This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Active Fixation Control to Predict Saccade Sequences

Calden Wloka Iuliia Kotseruba John K. Tsotsos
Department of Electrical Engineering and Computer Science
York University, Toronto, Canada
calden, yulia_k, tsotsos@cse.yorku.ca

Abstract

Visual attention is a field with a considerable history, with eye movement control and prediction forming an important subfield. Fixation modeling in the past decades has been largely dominated computationally by a number of highly influential bottom-up saliency models, such as the Itti-Koch-Niebur model. The accuracy of such models has dramatically increased recently due to deep learning. However, on static images the emphasis of these models has largely been based on non-ordered prediction of fixations through a saliency map. Very few implemented models can generate temporally ordered human-like sequences of saccades beyond an initial fixation point. Towards addressing these shortcomings we present STAR-FC, a novel multi-saccade generator based on the integration of central high-level and object-based saliency and peripheral lower-level feature-based saliency. We have evaluated our model using the CAT2000 database, successfully predicting human patterns of fixation with equivalent accuracy and quality compared to what can be achieved by using one human sequence to predict another.

1. Introduction

Most applications in computer vision function primarily in a passive way; algorithms are applied to static images or pre-recorded video sequences without control over what visual data is acquired next. However, it has long been recognized that eye movements are an integral aspect to human vision [33], with diverse functionality ranging from the enhanced extraction of features via microsaccadic motion [35] through high-level strategies for optimal information gathering [46]. It is this latter aspect which is of particular interest to the field of computer vision; active control over the acquisition of image data is fundamental to efficiently developing more robust and general computer vision solutions for unconstrained environments [57, 4].

Our work presented here develops and extends the Selective Tuning Attentive Reference model Fixation Controller (STAR-FC): an explicit model of human saccadic control [58]. In order to more easily compare with prior efforts in fixation prediction, we concentrate on the free-viewing paradigm, but nevertheless specify our control network in a manner which provides explicit extensibility for task-based tuning and top-down attentional control. By providing an extensive model of human fixation control which includes a number of aspects normally neglected by the

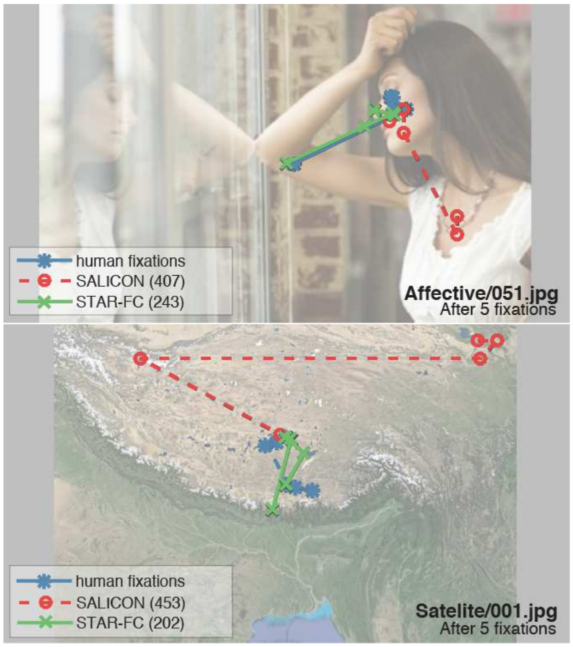


Figure 1: Two examples from the CAT2000 dataset with overlaid fixation sequences for the first five fixation points. The sequences predicted by our STAR-FC model are shown in green with X's marking the fixations, and SALICON predictions are shown in red with O's marking the fixation points. Human sequences (shown in blue) provided the closest match to both models. Euclidean distances between each model and the corresponding human sequence are noted in parentheses in the legend. Note that in both images, STAR-FC is significantly closer to human behavior than SALICON.

saliency literature, including an explicit transform to account for anisotropic retinal acuity, we are able to produce explicit fixation sequences with greater fidelity to those of humans than is seen by traditional saliency approaches (Figure 1). Altogether, we offer the following contributions:

1. A novel computational fixation model which outperforms traditional saliency map models for explicit sequences prediction.
2. A descriptive model of fixation control which may be used to better explore the function of early human attention.
3. A flexible design which may be parametrically tuned to match the specific experimental conditions under which eye tracking data is obtained.

1.1. Background

Eye movement control and early visual attention have frequently been conflated, particularly within the computational saliency literature. The term ”saliency map” was coined in [31] in the context of covert, pre-attentive visual processing. Due to the significant challenge of obtaining a suitable source of ground-truth data with which to validate a map of pre-attentive processing, focus shifted to predicting fixation locations [29]. Since then, many saliency algorithms have been proposed, ranging from information theoretic principles [10], efficient coding [21], spectral analysis [26], or processing pipelines driven largely by empirical performance [50], to name a few. One of the earliest machine learning efforts used a collection of low-, mid-, and high-level features as inputs to an SVM classifier in order to classify pixels as either salient or not [30]. More recently, however, deep learning networks have come to dominate the field [27, 36, 34, 41].

One schism which has formed within saliency research is whether the focus should be on locations or objects. Much of this split originated from the claim of Einhäuser *et al.* [15] that objects themselves actually predict fixations better than feature-based pixel saliency. This led to a number of approaches including those which seek to generate saliency maps based on class-generic object detectors (*e.g.* [3] and subsequent extensions to saliency [12]) and those which train and test saliency algorithms explicitly using object masks rather than fixation data (*e.g.*, [40]). However, there has been push-back against this object-centric view, with Borji *et al.* [7] arguing that the original findings of Einhäuser *et al.* were based largely on the metric used to measure performance. Given the focus of this paper on the explicit generation of saccade sequences, we test our algorithm performance against fixation data rather than object masks, but do take the view that there is a balance to be struck between pixel-level feature effects and higher-level object detection. This is discussed further in Section 2.1.

