VISUAL CORRELATES OF FIXATION SELECTION: A LOOK AT THE SPATIAL FREQUENCY DOMAIN

Neil D. B. Bruce, Daniel P. Loach, John K. Tsotsos

York University
Department of Computer Science and Centre for Vision Research
4700 Keele Street, Toronto, Ontario, Canada M3J 1P3

ABSTRACT

A representation for observing local image content is proposed for the purpose of considering the distinguishing characteristics of visual content that tends to draw a human observers gaze. Within this representation, the spectral profile distinguishing fixated from non-fixated locations is considered. Finally, the possibility of designing saliency operators based on the proposed local magnitude spectrum representation is explored, revealing a promising domain for predicting human gaze patterns.

Index Terms— spatial frequency, fourier transform, magnitude spectrum, attention, fixation

1. INTRODUCTION

The primate visual system is foveated and thus samples visual content at the center of fixation at a much higher resolution than in the periphery. Head movements and eye movements are made in such a manner that some regions of a scene receive intense scrutiny while others are relegated to only very low resolution sampling. In recent years, several attempts have been made at furthering the understanding of this selection process for its utility as a precursor to various operations of interest in image processing such as perceptually motivated compression or quality assessment.

It is undeniable that fixation selection is influenced appreciably by at least two factors: The properties of the surrounding environment, and the goals of the observer. For example, one might be far more likely to fixate faces in a crowd while looking for a friend, but would almost certainly be distracted by a bright flash of light, or vivid colors while doing so. In the literature, these two distinct components of the selection process are frequently referred to as top-down and bottom-up components respectively.

In this paper we consider the latter of these categories in order to address the following question: To the extent that selection of fixation points is stimulus driven, what sort of stimulus properties draw a human observers' gaze. Consideration of this problem has been the focus of some recent research efforts [1, 2]. Generally the approach that is taken in addressing this problem, is that of considering some basic feature measures on the image (e.g. contrast, edges etc.) and observing the extent to which such features are able to predict fixations.

One limitation of this sort of study, is that typically features are considered in isolation. That is, the extent to which edges, contrast, colors and other features are predictive of fixation points is typically considered for each feature independently. In reality, it is likely that some combination of these various features determines the criterion for fixation selection. It is this observation that forms the basis for the work presented here. It is expected that in considering local image content in a manner that simultaneously captures a rich array of orientation and spatial frequency content present in a local neighborhood of the scene, that this may elucidate the nature of stimuli that attract an observers gaze and as a by-product, afford a system for predicting fixation points. Some previous efforts that characterize saliency based on some combination of features have shown success [3, 4]. The distinction made in this work, is that i. Analysis is based on a *raw* representation of spatial frequency and orientation content rather than a combination of features eliminating dependence on a specific feature set ii. Because of consideration i. the features that give rise to fixation selection, or distinguish those points that are fixated from those that are not are more directly observable. iii. When viewed as a saliency operator, the operator proposed here is qualitatively different than any previous effort offering the possibility of improved performance, or at a minimum, a deeper understanding of what sort of model and/or stimuli is important in characterizing human gaze.

2. LOCAL MAGNITUDE SPECTRA

As described in the introduction, we seek a representation that allows direct observation of orientation and spatial frequency content within the image in question. The most obvious representation fitting this criterion, with a long history of use in signal processing, is the magnitude spectrum. It has been demonstrated that such spectra are able to adequately char-

We gratefully acknowledge NSERC for funding this research project.

acterize different contexts (e.g. forest, mountain, etc.), with the magnitude spectrum of each presenting its own characteristic shape [5]. It is also demonstrated in the same work that global magnitude spectra are able to distinguish between different categories of objects. One might imagine on the basis of these results, that a similar representation might be used to distinguish certain parts of a scene from others on the basis of their frequency components.

For an image $S(x, y)$ the magnitude spectrum $P(u, v)$ is given by $|\Im(u, v)| = (R(u, v)^2 + I(u, v)^2)^{\frac{1}{2}}$ where $\Im(u, v)$ is the 2D Fourier transform of $S(x, y)$ and $R(u, v)$ and $I(u, v)$ are the real and imaginary components of $\Im(u, v)$ respectively. Although this representation allows direct observation of spatial frequency and orientation content, it eliminates an important property inherent in the images themselves, that being the fact that pixels are localized in space. Ideally, the representation should be localized in space, and in frequency, but without a decomposition into independent features as in a Wavelet transform. For this reason the following representation is proposed: For an image $S(x, y)$ we may consider the magnitude spectrum, but within a localized region only. In this manner, for any given spatial location, the spatial frequency content is summarized in the form of a local magnitude spectrum. Consider the example of figure 1. An image is depicted, along with the magnitude spectra derived from a 120 by 120 window centered at 3 spatial locations. Note the variety in shape and the range of spatial frequencies present in each. A band of energy at a particular orientation corresponds to an edge orthogonal to the orientation of the band.

