Do Humans Fixate on Interest Points?

Akshat Dave, Rachit Dubey**
Nanyang Technological University
{aksh0006, rach0012}@ntu.edu.sg

Bernard Ghanem
King Abdullah University of Science and Technology
bernard.ghanem@kaust.edu.sa

Abstract

Interest point detectors (e.g. SIFT, SURF, and MSER) have been successfully applied to numerous applications in high level computer vision tasks such as object detection, and image classification. Despite their popularity, the perceptual relevance *of these detectors has not been thoroughly studied. Here, perceptual relevance is meant to define the correlation between these point detectors and free-viewing human fixations on images. In this work, we provide empirical evidence to shed light on the fundamental question: "Do humans fixate on interest points in images?". We believe that insights into this question may play a role in improving the performance of vision systems that utilize these interest point detectors. We conduct an extensive quantitative comparison between the spatial distributions of human fixations and automatically detected interest points on a recently released dataset of 1003 images. This comparison is done at both the global (image) level as well as the local (region) level. Our experimental results show that there exists a weak correlation between the spatial distributions of human fixation and interest points.*

1 Introduction

Detection of interest points has proven to be very useful in image processing for feature detection and extraction. Many forms of interest point detectors have been proposed which have found applications in mid/high level tasks such as image classification and object detection [6, 4, 8]. Researchers have evaluated the relative performance of these detectors [7]. However, these evaluations do not consider the relationship between automatically detected interest points and human fixations in images. The goal of this work is to shed light on this relationship through extensive empirical evaluations. Such evaluations provide insight on the perceptual relevance of popular interest point detectors.

Visual attention is a fundamental function of the human visual system (HVS). It permits efficient and parsimonious sampling of the input visual stimulus, whereby the HVS fixates on a sparse number of locations in the visual field. The phenomenon of visual attention in the HVS has been studied extensively in the past several decades. Understanding attention in images has many useful applications in the field of computer vision [3, 10]. When viewing an image, the HVS actively distributes more perceptual resources to salient locations than other locations in the visual field. This non-uniform allocation of resources makes high-level tasks (e.g. pattern recognition) computationally feasible. In fact, there is abundant empirical evidence of this non-uniformity especially in eye-fixation data (recorded image locations where eyes fixate). Here, we only consider free-viewing human fixations, i.e. locations in the image that a subject fixates on without requiring a particular visual task to be done by the subject. We do this to factor out higher level semantics used by the HVS in fixation. There is evidence that the HVS relies on bottom-up information (e.g. low-level image properties such as intensity contrast) for attention that is in turn used for high-level tasks including recognition [9]. Therefore, this paper takes an initial step towards understanding the relationship (if any) between free-viewing human fixations and popular interest points detectors that are deterministic functions of low-level (bottom-up) image information.

To the best of our knowledge, this work is the first quantitative comparison between popular interest point detectors and human fixation data. We focus on four widely used detectors, namely Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), Maximally Stable Extremal Regions (MSER) and Harris Corner Detector (HCD). Due to space limitations, we refer the reader to [7, 1] for a detailed description of these detectors. For comparison, we include Random Sampling of Edge Points (RSPE) generated by the Sobel edge detector. In Figure 1, we show detection results of the five aforementioned detectors as compared to eye fixations recorded on the same image.

The contributions of this paper are two fold. **(i)** It is the first quantitative comparison between interest point detectors and human fixations. **(ii)** From our experi-

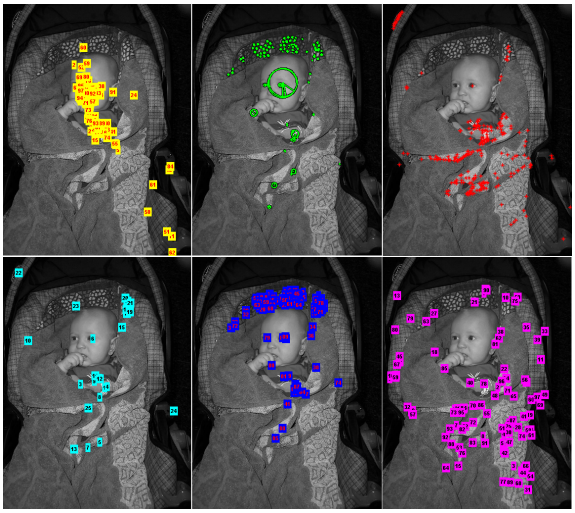

Figure 1: Top row (left to right) depicts fixation, SIFT, and HCD points on the same image. Bottom row (left to right) depicts MSER, SURF, and RSPE points.

mental results, we observe some interesting properties that describe the relationship between human fixations and automatically detected interest points. For example, we observe that there exists a weak correlation between their spatial distributions.