While our goal differs from the standard manner in which saliency algorithms are applied and evaluated, we compare performance against them in order to emphasize the importance of our novel perspective. Static saliency maps have previously been used to generate explicit fixation sequences, such as Itti and Koch’s [28] search system which couples Winner-Take-All (WTA) selection to a simple inhibition of return scheme. The connection between explicit eye movement patterns and saliency maps was explored from a different direction by [54], in which a saliency algorithm independent of the visual input was based on statistical regularities in eye movements. Despite the lack of visual processing, it nevertheless demonstrated comparable or better performance than the Itti-Koch-Niebur (IKN) saliency model [29], suggesting that fixation location may be driven as much by the underlying motor control of the eye as it is by visual information. Several efforts have been made to modulate a saliency map with a stochastic model, including a Levý Flight [8], a mixture of Gaussians learned from human fixation data [55], and a Markov process [43].

Outside of the saliency literature there are a number of eye movement control models. However, such models are usually dedicated to a specific subset of eye movements (e.g. smooth pursuit [48], the optokinetic reflex [14], or 3D gaze shifts [13]) or neural component (such as the role of the superior colliculus [62], cerebellum [44] or the basal ganglia [60]) without a clear path of extension or inclusion of attentional control. Tsotsos *et al.* [58] provide a more general formulation of attentional control with a focus on saccadic sequences. Nevertheless, the implementation of their model provides only a largely qualitative demonstration of efficacy over a single image. We build upon the theoretical formulation laid out by [58], extending the architecture to function over a broad range of natural images which allows for a quantitative analysis of performance. See Section 2.1 for a more thorough description of our model.

1.2. Applications of Fixation Prediction

Early interest in saccadic sequences was heavily influenced by Noton and Stark’s *scanpath theory* [47], which posited that the explicit spatiotemporal structure of eye movements drove memory encoding for visual patterns and subsequent retrieval. However, challenges to this view have arisen over the years, with experimental evidence showing that there is no recognition advantage conferred by the use of one’s own fixation locations versus those of another viewer nor by the retention of the temporal order of fixation [19]. These results certainly support the traditional approach to saliency evaluation which predominantly seeks to evaluate algorithms on prediction effectiveness over a static ground-truth fixation cloud, disregarding individual source and temporal characteristics of the fixations.

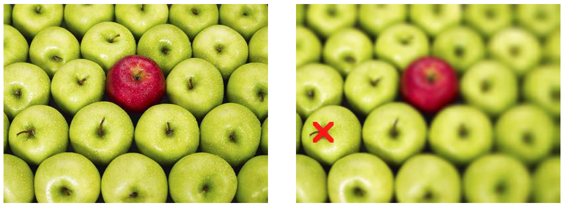
However, scanpath theory was largely devoted to the memory encoding and recall of images. Even if visual memory is not heavily influenced by scanpaths, there are nevertheless a number of applications for which explicit fixation sequence modeling and prediction is very valuable. For example, motivated by the very short window of consumer attention to most advertisements, commercial applications of saliency analysis already include predicted sequences of the first several fixations [2], despite validation using only traditional ROC methods which do not measure the efficacy of sequence prediction [1]. Understanding fixation locations has also gained recent interest in the area of science communication and policy making, particularly for graphical figures [25]. Even more so than in advertising, the sequence of fixations over a graphical figure becomes important for understanding whether and how viewers are understanding the information contained.

As previously mentioned, understanding the control of human eye movements may additionally be highly instructive in robotic visual systems with active camera control such as robotic search [49]. This is particularly useful for applications with anisotropic sensors which could be considered analogous to the anisotropy present within the human retina, such as omnidirectional camera systems which introduce a high degree of spatial distortion unevenly across the visual field [22] or a two-camera visual input system which combines high- and low-resolution streams to effectively maintain a wide field of view without sacrificing the ability to acquire high acuity detail over a targeted region [16]. Furthermore, as robotic applications increase their focus on social interactions, it becomes important not only to accurately attend to relevant information during an interaction, but also to provide socially important cues through body language such as gaze location [42]. Robotic modelling of joint attention has previously been improved through the application of saliency [65], and can likely be further improved with a more complete gaze model. Accurate modelling of joint attention between parties has wide reaching ramifications, from self-driving vehicles [32] to the handover of physical objects [45].

2. Methods

2.1. System Architecture

Our gaze control model extends and generalizes the approach initially taken by Tsotsos *et al*. [58]. Their original implementation provided much of the theoretical basis for the design of our model, but was only qualitatively tested against the seminal eye tracking work of Yarbus [64]. Without compromising the theoretical motivations of the previous work, we have modified the network structure to generalize across natural images and thereby allow quantitative testing of the model performance. Figure 3 provides a schematic of our implementation.