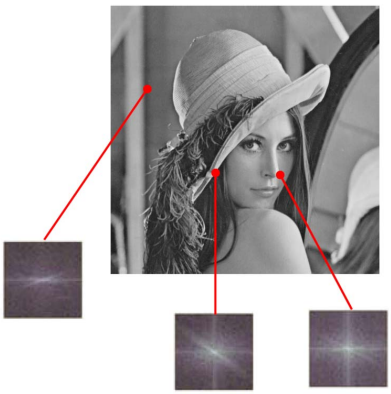

Fig. 1. An image along with the local magnitude spectra corresponding to three local neighborhoods.

3. EYE TRACKING DATA

Eye tracking data was collected for the purpose of considering the extent to which local spatial frequency content informs on salient visual content. Data was collected for a set of 250 grayscale images, from 10 subjects each viewing 50 images (5 sets of 50 images, with 2 subjects viewing each set). Images were randomly chosen from the Corel stock photo database and presented in random order for 4 seconds each with a mask between each pair of images for 2 seconds. Analysis was based on the glint-pupil vector data obtained from an Arrington Research ViewPoint EyeTracker. Images were presented on a 21 inch CRT monitor at a resolution of 1024 x 768, with participants positioned at a distance of 70 cm. Participants were naive to the purpose of the study and were instructed simply to observe the images.

4. CORRELATES OF FIXATION SELECTION

Given the representation described in section 2, one can imagine a variety of quantities to consider in relation to fixation data. The most obvious consideration is perhaps the comparison of magnitude spectra derived from locations fixated by human observers versus those corresponding to patches sampled randomly. A qualitative comparison may be achieved in observing figure 2. Figure 2 demonstrates a comparison of the average magnitude spectrum of regions centered at the approximately 3000 fixated regions appearing in the data set as compared with the average magnitude spectrum of 62500 randomly selected patches in the form of an arithmetic difference.

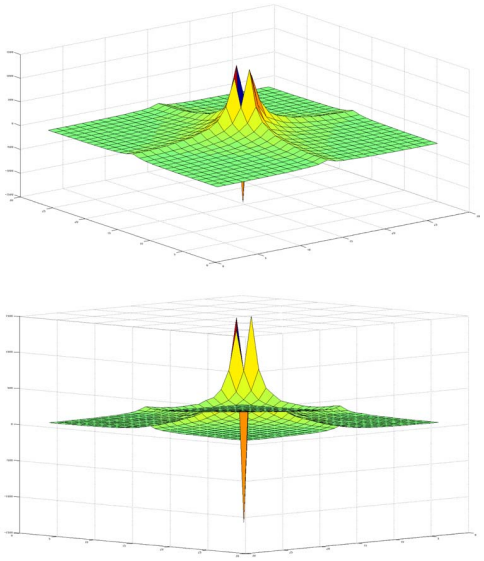

Fig. 2. Two views of the difference between the average magnitude spectrum of fixated points versus the average of non-fixated regions. The centre "hole" corresponds to the origin and the elongated peaks moving to higher spatial frequencies correspond to vertical and horizontal structure.

In observing the difference in figure 2, we notice that there is relatively more mid-range spatial frequencies in fixated regions, especially oriented vertically and horizontally as well

as a great deal less content of very low spatial frequency in fixated patches. This agrees with our intuition since people tend not to fixate *empty* regions of the sky, or a blank wall; regions with a greater proportion of very low spatial frequency content. Figure 2 then summarizes the envelope of spatial frequency and orientation content that distinguishes fixated from non-fixated coordinates. This consists of the minimum possible content at the very low spatial frequency end, and the maximum possible at mid range spatial frequencies especially oriented horizontally and vertically. Of note is the similarity of this profile to a Laplacian-of-Gaussian filter. As in the study of Tatler et al., it is interesting to consider the extent to which measures based on the local magnitude spectra are predictive of points fixated in the experimental data [1]. In this light, the following quantities are considered.