2 Comparative Methodology

We denote the set of images belonging to a fixation point dataset as I. Each image $i \in I$ has an associated list of fixation points $F_{\mathbf{x}}(i)$, where $\mathbf{x}$ denotes the image location of the fixation. We set this list as ground truth for all comparisons. We also define the list of interest points generated by interest point detector j as $\Phi_{\mathbf{x}}(i,j)$, $j \in D = \{\text{HCD}, \text{SIFT}, \text{SURF}, \text{MSER}, \text{RSPE}\}$.

Our goal is to compare the point sets returned by detectors in D to the ground truth fixations. We do this by comparing the spatial distribution of $\Phi_{\mathbf{x}}(i,j)$ (interest points generated by detector j in image i) to the distribution of $F_{\mathbf{x}}(i)$ (fixation points in image i). Instead of assuming a parametric distribution which might not fit the underlying data well, we estimate these distributions in a parameter-free data-driven manner using kernel density estimation (KDE). We denote the KDE density of $F_{\mathbf{x}}(i)$ and $\Phi_{\mathbf{x}}(i,j)$ as $P_F(\mathbf{x}|i)$ and $P_\Phi(\mathbf{x}|i,j)$ respectively. Note that the value of $\Phi_{\mathbf{x}}(i,j)$ (or $P_F(\mathbf{x}|i)$) at an image location $\mathbf{x}$ is computed as a linear combination of kernel values evaluated at each interest point (or fixation) location in image i. It is explicitly expressed in Eq (1), where $\mathbf{x}_k$ denotes the location of the k^{th} interest point generated by the j^{th} detector when applied to image i. In our experiments, we use the RBF kernel function and set the smoothing parameter h to its corresponding optimal value.

$$P_\Phi(\mathbf{x}|i,j) = \frac{1}{Nh}\sum_{k=1}^{N} K\left(\frac{\mathbf{x}-\mathbf{x}_k}{h}\right) \tag{1}$$

We compare these KDE densities at a global and local level. Globally, we use the Bray-Curtis histogram similarity measure [2] to compare the distributions. Locally, we formulate the comparison as a binary classification problem and evaluate each detector using receiver operator characteristic (ROC) curves.

2.1. Global Comparison

To determine how close the KDE densities are over the whole range of each image in I, we compare the distributions using a global measure. To do this, we use the Bray-Curtis similarity (BCS) measure for histograms, as defined in the expression below. There exist many histogram-to-histogram similarity measures; however, we choose BCS for its simplicity and its direct mapping to the interval $[0,1]$, where a larger BCS value indicates that the distributions are more similar. Note that BCS is similar in nature to the χ^2 distance.

$$BCS\left(P_F, P_\Phi\right) = 1 - \frac{\sum_{\mathbf{x}} |P_F(\mathbf{x}|i,j) - P_\Phi(\mathbf{x}|i,j)|}{\sum_{\mathbf{x}} \left(P_F(\mathbf{x}|i,j) + P_\Phi(\mathbf{x}|i,j)\right)}$$

2.2. Local Comparison

Distributions $P_F(x,y|i)$ and $P_\Phi(\mathbf{x}|i,j)$ are not likely to overlap in the entire range of image i; however, this overlap might occur in certain portions of the image. Therefore, to provide a fair quantitative comparison between interest point detections and human fixations, it is necessary to conduct a local comparison between their spatial distributions.

To compare the sets of interest and fixation points locally, we formulate the comparison as a binary classification problem, whereby the detected interest points are used to discriminate between fixation and non-fixation points. Each detector generates interest points, which compete to describe the fixation points. Since interest points are not labeled as fixation or non-fixation points, we hypothesize that an interest point $\mathbf{x}$, generated by detector j, is deemed a fixation in image i if $P_F(\mathbf{x}|i,j) \geq \tau$. This decision function defines the binary classifier for detector j on image i. By averaging the classification performance of this classifier over all images in I, we obtain an average true positive and false positive rate for a given threshold τ. By varying τ, we can generate an ROC curve that depicts the overall performance of each detector in predicting fixations. To

quantitatively compare detectors, we compute the area under the ROC curve for each detector. Clearly, the larger this area, the more correlated the interest points are to human fixations.