(a) Original image (b) With retinal transform

Figure 2: An example of the retinal transform: (a) original image; (b) fixated in the location marked with a red 'X'

The primary motivation of our architecture is to construct a set of interactive modules which introduce an iterative temporal component to fixation prediction (*i.e*. an active approach to perception). When humans visually explore an image, each fixation is made in the context of the prior fixations, introducing a confounding difficulty for any static map attempting to predict fixation locations passively. Although it has long been pointed out that saliency maps predict fixations with differing efficacy over time [53], static maps predicting a probabilistic distribution of the likelihood of any particular region being fixated remain standard practice in saliency research [38, 27, 37]. In order to better simulate the temporal dependence of fixation order, STAR-FC processes an input image iteratively through a chain of interacting modules:

1. Retinal transform: Based on the cone distribution from [23] and rod distribution from [61], we recreate the acuity field of the human eye through anisotropic blurring centered on the current fixation point. Each pixel in the image is sampled from the appropriate level of a Gaussian pyramid depending on the distance from fixation, increasing blur with distance from fixation (see Figure 2). Further details are provided in the supplementary material.
2. Central-peripheral split: To represent the different levels of cortical representation devoted to central versus peripheral processing, we split the image into two processing streams. Peripheral attentional capture is heavily dependent on low-level features, whereas central attentional capture is allocated at a higher level abstraction and tends to be more object-based (see [58] for justification). In the proposed architecture this is achieved by using a bottom-up algorithm based on low-level features (e.g. AIM [9], BMS [66], etc.) in the peripheral field and applying a CNN-based bottom-up saliency algorithm such as SALICON in the central field. The radius of the central attentional field is set to 12.5 degrees.
3. Conspicuity map: The central and peripheral processing streams are recombined into a single map; this is our closest correlate to the original concept of a saliency map [31]. Since there is no standard

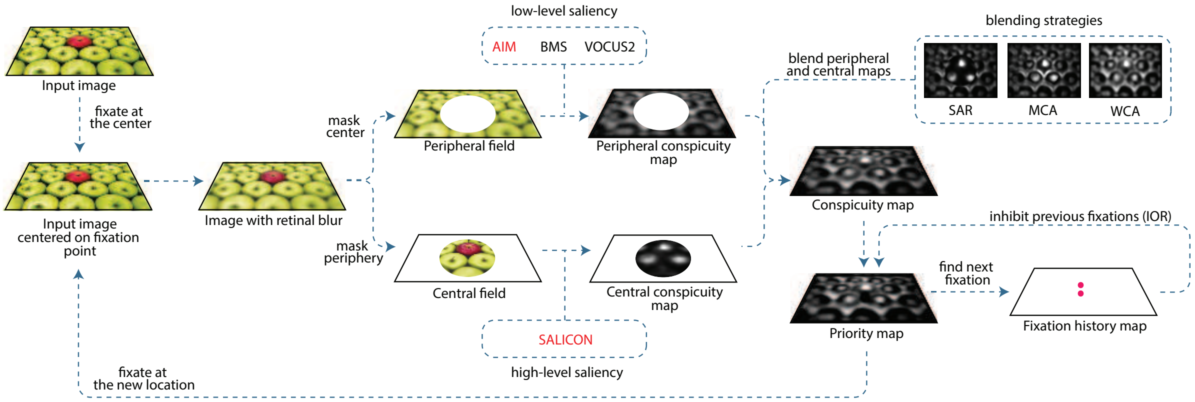


Figure 3: STAR-FC system architecture: Input images are first centrally fixated. A retinal transform is applied at the current fixation, and conspicuity is calculated within two streams: a peripheral stream which is dominated by low-level features, and a central stream which includes high-level and abstract features such as those learned by deep networks. The peripheral and central streams are then fused into a conspicuity map. The priority map combines the conspicuity map and input from a history map of all previous fixations (as well as any task-specific biases not further detailed here), providing an inhibition of return (IOR) mechanism. The next fixation point is selected from the maximum of the priority map, at which point the fixation is shifted to this new target location and the process repeats.

procedure for performing this integration, we experimented with three strategies (subsequently labeled in the text as STAR-FC_SAR, STAR-FC_MCA, and STAR-FC_WCA):

(a) *Separate Activation Regions* (SAR): A binary mask is applied to both the peripheral and central attentional maps to confine activations to their respective fields. A narrow overlap region is included within which the maximum value of either the peripheral or central activation is retained (originally proposed in [58]).

(b) *Maximum Central Activation* (MCA): The central attentional map is masked as in SAR, but no mask is applied to the peripheral map. The central region of the conspicuity map is equal to the maximum activation of either the peripheral or central maps.

(c) *Weighted Central Activation* (WCA): The peripheral and central attentional maps are combined as follows:

$$CM_{ij} = \begin{cases} \frac{r_c - r_p}{r_c} C_{ij} + \frac{r_p}{r_c} Pij, & \text{if } r_p < r_c \\ (1 + \frac{r_p - r_c}{r_{\max} - r_c})[1 - g_p]P_{ij}, & \text{otherwise} \end{cases} \tag{1}$$

where CM_{ij} is the conspicuity map value at pixel (i, j), C and P are the central and peripheral maps, respectively, r_c refers to the radius of the central field in pixels, r_p is the distance to the center in pixels and $r_{\max}$ is the maximum distance from the center in pixels. Here, an optional peripheral gain factor g_p is introduced to increase the importance of peripheral features most affected by the retinal transform.

4. Priority map: This combines the bottom-up activity of the conspicuity map with top-down spatial modulation, consistent with the proposed neural model of [18]. In our experiments this map only includes an inhibition of return (IOR) mechanism due to our focus on free-viewing. However, it could potentially be extended to incorporate other forms of modulation.
5. Fixation history map: This processing layer stores a history of previously fixated locations in image coordinates. These locations are inhibited with a circular zone of inhibition. Following [58] the radius of IOR is set to 1.5 degrees with suppression being maximal at the point of previous fixation and linearly decreasing towards the edge. IOR decays linearly within 100 fixations. In this paper IOR is applied by subtracting the fixation history map from the priority map.
6. Saccade control: This module is responsible for finding a new target within the priority map using a WTA scheme, shifting the gaze to a new location (by re-applying the retinal transform centered on the new fixation coordinates), as well as updating the fixation history map.

