Towards fine-grained fixation analysis: Distilling out context dependence

Conference Paper · March 2014

DOI: 10.1145/2578153.2578167

1 author:

Neil D. B. Bruce
University of Manitoba
46 PUBLICATIONS **1,103** CITATIONS

SEE PROFILE

Towards Fine-Grained Fixation Analysis: Distilling out Context Dependence

Neil D.B. Bruce*
Department of Computer Science, University of Manitoba

Abstract

In this paper, we explore the problem of analyzing gaze patterns towards attributing greater meaning to observed fixations. In recent years, there have been a number of efforts that attempt to categorize fixations according to their properties. Given that there are a multitude of factors that may contribute to fixational behavior, including both bottom-up and top-down influences on neural mechanisms for visual representation and saccadic control, efforts to better understand factors that may contribute to any given fixation may play an important role in augmenting raw fixation data. A grand objective of this line of thinking is in explaining the reason for any observed fixation as a combination of various latent factors. In the current work, we do not seek to solve this problem in general, but rather to factor out the role of the holistic structure of a scene as one observable, and quantifiable factor that plays a role in determining fixational behavior. Statistical methods and approximations to achieve this are presented, and supported by experimental results demonstrating the efficacy of the proposed methods.

CR Categories: I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Perceptual reasoning;

Keywords: eye tracking, saccades, gaze analytics, context

1 Introduction

Understanding gaze behavior has been a topic of much interest since the seminal work of Yarbus [Yarbus et al. 1967], and remains an important topic both within basic science, and various application domains. It is evident that there exist a multitude of factors that contribute to visual sampling of ones environment associated with either stimulus properties and context, task, or reward driven behavior in general [Chen and Zelinsky 2006; Hayhoe and Ballard 2005; Rothkopf et al. 2007]. Numerous models have attempted to predict fixation data through algorithmic means based on stimulus salience, however alternative efforts demonstrate the complexity of processes involved, and the importance of having a broader perspective on elements that factor into observed fixation behavior.

One area that stands to benefit from further investigation at the level of data analysis, is in attempting to attach meaning to fixations towards a representational construct that is more fine-grained than spatial or spatio-temporal coordinates. Some recent efforts view fixational behavior within a categorical construct, with fixations determined to be *ambient* vs. *focal* [Follet et al. 2011; Pannasch and Velichkovsky 2009], or corresponding to fixation towards the center-of-mass of a pool of stimuli [Findlay 1982; Clark 1999]. A variety of efforts have also sought to shed light on observed central bias of fixations [Tatler 2007; Tseng et al. 2009].

*e-mail:bruce@cs.umanitoba.ca

ETRA 2014, March 26 – 28, 2014, Safety Harbor, Florida, USA.

ACM 978-1-4503-2751-0/14/03 $15.00

The current work is a preliminary effort in moving towards the larger goal of attaching meaning to individual observed fixations. To this end, we aim to determine the extent to which observed fixations, or fixation density is consistent with what one would expect in considering the holistic structure or envelope of a scene. Consider for example a cross-walk where in one instance a person is beginning to cross a busy street and an identical view where a person is close to reaching the other side of the street. In each instance the holistic scene structure is identical, but in each case one might expect a relatively higher proportion of fixations in the vicinity of the person, independent of fixations made to other structure within the scene. Given a suitable construct for modeling the expected density of fixations as predicted by a coarse-grained model of scene structure, one should be able to determine that fixations observed in the vicinity of the person crossing in each instance, are due to something local within the scene.

To this end, in this manuscript we present a method for modeling the dependence between observed fixation densities, and holistic scene structure towards the goal of factoring out the contribution to observed fixation density that would be expected from holistic scene structure. In section 2, we describe the statistical and computational considerations involved in modeling the desired relationship. This includes the details of producing a holistic representation of a scene, and means of overcoming combinatorics that would imply the need for an exceedingly large data set in the absence of some of the strategies proposed. Section 3 demonstrates a number of results of context-independent fixation density maps as compared with standard determinations of fixation density. Finally, implications of the results are discussed, along with other fruitful avenues for fine-grained fixation analysis and broader implications of this line of investigation.

2 Methods

The following describes the representation employed to characterize the *holistic* nature of the scene, and statistical methods employed to characterize context independent fixation density.