As such, we provide two approaches to compare interest point detections with human fixations. These approaches quantify how similar the underlying spatial distributions are for both types of points, and thus, how perceptually relevant each detector is to human attention.

3 Experiments and Results

The ground truth data used for comparison is obtained from the recently compiled eye fixation dataset available in [5]. This free-viewing dataset comprises 1003 images of diverse semantic content. In this section, we report the results of comparing the five interest point detectors with human fixation based on the local and global criteria described in Sections 2.1 and 2.2.

The parameters for each detector are set apriori such that the number of interest points detected in each image is similar to the number of fixations (refer to Table 1). We keep them fixed for all images in the dataset.

Detector	Parameter	Value
HCD	Max Corners	250
MSER	MaxVariation	0.25
	MinDiversity	0.25
	Delta	37
SIFT	Peak Threshold Edge	10
	Threshold	5
SURF	Hessian Response Threshold	200
RSPE	Threshold	auto

Table 1: Parameters for the 5 interest point detectors

3.1. Global Comparison

We compute the BCS measure between each detector and the ground truth on each image in the dataset. To evaluate the overall performance of detector j, the histogram of its BCS measures to ground truth is constructed (refer to Figure 2). Here, we add the performance of a baseline detector (denoted RSP) that randomly samples points in the image as interest points. Since the maximum BCS value is 1, a mean closer to 1 indicates a more perceptually relevant detector. From the plots in Figure 2, we observe that the BCS of each detector follows a similar trend. Interestingly, all the histograms resemble a normal distribution. In Table 2, we rank the different interest point detectors based on the mean BCS value (μ) over all images in the dataset.

As expected, the BCS of the RSP detector follows a normal distribution. Surprisingly, the other detectors have very similar performance (mean and variance) to that of the RSP. The RSPE detector performs marginally better than the other detectors with $\mu = 0.3869$ followed by SURF, SIFT, and MSER. HCD ranks last. As evident from the values of μ and the histograms, all the detectors except HCD give marginally better results than the RSP detector. However, when compared to human fixation, the performance of all detectors is relatively low, with similarity ranging from 30% to 39% . By taking this similarity as a measure of perceptual relevance, we observe that the detectors are weakly relevant to human attention.

Ranking	Detector	Mean BCS Value (μ)
1	RSPE	0.3869
2	SURF	0.3787
3	SIFT	0.3358
4	MSER	0.3350
5	**RSP**	**0.3340**
6	HCD	0.3049

Table 2: Ranking of detectors based on mean BCS value

3.2. Local Comparison

For local comparison, each detected interest point is checked to see if it qualifies as a human fixation in each image. By varying the threshold τ in Section 2.2, we generate an average ROC curve over the whole dataset for each detector. In Table 3, we rank the different detectors based on area under the ROC curve. These results show a similar trend to the results of the global comparison. The performance of the detectors follows a similar trend, is quite low, and is quite similar to the that of the baseline RSP detector. Since the areas in Table 3 are smaller than 0.5, we conclude that the interest point detectors do not hold much power in discriminating between fixation and non-fixation points.

Ranking	Detector	Area Under ROC Curve
1	RSP	0.5
2	RSPE	0.4978
3	MSER	0.4974
4	SIFT	0.4939
5	SURF	0.4879
6	HCD	0.4501

Table 3: Ranking of detectors based on ROC curve

4 Discussion

Our experimental results show that the interest point detectors give similar results for under both local and global comparisons. Globally, the BCS histograms of