As mentioned, our work has been heavily influenced by the proposed control architecture in [58], but makes a number of important modifications and extensions. The original approach utilizes manually-derived face filters in the central field, specific to the single test image used for illustration. In order to generalize performance across natural images, we remove the custom face filters and instead incorporate, as part of the central field, a deep convolutional neural network (CNN), namely the SALICON saliency detection model [27]. In our implementation we use a C++ conversion of the OpenSALICON [56].

Our choice of using a CNN-based saliency algorithm is motivated by the idea that such saliency models can be viewed as processing incoming visual information analogous to a full forward pass through the visual hierarchy in order to produce high-level feature abstraction and object-based conspicuity allocation [39]. This is consistent with the theoretical aims of the central field put forth in [58]. SALICON was chosen due to the availability of an open-source implementation, but our formulation is agnostic to the specific saliency representations used in its construction.

Furthermore, we experiment with several bottom-up saliency algorithms to demostrate the effect of using different low-level features for computing peripheral attentional maps. In addition to AIM, which was also used in [58], we tested BMS [66, 67] and VOCUS2 [20].

Despite the fact that BMS significantly outperforms AIM on the CAT2000 dataset using the saliency metrics of the MIT Saliency Benchmark [11], when utilized in the peripheral component of STAR-FC both BMS and VOCUS2 achieve much worse fidelity to human fixation patterns compared to AIM, leading us to focus most of our tests on optimizing the AIM-based architecture. See the supplementary material for further information on the different STAR-FC variants we tested.

Finally, we define two additional strategies for combining the central and peripheral attentional maps aiming to alleviate the sharp border between the central and peripheral fields. This allows our architecture to more smoothly transition its activity across the visual field.

Although virtually any saliency algorithm can be used within the proposed architecture, both the choice of saliency algorithms for the central/peripheral fields and strategy for combining them have a dramatic effect on the produced fixation sequences. This will be discussed in more detail in Section 3.

2.2. Fixation Dataset

We evaluated model performance over the CAT2000 dataset[1] [6]. This dataset was chosen due to several positive attributes: it contains twenty different image categories (thereby representing a wide spectrum of visual stimuli), as well as one of the widest fields of view which we are aware of for a free-viewing eye tracking dataset (approximately 45°). Larger fields of view better approximate natural scene exploration, and are also likely to be more greatly impacted by considerations of retinal anisotropy and motoric bias than a comparable dataset gathered over a narrow field of view.

2.3. Evaluation Metrics

One major challenge in this work was determining the best method for evaluation. The output of our fixation control model is not directly comparable to that of saliency algorithms designed to predict human fixations, as we output a sparse set of explicitly predicted locations rather than a smooth map which can be treated as a probability distribution for likely fixation points over an image [38]. However, as mentioned in Section 1.2, there are applications for which an explicit sequence of fixation points is preferable to a probabilistic heat-map which lacks temporal structure.

Given that the innovation of our work rests on providing an explicit, temporally ordered fixation sequence rather than on a novel representation of saliency, we focus on evaluation metrics which reflect the spatiotemporal structure of sequences. In order to compare against the static maps which are the standard output of saliency algorithms, we sampled fixation sequences from the maps by applying an iterative WTA procedure. IOR was applied to each selected location using the same parameters as those of our fixation control model. This technique is consistent with previous work which samples loci of attention from saliency maps [28].

Although saccade amplitude distributions provide a relatively coarse measure with which to compare fixation sequences (as there is no representation of positional differences over the visual field), they do provide a representation of the motoric bias in the prediction. An early criticism of saliency algorithms was that they fail to account for inherent motor biases in how humans move their eyes [54], and it has been suggested that this motor bias could implicitly contribute to the persistent challenge of center bias in saliency research [63]. We therefore examine this aspect of model function in Section 3.1, demonstrating a much more human-like distribution of saccade amplitude with our model than is found from the predictions of sampled from static saliency maps.

To more explicitly explore the prediction performance of our model, we utilize trajectory-based scoring methods. These metrics focus on measuring the deviation between two spatiotemporal sequences. Trajectory comparison is a common problem in a wide range of fields, and can often rely on a number of different constraints or assumptions. Three common classifications of trajectory metrics are *network-constrained*, *shape-based*, and *warping-based* [5]. Network-constrained methods rely on an underlying path structure (such as a road network), and were therefore not appropriate for our purposes. However, both shape-based (which measure the spatial structure of trajectories) and warping-based (which take into account the temporal structure as well as the spatial) can provide meaningful in-

[1] A number of fixations included in the individual sequences of observers for the CAT2000 dataset were outside the bounds of the image. In order to prevent spurious comparisons with out of bound fixations while still ensuring cohesive sequences, we groomed the CAT2000 data by truncating any sequence which went out of bounds to the final in-bounds fixation location. If this truncation left the sequence with fewer than ten total fixations, it was discarded completely. Of 36000 total recorded fixation sequences, this criterion led to the elimination of 6257 sequences.

sight for saccadic sequences, and we therefore utilized the following set in order to provide a comprehensive sense of performance (trajectory-based score results are found in Section 3.2):

- *Euclidean Distance (ED)*: ED is one of the most common and basic warping-based trajectory metrics, and is calculated by matching two sequences in temporal order and computing the average pairwise distance between corresponding fixation points.
- *Fréchet Distance (FD)*: FD, also referred to as the 'dog-walking distance', represents the maximum distance at any given point in time over the length of two trajectories.
- *Hausdorff Distance (HD)*: HD is the maximum distance of a point in one sequence to the nearest point in a second sequence. Unlike ED and FD, HD is purely spatial and does not take sequence order into account.