2.1 Representing Context

Representing the holistic structure of a scene is a problem that has been explored in the context of computer vision for the purposes of scene categorization [Oliva and Torralba 2001] among other tasks. One means of producing such a representation that has appeared in various machine vision applications, and also in studies involving perception is the *Gist* model [Oliva and Torralba 2006]. This representation is derived in modeling spatial regularities over complete images to produce a low dimensional representation of the entire scene. This is accomplished by pooling the response of low level feature detectors over subregions of the image, and subsequent application of PCA to derive global receptive fields that comprise a holistic representation of a scene. This representation has been shown to be capable of categorizing the type of scene (indoor, outdoor, forest, city, etc.) [Oliva and Torralba 2006], and also useful in augmenting performance of models for predicting gaze locations [Torralba et al. 2006].

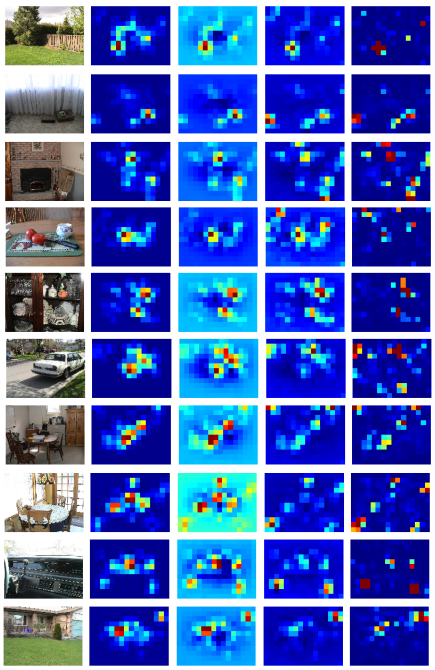

Figure 1: *Spatially quantized representations of fixation density. From left to right: Original Image, unprocessed fixation density, mean subtracted fixation density, whitened fixation density, context-independent density*

2.2 Statistical Methods

To derive a context independent determination of fixation density, we formulate the problem as: $p(D_{xy}|C, l_1, l_2, ..., l_n) \approx 1$ where D_{xy}[1] represents observed fixation density at position x, y, C is a vector representation of context given by the scene *Gist* and $l_1, ..., l_n$ are other latent factors (stimulus salience, task bias, reward seeking, etc.) that are assumed to contribute to observed fixation behavior. This further assumes that subject to knowledge of context, and a set of latent factors, that one is able to perfectly explain the observed fixation density. This does not however assume any knowledge of what these latent factors are, or how they are interrelated. Data from individual observers may be somewhat stochastic in nature, but as the sample size (participants in a user study) increases, any additional change in fixation density from a subsequent viewer is likely to be minimal. Equation 1 (specifically approximate equality to 1) reflects the assumption that density pooled across observers, is presumed to be well explained if one had access to complete knowledge of measurements of context, and a number of other unknown latent factors.

For a context independent representation of fixation density, we are interested in $p(D_{xy}|l_1, l_2, ..., l_n)$. The first assumption made in the following formulation, is that C and $l_1, ..., l_n$ can be treated as independent. In practice, there is likely to be some interaction between latent factors and context (e.g. rapid priming effects), however this assumption is a necessary approximation to disentangle context from a nondescript set of latent factors.

[1]By abuse of notation $p(D_{xy})$ refers to $p(D_{xy} = d_{xy})$ where d_{xy} represents the value of the observed density for the variable D_{xy}. This consideration applies equally to C and $\{l_k\}$.

Given that C and $l_1, ..., l_n$ are assumed independent,

$$p(D_{xy}|C, l_1, ..., l_n) \approx p(D_{xy}|C)p(D_{xy}|l_1, l_2, ..., l_n) \approx 1 \quad (1)$$

it follows that

$$p(D_{xy}|l_1, ..., l_n) \approx \frac{1}{p(D_{xy}|C)} = \frac{p(C)}{p(D_{xy}, C)} \quad (2)$$

Estimation of $p(D_{xy}, C)$ poses a challenging problem given that D_{xy} is relatively sparse, and only one observation C is provided per image for which eye tracking data is available. This would seem to present a situation where the database of images, and fixation data from experimental participants would need be exceedingly large (to a level that is impractical). To overcome this problem, we present three steps that allow for the desired quantity to be estimated given a data set that is of a practical size, while also allowing the end result to be visualized at the same resolution as the images displayed during capture of fixation data.

