participants see an image followed by b) a quickly-flashed code chart; c) they self-report the three-letter code they saw last.

4 Multi-duration saliency model

In order to predict changes in human attention over time, we introduce the Multi-Duration Saliency Excited Model (MD-SEM), a new architecture adapted to multi-duration saliency. Our model is capable of producing saliency maps at T different viewing durations (T=3 here). The architecture is based on two concepts: (1) a powerful encoder-decoder architecture built on an ImageNet-pretrained Xception network [5], and (2) a Temporal Excitation Module, a novel block that applies a time-based re-weighting to saliency feature maps with a minimal increase in parameters. Our model (Fig. 3) achieves first place on the LSUN 2017 SALICON saliency challenge with 57% less parameters than SAM [6], and is the current state-of-the-art on CodeCharts1K, our multi-duration saliency dataset.

Network loss: Following previous works, the network's core loss is defined as a weighted combination of Kullback Leibler divergence (KL), Pearson's Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS) (see [2] for formulas) and a novel loss called Correlation Coefficient Match (CCM) that forces our network to reproduce differences and similarities between saliency maps at

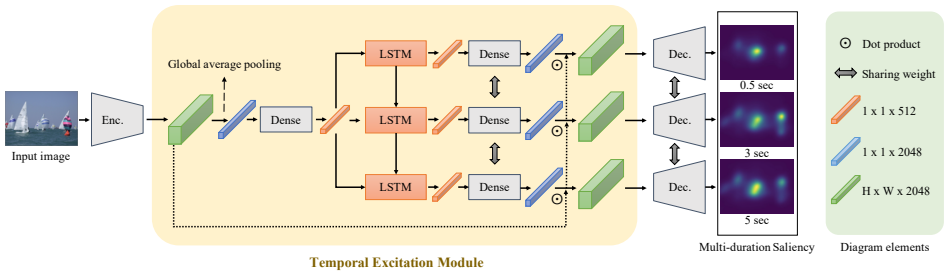

Figure 3: Multi-Duration Saliency Excited Model (MD-SEM) architecture. In order to predict saliency across durations, our novel module leverages LSTM cells to generate scaling vectors that re-weight the feature maps differently for each timestep.

Model	*NSS* ↑	*CC* ↑	*KL* ↓	*SIM* ↑	*CCM* ↓
SAM-MD w/o CCM	2.700	0.744	0.434	0.616	0.091
SAM-MD w/ CCM	2.765	**0.778**	0.401	**0.641**	0.073
MD-SEM w/o CCM	2.778	0.754	0.565	0.598	0.061
MD-SEM w/ CCM	**2.875**	0.773	**0.392**	0.633	**0.041**

Table 1: MD-SEM results on CodeCharts1K with and without CCM loss. Results are averaged across durations. NSS and CC are effective saliency metrics (they are symmetric to false positives and false negatives [2]).

different durations. We calculate the CCM loss by computing Pearson's Correlation Coefficient (CC) on pairs of saliency maps at adjacent durations, then computing the difference between the ground-truth and predicted CCs and averaging over all pairs of durations. We show in Table 1 that this novel loss significantly boosts model performance when evaluating multi-duration predictions.

5 Evaluation

On CodeCharts1K, our final model achieves an NSS of 2.875 and CC of 0.773 averaged across the three durations (Table 1). This is compared to an average ground-truth human score of 0.843 on the same dataset. Example predictions can be seen in Fig. 4. As our model is first-of-its-kind in its ability to predict multi-duration saliency, we benchmark against training T separate copies of a standard saliency network, SAM [6], one copy per viewing duration (SAM x3), as well as a modified version of SAM with our temporal excitation module (SAM-MD). The results of these benchmarks are in Table 2. Not only is MD-SEM better at approximating human gaze and differentiating across durations, but it also uses significantly fewer parameters than the other baselines. Finally, we show that our model performs at state-of-the-art for traditional single-duration saliency by achieving first place on the LSUN 2017 challenge (Table 2).