3. Results

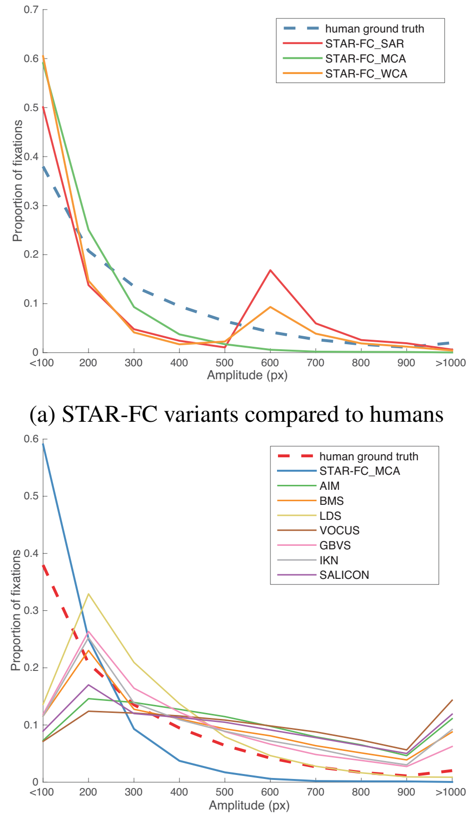
We compare the performance of our STAR-FC with a range of established saliency models: AIM [10], BMS [66], GBVS [24], LDS [17], SALICON [27, 56], SSR [51], and VOCUS2 [20]. For additional comparisons see the supplementary material. Results for saliency models modulated by motoric distributions [8, 55, 43] were not available for comparison on the CAT2000 at the time of publication.

3.1. Spatial Distributions

Saccadic amplitude distributions are shown in Figure 4. As can be seen in Figure 4a, the original central-peripheral integration strategy of Separate Activation Regions (SAR) used in [58] has a tendency to create a bimodal distribution not seen in the human data. This is likely due to the fact that the retinal anisotropy creates a biased gradient to the output of both the central and peripheral fields, meaning that near the border of the two the central field is weakest and the peripheral field is strongest. In order to facilitate a smoother transition of activation across the visual field, we tested two other integration strategies (described in Section 2.1): Maximum Central Activation (MCA) and Weighted Central Activation (WCA).

Our motivation to allow for the low-level feature representation of the peripheral map to affect the central region but not the other way around is based on the fact that there do appear to be fundamental perceptual limitations in object perception and feature binding within peripheral vision [52], whereas low-level features do seem to have a persistent role in attentional guidance [39].

Despite blending peripheral and central activations in a smoothly merging fashion, the WCA strategy leads to an activation pattern remarkably similar to the original SAR strategy. This is likely due to the fact that a weighted blending penalizes the chances of both algorithms within the mid-central region to attract attention unless they both happen to achieve a high score, essentially requiring a target to attract both high and low level attention simultaneously.

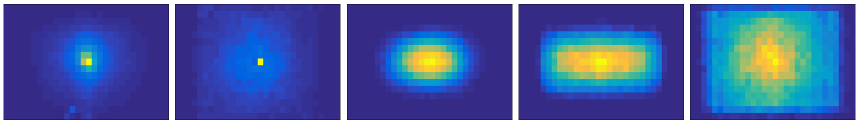


(a) STAR-FC variants compared to humans

(b) Traditional saliency algorithms, STAR-FC, and human distributions

Figure 4: Plots of the saccadic amplitude distributions over the CAT2000 dataset. Saccade lengths were assigned to bins of pixel ranges and the proportion of saccades falling in each bin are shown in the figures: (a) shows the effect of the different STAR-FC configurations on the resultant saccadic amplitude distribution (contrasted with the human distribution shown with a dashed line); (b) shows the distributions of traditional saliency algorithms contrasted with the MCA variant of STAR-FC and the human distribution.

The closest distribution pattern to that of humans was achieved by the MCA integration strategy, and it is therefore the variant reported in Figure 7 and Table 1. Although it does match the human distribution more closely than WCA and SAR variants, MCA appears to over-emphasize short saccades, having a much shallower tail than seen in the distribution of human observers. As previously mentioned, one likely contribution to this over-emphasis is the difficulty of many algorithms which have not been explicitly designed or trained to deal with signal degradation to function effec-



(a) Human (MSE Score)

(b) STAR-FC, (0.002)

(c) LDS, (0.027)

(d) GBVS, (0.072)

(e) SALICON, (0.122)

Figure 5: 2D histograms of fixation locations over the CAT2000 dataset. Mean-squared-error (MSE) scores between model and human distributions are shown in parentheses under each model name; as can be seen, STAR-FC is an order of magnitude closer to the human distribution than the closest competing saliency model.