1. **Overcoming the sparsity present in the fixation data:** In the current work, we have examined images presented at a resolution of 1024 x 768 [Bruce and Tsotsos 2005] with a viewing time of 4s in a task-free examination paradigm. This resolution implies 786432 possible image locations that initially have a binary value (1 if fixated, 0 otherwise). These fixation images are converted to fixation density maps as in [Bruce and Tsotsos 2005; Bruce and Tsotsos 2009] to present a topographical representation of fixation density that is less sparse than raw fixation coordinates, and lower in dimensionality when vectorized. Subsequently, these density maps are down-sampled by a factor of 64 to yield a 16x12 representation of fixation density, with values in the down-sampled map given by averaging the density within 64x64 sub-blocks of the original 1024x768 fixation density map.

2. **Overcoming the curse of dimensionality:** Given a relatively high-dimensional vector in C, as discussed, representing $p(D_{xy}, C)$ in it's raw form poses a challenge. To this end, we seek to remove any redundancy in C across image samples, and to make the dimensions of C as independent as possible. The desirable property that is sought, is that given the n dimensional vector $C = c_1, ..., c_n$ we seek a transformation S such that $CS = C' = c'_1, ..., c'_n$ with the property that

$$p(c_1, ..., c_n) = p(c'_1, ...c'_n) = \prod_{k=1,...,n} p(c'_k)$$

This allows the n dimensional joint likelihood estimate to be treated as a product of n one dimensional density estimates. To derive a transformation with the desired properties, Cardoso's joint approximate diagonalization [Cardoso 1998] is applied to the vectors C_k with $k = 1, ..., 120$ corresponding to the test images, with PCA as a pre-processing step retaining 97% variance corresponding to the first 40 principal components. The joint likelihood of each $p(D_{xy}, c'_k)$ is approximated by a fit to a 2D Gaussian probability density function. $p(c'_k)$ is represented in similar fashion by fitting a 1D Gaussian probability density function to observed data.

3. **Restoring the spatial resolution of the fixation density map:** Given that the value computed in equation 2 is derived from fixations within the subsampled region D_{xy}, one might take advantage of the topology of the original fixations within each region corresponding to the sub-block D_{xy} for the raw fixation data. To re-map the coarse resolution output corresponding to the rightmost quantity appearing in equation 2, this value is divided evenly among fixations in the original high resolution binary map appearing within the

corresponding sub-block. That is, the value of $p(D_{xy}|c'_1, ..., c'_n)$ is divided evenly among all fixated positions at the original resolution that fall within the 64x64 spatial sub-block corresponding to D_{xy}. This is then transformed to a density map in the same fashion applied by Bruce and Tsotsos [Bruce and Tsotsos 2005; Bruce and Tsotsos 2009]. While fixations within any sub-block might contribute disproportionately to the resultant context independent density, this forms a reasonable basis for approximating the topology of the resulting fixation density at the original resolution.

4. **Visualization as fixation heatmaps:** To facilitate comparison across fixation density maps, and between fixation density maps and context-free fixation maps, the contrast is normalized via histogram equalization. In overlays, the image and fixation density (expressed as a heat-map) are added together and each given a weighting of 0.5.

3 Results

In this section, we present results demonstrating predicted context-independent fixation density as determined by the methods discussed thus far. Results are shown for two stages, specifically the resulting density heatmaps at a low resolution following stage 2 in section 2, and the final results following stage 4.

Figure 1 demonstrates the predicted context independent density measure at a the spatially quantized resolution. The 5 columns appearing in the figure correspond to the original image, the raw fixation density at the lower spatially quantized resolution, the low-resolution fixation density with mean density for each D_{xy} across all samples subtracted from individual D_{xy}, D_{xy} whitened across samples (mean subtracted, and with densities normalized so that the variance for D_{xy} is 1 across samples), and finally the estimate of $p(D_{xy}|\{l_k\})$. It is important to note that this final quantity is derived from the joint likelihood between the densities appearing in column 2, and the diagonalized context feature vector. The additional visualizations serve to compare the result with alternative representations that control for spatial bias only, and not an explicit representation of context.