6 Applications

Compression and rendering: Just as saliency has been used as a mechanism to prioritize the visual content to preserve and render, multi-duration saliency can add a temporal aspect to these applications. For instance, if an image is expected to be viewed for shorter periods of time, fewer visual elements

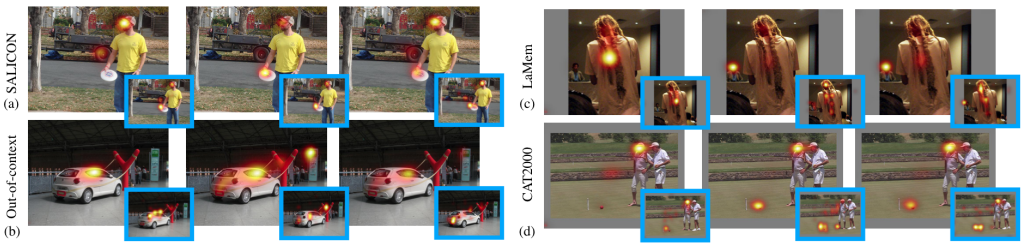

Figure 4: Saliency predictions of MD-SEM. Insets with blue borders contain ground-truth gaze locations collected using our CodeCharts UI. a) The "boomerang" pattern of saliency: starts out at face (0.5 sec), moves to object (3 sec), and back to face (5 sec). b) The boomerang pattern applied to objects: gaze starts at central object, moves to new salient location, then back to initial object. c) More distant objects (especially faces) become the objects of focus at later durations. d) Saliency distributes over time from faces to multiple scene elements.

Model	*NSS* ↑	*CC* ↑	*Params* ↓
SAM x3	2.020	0.803	210.3M
SAM-MD	2.057	0.792	70.1M
MD-SEM	**2.061**	**0.811**	**30.9M**

Model	*NSS* ↑	*CC* ↑	*KL* ↓
SAM-res [6]	1.990	**0.899**	0.610
EML-Net [8]	2.050	0.886	**0.520**
SalNet [10]	1.859	0.622	-
CEDNS	2.045	0.862	1.026
MD-SEM	**2.058**	0.868	0.568

Table 2: **Left:** Comparison to state-of-the-art models on SALICON-MD. Our model outperforms SAM-MD (SAM adapted by ourselves for multi-duration prediction) with 57% less parameters. **Right:** Comparison to state-of-the-art on SALICON test set (LSUN 2017 Challenge).

Figure 5: a) Visualized are image regions predicted to attract gaze at different viewing durations (accumulated over time). A better understanding of saliency across time can facilitate saliency-driven applications like image compression, transmission, and rendering to take viewing duration into account. b) Captions generated by passing saliency-enhanced images to an image captioning model [11], using saliency at different durations to prioritize image content.

need to be rendered (or preserved, in the case of compression). In Fig. 5a we provide a visualization of which visual content would be prioritized at different viewing durations for such applications.

Captioning: Our multi-duration saliency maps offer a closer approximation to how humans view images and provide an opportunity to focus attention on the most relevant regions for a given viewing duration. Here, we used our saliency predictions to focus an image captioning model [11] on image regions that should stand out at different viewing times. Removing the non-salient visual clutter produces promising initial results (Fig. 5b).

References

[1] Zoya Bylinskii et al. *MIT Saliency Benchmark*. `saliency.mit.edu/datasets.html`.

[2] Zoya Bylinskii et al. "What do different evaluation metrics tell us about saliency models?" In: *IEEE transactions on pattern analysis and machine intelligence* 41.3 (2019), pp. 740–757.

[3] Zoya Bylinskii et al. "Where should saliency models look next?" In: *European Conference on Computer Vision*. Springer. 2016, pp. 809–824.

[4] Moran Cerf, E Paxon Frady, and Christof Koch. "Faces and text attract gaze independent of the task: Experimental data and computer model". In: *Journal of vision* 9.12 (2009), pp. 10–10.

[5] François Chollet. "Xception: Deep learning with depthwise separable convolutions". In: *IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[6] Marcella Cornia et al. "Predicting human eye fixations via an lstm-based saliency attentive model". In: *IEEE Transactions on Image Processing* 27.10 (2018), pp. 5142–5154.

[7] Adam Geitgey. *Face Recognition*. `http://github.com/ageitgey/face_recognition`.

[8] Sen Jia. "Eml-net: An expandable multi-layer network for saliency prediction". In: *arXiv preprint arXiv:1805.01047* (2018).

[9] Ming Jiang et al. "Salicon: Saliency in context". In: *IEEE conference on computer vision and pattern recognition*. 2015, pp. 1072–1080.

[10] Junting Pan et al. "Shallow and deep convolutional networks for saliency prediction". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 598–606.

[11] Steven J Rennie et al. "Self-critical sequence training for image captioning". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7008–7024.

[12] Dmitry Rudoy et al. "Crowdsourcing gaze data collection". In: *Proceedings of ACM Collective Intelligence Conference* (2012).

[13] Juan Xu et al. "Predicting human gaze beyond pixels". In: *Journal of vision* 14.1 (2014).