Model	AUC ED	AUC HD	AUC FD	MSE
Human	632	844	1004	0
STAR-FC	**630**	**841**	**1006**	**0.002**
LDS	762	918	1067	0.027
GBVS	1068	1239	1415	0.072
BMS	1253	1447	1629	0.102
SALICON	1281	1471	1680	0.122
AIM	1313	1525	1758	0.161
VOCUS2	1347	1551	1781	0.183
SSR	1557	1755	1966	0.183
center	1875	2156	2156	0.008

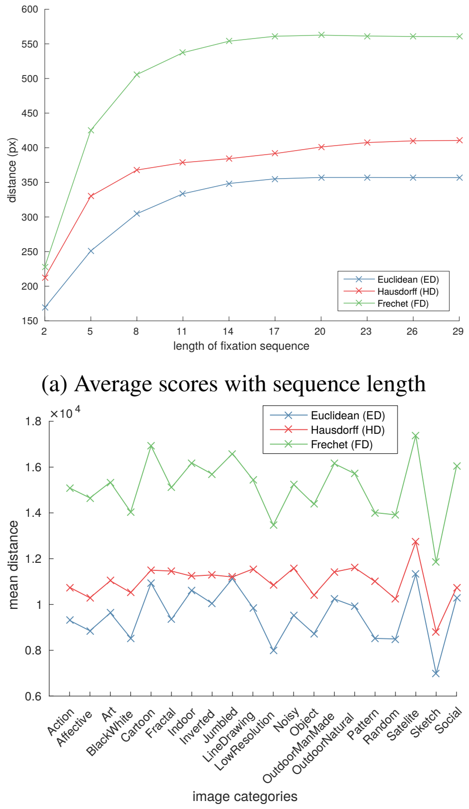
Table 1: Algorithm performance. Area-under-the-curve (AUC) scores are reported over the first five fixations for each plot in Figure 7. The last column shows the mean-square-error for the spatial histogram of predicted fixations versus the distribution of human fixations over the entire dataset. Note that our model (in bold) matches the inter-subject error of human observers.

tively across the retinal transform.

In contrast to the STAR-FC amplitude distributions, virtually all static saliency maps are skewed in the opposite direction with distributions which are much flatter than those seen with human data. Many algorithms do retain a small preference for shorter saccades, but this could also be an outcome of compositional bias in the underlying images. 2D histograms of fixation location produced with 64×64 sized blocks across the full CAT2000 dataset are shown for humans along with the MCA variant of STAR-FC and several representative saliency algorithms in Figure 5. As can be seen, there does appear to be a consistent spatial bias toward the center of the image which, at least in part, likely represents the underlying composition of the dataset images. Likewise, the saliency algorithms with the closest spatial distribution to the human distribution do tend to have a greater propensity for shorter saccades (Figure 4).

3.2. Trajectory Scores

Figure 6 shows the results of computing pair-wise scores across all combinations of human sequences for each image from the CAT2000 dataset. Figure 6a shows that the different trajectory metrics all tend to drift toward a saturated value; ED, FD, and HD all get larger as sequences diverge through time. Additionally, it has been shown that saliency tends to correlate best with early fixations [53], and both saliency correlation and inter-observer consistency degrade largely after the first five fixations. We therefore restrict our analysis to only this interval. Analysis of the full sequences may be found in the supplementary material.

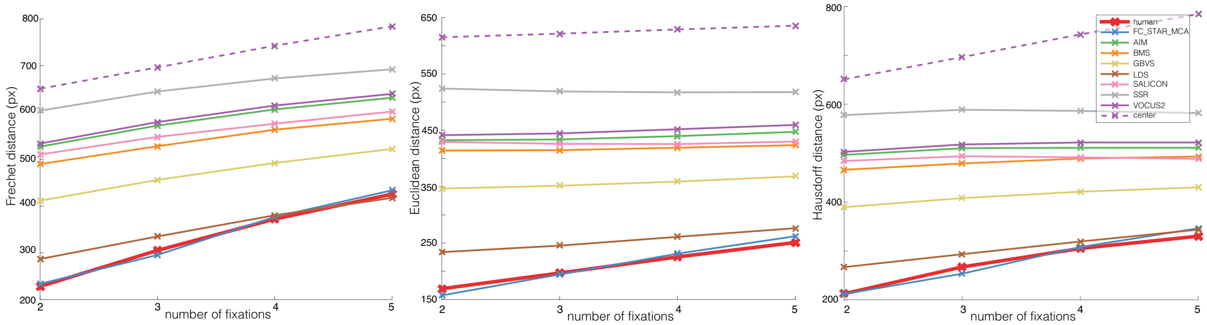


(a) Average scores with sequence length

(b) Average total sequence score by category

Figure 6: Average scores computed for all metrics over pair-wise matches of human sequences: (a) shows that as sequence length increases observer agreement tends to diverge, leading to a saturation in score values for each metric; (b) shows average sequence score per category, showing agreement with [6] about which categories tend to have greatest inter-observer consistency.

Figure 6b shows the category-wise average total



(a) Mean FD

(b) Mean ED

(c) Mean HD

Figure 7: A comparison of fixation prediction scores for static saliency maps and STAR-FC. A sequence formed by always picking the center pixel is shown in a dashed line to provide a performance baseline.

sequence scores per category. Here we can see that the trajectory metrics largely agree with the analysis in [6] on which categories have the greatest inter-observer consistency (Sketch, Low Resolution, and Black and White), and which categories tend to have poor inter-observer consistency (Satellite, Jumbled, and Cartoon).

We compare STAR-FC against a wide selection of saliency algorithms in Figure 7, showing that STAR-FC consistently achieves trajectory scores more in line with human sequences over the critical range of the earliest fixations, followed by LDS and GBVS (see Table 1 for numerical scores). In fact, STAR-FC is the only model which is able to achieve near-parity with the natural heterogeneity found within human observers.