Figure 2 demonstrates for each triplet of images, the original image (left), the raw fixation density heat map (middle), and the context-independent density heat map (right). Note that the heatmaps derived from the proposed method appear to be more strongly targeted on objects, or structure within the images that one might expect would garner overt attention. This may provide for better diagnosticity in examining targets that garner overt attention than the raw data presents.

3.1 Evaluation

Evaluation of the proposed visualization methodology poses a challenge, as one is working with data at the level of raw fixations to produce an alternate heat-map to characterize fixation density. As one means of casting further light on differences between the original representation, and context-free representation, we have analyzed the resulting heat-maps subject to user annotation of the image set. Borji et al. [Borji et al. 2013] provide annotations resulting from a user study in which 70 participants were requested to circumscribe the most salient object in the scene. The raw fixation density map, and context-free fixation density maps were assessed with respect to the proportion of density in the heatmaps lying within regions deemed to be salient by user annotation. In this analysis, we considered pixels within regions of each image annotated as salient by at least 3/70 participants (approximately 5%). For the original fixation density maps, on average 35% of the fixation density was within the *salient regions*, while on average 44% was within the *salient regions* for context-free density. This result also holds for other choices for the minimum number of observers with overlapping annotations in common for a given image. (All of 1-10 observers showing a greater proportion of density within *salient regions* for the context-free densities). This analysis, and qualitative observations on heatmap-salient region correspondence reveal that while salient regions are well characterized by fixation density, the context-free representation provides a more focal determination of regions deemed to be salient by observers. An additional observation in the case of larger objects, is that within the context-free heatmap, the focus may be more on key features of large objects, or the center of mass of objects.

4 Discussion

There are many benefits that may be had in analyzing gaze data with the goal of attaching greater meaning to individual fixations. This ranges from augmented visualization in examining behavior across a variety of application domains, to further analytics tied to higher level cognitive routines.

In the current work we have demonstrated a method that attempts to distill out fixation density that is not well explained by the spatial envelope of a scene. While there is no *exact* ground truth that might serve to verify that the methodology succeeds at this, the visualization results and density within *salient regions* suggest that the methods present a means of identifying focal points that tend to be specific items, or landmarks of interest within the scene. At a more general level, the methodology presented may have utility in developing more sophisticated models for attaching meaning to fixation data in extending the current work.

A rather ambitious goal in the analysis of fixation data, is that of solving a generalized version of the inverse Yarbus problem [Greene et al. 2012; Abolhassani and Clark 2011] (i.e. inferring intent from gaze data). It is worth noting that efforts that serve to ascribe greater meaning to fixations by way of more forensic analysis of eye tracking data, image properties, and temporal dynamics may allow for a corpus of data that is sufficiently rich to allow stronger determinations of this kind to be made.

Acknowledgements

The authors gratefully acknowledge the financial support of NSERC Canada and the University of Manitoba.

References

ABOLHASSANI, A. H., AND CLARK, J. J. 2011. Realization of an inverse yarbus process via hidden markov models for visual-task inference. *J of Vision 11*, 11, 218–218.

BORJI, A., SIHITE, D. N., AND ITTI, L. 2013. What stands out in a scene? a study of human explicit saliency judgment. *Vision research 91*, 62–77.

BRUCE, N., AND TSOTSOS, J. 2005. Saliency based on information maximization. In *Advances in neural information processing systems*, 155–162.

BRUCE, N. D. B., AND TSOTSOS, J. K. 2009. Saliency, attention, and visual search: An information theoretic approach. *J of Vision 9*, 3.

CARDOSO, J.-F. 1998. Multidimensional independent component analysis. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 4, IEEE, 1941–1944.