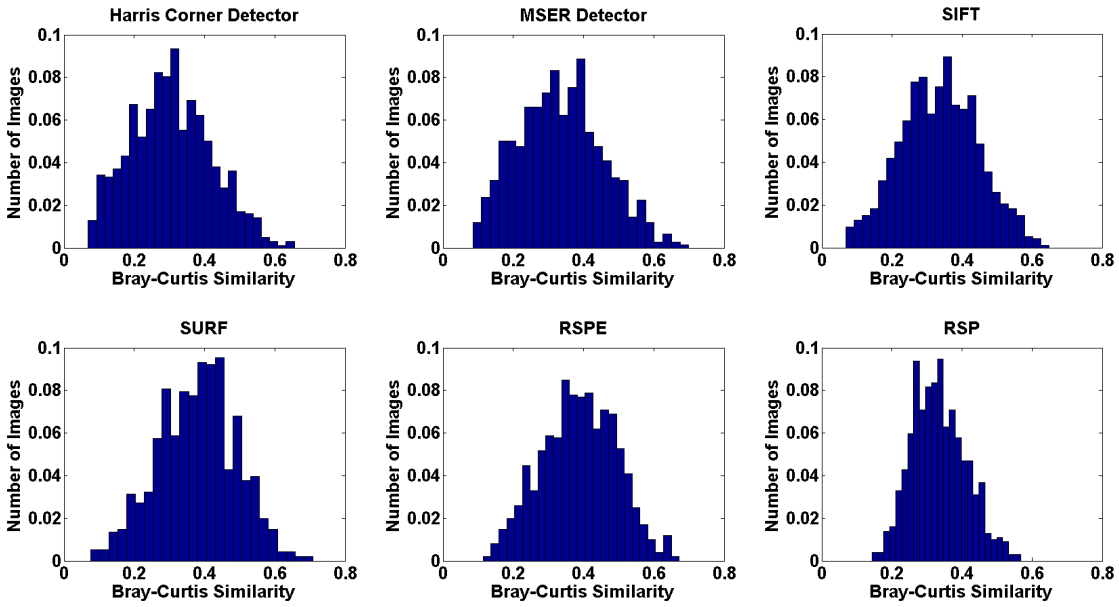

Figure 2: Normalized histogram of BCS measure (over 1003 images) between fixations and 5 interest point detectors: HCD, MSER, SIFT, SURF, and RSPE. As a baseline detector, we also show the BCS results of randomly sampling points (RSP) as interest points in the dataset images.

the different detectors are quite similar to each other and resemble Gaussians with similar variances. Moreover, all these histograms are very similar to that of the RSP histogram. Locally, the ROC curves of the interest point detectors are found to be similar to the ROC of the RSP. Furthermore, even at the local comparison level, there is a significant similarity between the performances of different detectors. Most of the detectors tend to be slightly inferior in performance when compared with the RSP (evident from the values of the areas under the ROC). Overall, the interest points generated by these detectors show a weak correlation with the human fixations in terms of their respective spatial distributions.

Our results show that while the performance of the interest point detectors is better than random in most cases, there is still significant room for improvement in the "perceptual relevance" of interest points detected. This work provides empirical evidence that the points detected by popular interest point detectors are quite dissimilar to human fixation points and thus, have low perceptual relevance to human attention.

5 Conclusion

In this paper, we present an extensive quantitative comparison between interest point detectors and human fixation points. We evaluate the perceptual relevance of four widely used detectors, SIFT, SURF, MSER, HCD along with RSPE under global and local criteria. Our empirical comparison give us a holistic picture of the behavior of these detectors. A general trend is observed in the spatial distributions of the interest point detectors. As seen from the experimental results, there exists a weak correlation between the spatial distribution of fixation and interest points. This work is a stepping stone towards the development of a biologically inspired interest point detector, which learns from where humans look to possibly improve upon the performance of automatic vision systems that utilize interest point detectors.

References

[1] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417. 2006.

[2] J. R. Bray and J. T. Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325 – 349, 1957. Ecological Society of America 1957.

[3] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou. A visual attention model for adapting images on small displays. *Multimedia Systems*, 9:353–364, 2003.

[4] M. S. L. C. S. J, Zhang; M. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73:213–238, 2006.

[5] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106 –2113, 2009.

[6] D. G. Lowe. Object recognition from local scale-invariant features. *ICCV*, 2(8):1150–1157, 1999.

[7] C. Mikolajczyk, Krystian; Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[8] F. . T. B. Nowak, Eric ; Jurie. Sampling strategies for bag-of-features image classification. *ECCV*, (3954):490–503, 2006.

[9] Y. Ostrovsky, E. Meyers, S. Ganesh, U. Mathur, and P. Sinha. Visual Parsing After Recovery From Blindness. *Psychological Science*, 20(12):1484–1491, 2009.

[10] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *SIGCHI conference on Human Factors in computing systems*, pages 771–780, 2006.