As is made clear in Figure 5, human fixations over CAT2000 are strongly biased toward the center, a distribution which is well-matched by STAR-FC. The best performing saliency algorithms (LDS [17] and GBVS [24]) likewise have correspondingly stronger biases toward predicting fixations near the image center. We therefore also tested the ”center” model, which is simply a sequence which always selects the central pixel for every fixation. This selection will minimize the upper error bound for all trajectory metrics, and can be qualitatively thought of as a similar performance baseline to a centered Gaussian for more traditional saliency metrics [30]. Nevertheless, as Figure 7 shows, the center model consistently achieves the worst score in all metrics, confirming that while a centrally focused distribution of fixation locations is appropriate for the CAT2000 dataset, it is not a sufficient characteristic to score well.

4. Conclusion

Our Fixation Control model provides a powerful tool for predicting explicit fixation sequences. demonstrating fidelity to human fixation patterns equivalent to that of using one person’s fixation sequence to predict another. This performance is significantly better than what can be achieved by sequence sampling from static saliency maps (see Table 1). Our model will allow improved performance in saliency applications relying on explicit fixation prediction, including for commercial [2] and science communication [25] purposes. In addition to its performance, our model is also constructed to provide a descriptive model of fixation control, allowing further research into the interaction of the different cognitive control architectures which link gaze to higher order visual cognition [59].

While it is clear that retinal anisotropy has a significant effect on human visual performance, very few computational algorithms are developed with the aim of dealing with anisotropic acuity. This creates a significant challenge to accurately detect and ascribe conspicuity values across the visual field, and our model’s incorporation of retinal anisotropy represents an interesting platform for exploring this area of research.

Additionally, free-viewing over static images represents only a very narrow range of task for which fixation prediction provides valuable information. Fixation prediction over video and under task demands are highly challenging domains for which explicit fixation control may prove extremely valuable.

Acknowledgments

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Office of Naval Research (USA), the Air Force Office for Scientific Research (USA) and Canada Research Chairs Program. We would also like to thank Toni Kunić for technical assistance.

References

[1] 3M Commercial Graphics Division. *3M Visual Attention Service Validation Study*, 2010. 3

[2] 3M Visual Attention Service. *3M White Van VAS Sample Report*, 2015. Version 5.2. 3, 8

[3] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010. 2

[4] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos. Revisiting active perception. *Autonomous Robots*, 2017. 1

[5] P. Besse, B. Guillouet, J.-M. Loubes, and F. Royer. Review and perspective for distance based clustering of vehicle trajectories. *IEEE Transactions on Intelligent Transportation Systems*, 2016. 5

[6] A. Borji and L. Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581. 5, 7, 8

[7] A. Borji, D. N. Sihite, and L. Itti. Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data. *Journal of Vision*, 13(10):18, 2013. 2

[8] D. Brockmann and T. Geisel. The ecology of gaze shifts. *Neurocomputing*, 32-33:643 – 650, 2000. 2, 6

[9] N. D. Bruce and J. K. Tsotsos. Attention based on information maximization. *Journal of Vision*, 9(7), 2007. 3

[10] N. D. B. Bruce and J. K. Tsotsos. An information theoretic model of saliency and visual search. In E. R. L. Paletta, editor, *International Workshop on Attention and Performance in Computer Vision (WAPCV)*, pages 171–183, 2007. 2, 6

[11] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. http://saliency.mit.edu/. 5

[12] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *ICCV*, 2011. 2

[13] J. Crawford and E. Klier. Neural control of three-dimensional gaze shifts. In S. P. Liversedge, I. D. Gilchrist, and S. Everling, editors, *The Oxford Handbook of Eye Movements*, pages 339–356. Oxford University Press, 2011. 2

[14] C. Distler and K.-P. Hoffmann. The optokinetic reflex. In S. P. Liversedge, I. D. Gilchrist, and S. Everling, editors, *The Oxford Handbook of Eye Movements*, pages 65–83. Oxford University Press, 2011. 2

[15] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18, 2008. 2

[16] J. Elder, Y. Hou, R. Goldstein, and F. Dornaika. Attentive panoramic visual sensor, Oct. 31 2006. US Patent 7,130,490. 3

[17] S. Fang, J. Li, Y. Tian, T. Huang, and X. Chen. Learning discriminative subspaces on random contrasts for image saliency analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 28(5):1095–1108, 2017. 6, 8

[18] J. H. Fecteau and D. P. Munoz. Salience, relevance, and firing: a priority map for target selection. *Trends in Cognitive Sciences*, 10:382–390, 2006. 4

[19] T. Foulsham and A. Kingstone. Fixation-dependent memory for natural scenes: An experimental test of scanpath theory. *Journal of Experimental Psychology: General*, 142(1):41, 2013. 2

[20] S. Frintrop, T. Werner, and G. M. Garca. Traditional saliency reloaded: A good old model in new shape. In *CVPR*, 2015. 5, 6

[21] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, 2012. 2

[22] J. Gaspar, N. Winters, and J. Santos-Victor. Vision-based navigation and environmental representations with an omnidirectional camera. *IEEE Transactions on Robotics and Automation*, 16(6):890–898, 2000. 3

[23] W. S. Geisler and J. S. Perry. Real-time foveated multiresolution system for low-bandwidth video communication. In *Proc. SPIE*, volume 3299, pages 294–305, 1998. 3

[24] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, volume 19, pages 545–552, 2007. 6, 8

[25] J. Harold, I. Lorenzoni, T. F. Shipley, and K. R. Coventry. Cognitive and psychological science insights to improve climate change data visualization. *Nature Climate Change*, 6(12):1080–1089, 2016. 3, 8