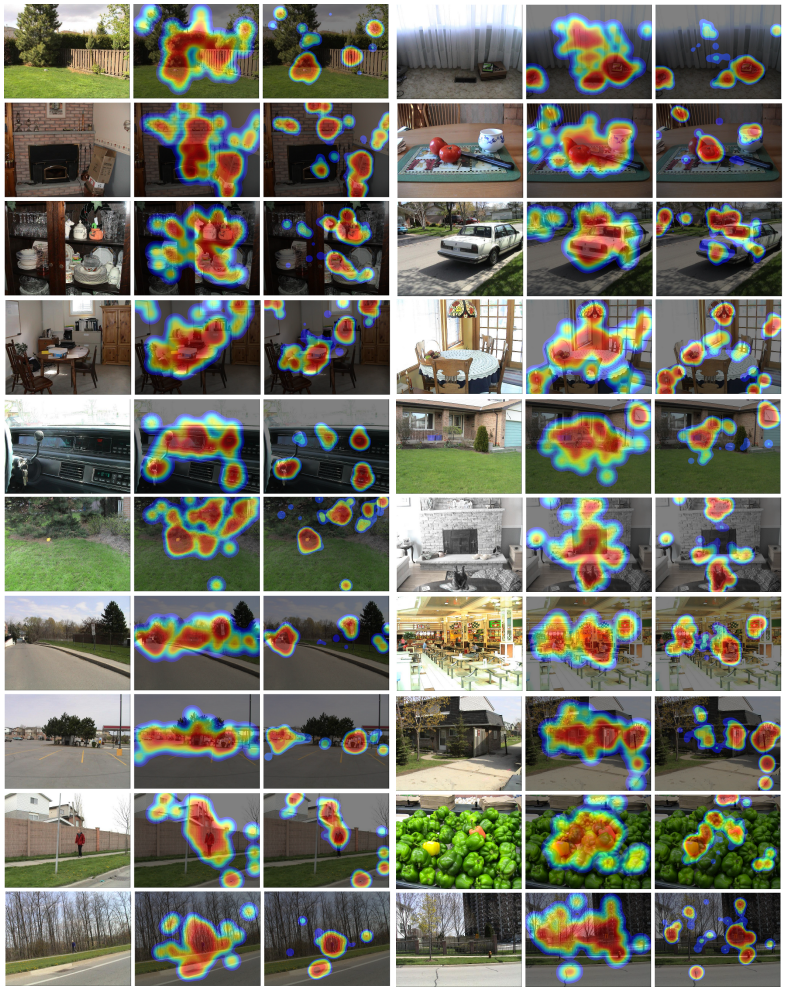

Figure 2: *Visualization of the final output for the proposed methods. For each set of three: Original image (left), corresponding unprocessed fixation density (middle), derived context-independent density (right). Contrast is normalized for each* heat map *via histogram equalization.*

CHEN, X., AND ZELINSKY, G. J. 2006. Real-world visual search is dominated by top-down guidance. *Vision research 46*, 24, 4118–4133.

CLARK, J. J. 1999. Spatial attention and latencies of saccadic eye movements. *Vision Research 39*, 3, 585–602.

FINDLAY, J. M. 1982. Global visual processing for saccadic eye movements. *Vision research 22*, 8, 1033–1045.

FOLLET, B., LE MEUR, O., AND BACCINO, T. 2011. New insights into ambient and focal visual fixations using an automatic classification algorithm. *i-Perception 2*, 6, 592.

GREENE, M. R., LIU, T., AND WOLFE, J. M. 2012. Reconsidering yarbus: A failure to predict observers task from eye movement patterns. *Vision research 62*, 1–8.

HAYHOE, M., AND BALLARD, D. 2005. Eye movements in natural behavior. *Trends in cognitive sciences 9*, 4, 188–194.

OLIVA, A., AND TORRALBA, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision 42*, 3, 145–175.

OLIVA, A., AND TORRALBA, A. 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research 155*, 23–36.

PANNASCH, S., AND VELICHKOVSKY, B. M. 2009. Distractor effect and saccade amplitudes: Further evidence on different modes of processing in free exploration of visual images. *Visual Cognition 17*, 6-7, 1109–1131.

ROTHKOPF, C., BALLARD, D. H., AND HAYHOE, M. 2007. Task and context determine where you look. *J of Vision 7*, 14.

TATLER, B. W. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *J of Vision 7*, 14.

TORRALBA, A., OLIVA, A., CASTELHANO, M. S., AND HENDERSON, J. M. 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review 113*, 4, 766.

TSENG, P.-H., CARMI, R., CAMERON, I. G., MUNOZ, D. P., AND ITTI, L. 2009. Quantifying center bias of observers in free viewing of dynamic natural scenes. *J of Vision 9*, 7.

YARBUS, A. L., HAIGH, B., AND RIGSS, L. A. 1967. *Eye movements and vision*, vol. 2. Plenum press New York.