[26] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34:194–201, 2012. 2

[27] X. Huang, C. Shen, X. Boix, and Q. Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, 2015. 2, 3, 4, 6

[28] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000. 2, 5

[29] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20:1254–1259, 1998. 2

[30] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009. 2, 8

[31] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985. 2, 3

[32] I. Kotseruba, A. Rasouli, and J. K. Tsotsos. Joint attention in autonomous driving (JAAD). *arXiv preprint arXiv:1609.04741*, 2016. 3

[33] E. Kowler. Eye movements: The past 25 years. *Vision Research*, 51:1457–1483, 2011. 1

[34] S. S. Kruthiventi, K. Ayush, and R. V. Babu. DeepFix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 2017. 2

[35] X. Kuang, M. Poletti, J. D. Victor, and M. Rucci. Temporal encoding of spatial information during active visual fixation. *Current Biology*, 22(6):510 – 514, 2012. 1

[36] M. Kümmerer, L. Theis, and M. Bethge. Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. In *ICLR Workshop*, 2015. 2

[37] M. Kümmerer, T. S. Wallis, and M. Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016. 3

[38] M. Kümmerer, T. S. A. Wallis, and M. Bethge. Information-theoretic model comparison unifies saliency metrics. *PNAS*, 112(52):16054–16059, 2015. 3, 5

[39] M. Kümmerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge. Understanding low- and high-level contributions to fixation prediction. In *ICCV*, 2017. 5, 6

[40] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *CVPR*, 2016. 2

[41] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In *CVPR*, 2015. 2

[42] N. Mavridis. *Grounded situation models for situated conversational assistants*. PhD thesis, Massachusetts Institute of Technology, 2007. 3

[43] O. L. Meur and Z. Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research*, 116:152 – 164, 2015. Computational Models of Visual Attention. 2, 6

[44] F. Miles. The cerebellum. In R. Carpenter, editor, *Eye Movements*, pages 224–243. CRC Press, 1991. 2

[45] A. Moon, D. M. Troniak, B. Gleeson, M. K. Pan, M. Zheng, B. A. Blumer, K. MacLean, and E. A. Croft. Meet me where I'm gazing: How shared attention gaze affects human-robot handover timing. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*, 2014. 3

[46] J. Najemnik and W. S. Geisler. Optimal eye movmement strategies in visual search. *Nature*, 434:387–391, 2005. 1

[47] D. Noton and L. Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311, 1971. 2

[48] J. Pola and H. J. Wyatt. Smooth pursuit: response characteristics, stimuli and mechanisms. In R. Carpenter, editor, *Eye Movements*, pages 138–157. CRC Press, 1991. 2

[49] A. Rasouli and J. K. Tsotsos. Visual saliency improves autonomous visual search. In *Canadian Conference on Computer and Robot Vision (CRV)*, 2014. 3

[50] N. Riche, M. Mancas, B. Gosselin, and T. Dutoit. RARE: A new bottom-up saliency model. In *IEEE International Conference on Image Processing (ICIP)*, pages 641–644, 2012. 2

[51] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009. 6

[52] H. Strasburger, I. Rentschler, and M. Jüttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11:1–82, 2011. 6

[53] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45:643–659, 2005. 3, 7

[54] B. W. Tatler and B. T. Vincent. The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17:1029–1054, 2009. 2, 5

[55] H. R. Tavakoli, E. Rahtu, and J. Heikkilä. Stochastic bottomup fixation prediction and saccade generation. *Image and Vision Computing*, 31(9):686 – 693, 2013. 2, 6

[56] C. L. Thomas. Opensalicon: An open source implementation of the salicon saliency model. Technical Report TR-2016-02, University of Pittsburgh, 2016. 4, 6

[57] J. K. Tsotsos. On the relative complexity of active vs. passive visual search. *International Journal of Computer Vision*, 7(2):127–141, 1992. 1

[58] J. K. Tsotsos, I. Kotseruba, and C. Wloka. A focus on selection for fixation. *Journal of Eye Movement Research*, 9:1–34, 2016. 1, 2, 3, 4, 5, 6

[59] J. K. Tsotsos and W. Kruijne. Cognitive programs: Software for attention's executive. *Frontiers in Psychology*, 5(1260), 2014. 8

[60] C. Vokoun, S. Mahamed, and M. Basso. Saccadic eye movements and the basal ganglia. In S. P. Liversedge, I. D. Gilchrist, and S. Everling, editors, *The Oxford Handbook of Eye Movements*, pages 215–234. Oxford University Press, 2011. 2

[61] A. B. Watson. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of Vision*, 14(7):1–17, 2014. 3

[62] B. White and D. Munoz. The superior colliculus. In S. P. Liversedge, I. D. Gilchrist, and S. Everling, editors, *The Oxford Handbook of Eye Movements*, pages 195–214. Oxford University Press, 2011. 2

[63] C. Wloka and J. Tsotsos. Spatially binned roc: A comprehensive saliency metric. In *CVPR*, 2016. 5

[64] A. L. Yarbus. *Eye Movements and Vision*. Plenum Press, 1967. 3

[65] Z. Ycel, A. Salah, C. Mericli, T. Mericli, R. Valenti, and T. Gevers. Joint attention by gaze interpolation and saliency. *IEEE Transactions on Cybernetics*, 43(3):829–842, 2013. 3

[66] J. Zhang and S. Sclaroff. Saliency detection: A Boolean map approach. In *ICCV*, 2013. 3, 5, 6

[67] J. Zhang and S. Stan. Exploiting surroundedness for saliency detection: A Boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 5(38):889–902, 2016. 5