- Let A be the average of the magnitude spectra of all local neighborhoods in all 250 images.
- Let I^k be the average of the magnitude spectra of all local neighborhoods in image k.
- Let F be the average of the magnitude spectra of all local neighborhoods surrounding fixation points.
- Let FI^k be the average difference between the magnitude spectra of fixated points in image k, and I^k. FI refers to the average of $FI^k \; \forall \; k$

In each case the difference described is based on magnitude spectra and the resultant quantity is equated to a measure of saliency (π). The saliency maps are the result of computing the following measures $\forall \; i, j \in S$. Resultant saliency maps are filtered with a Gaussian approximation of the dropoff in visual acuity from any fixation point so that saliency in sampling takes into account foveation.

The following operators will from hereon be referred to as type 1 to type 5 respectively and x is $\frac{1}{2}$ the height/width of A,I and F minus 1 (27-1)/2 = 13 in this implementation.

1. $\pi_{i,j} = \sum_{i-x}^{i+x} \sum_{j-x}^{j+x} ||\Im(S_{i\pm x, j\pm x})| - A|$. This quantity describes the extent to which each image patch resembles the *average image patch*. One might expect that the distance of a local magnitude spectrum from the average magnitude spectrum will predict uncommon image patterns which may equate to salient content.
2. $\pi_{i,j} = \sum_{i-x}^{i+x} \sum_{j-x}^{j+x} ||\Im(S_{i\pm x, j\pm x})| - F|$. The proximity of each local neighborhood to a *typical* fixated region. Patches that more closely resemble fixated patches may indicate content of interest.
3. Type 1 - Type 2. Since F is still similar to the typical $1/f$ spectrum of natural images, it may still be that distance from this typical form correlates with salient content. However, drawing such samples from fixation points is likely to pull the local spectra and those of salient patches closer together. For this reason, it makes sense to consider the difference as a measure of the saliency attributable to the similarity of any given patch to the average of fixated patches.
4. $\pi_{i,j}^k = \sum_{i-x}^{i+x} \sum_{j-x}^{j+x} ||\Im(S_{i\pm x, j\pm x})| - I^k|$. A measure of how any given neighborhood compares to its context. One might expect those neighborhoods with a large difference from the rest of the image to predict unusual or unexpected content.
5. $\pi_{i,j}^k = \sum_{i-x}^{i+x} \sum_{j-x}^{j+x} ||\Im(S_{i\pm x, j\pm x})| - I^k - FI|$. The extent to which the type 4 quantity resembles the average difference between fixated patches and their context. The raw difference in itself may be less prescient than the extent to which this measure resembles the *typical* difference between fixated regions and their context.

5. RESULTS

In this section, qualitative and quantitative aspects of the operators described in the preceding section are considered. Figure 3 demonstrates for a few examples the output of operators of types 1 through 5. In each case the original image is at the top left, followed by the output of type 1 and 2 operators in top middle and top right. The bottom row from left to right depicts types 3, 4 and 5 respectively.

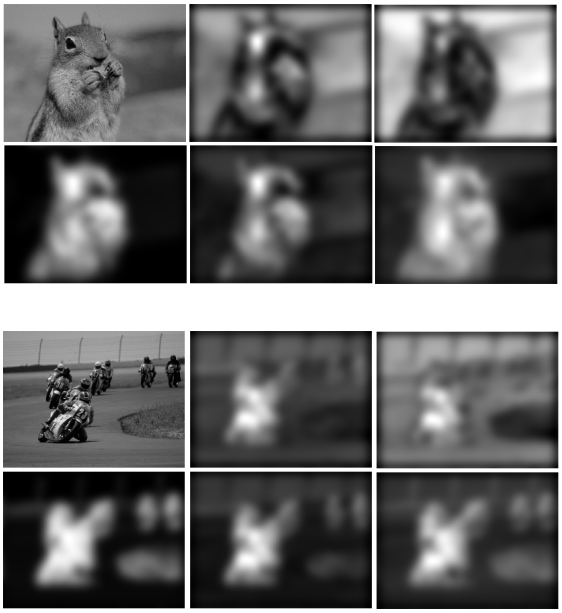

Fig. 3. Two example images depicting the saliency maps associated with the 5 types of operators. From left to right: Top: Original Image, Type 1 and Type 2 Filters. Bottom: Filter types 3-5 respectively.

As in the study of Tatler et al. [1] we also consider the extent to which each measure is predictive of fixated regions

of an image. The method proposed in the study of Tatler et al. is that of choosing a large variety of thresholds for each saliency operator. Once thresholded the saliency map can be treated as a binary classifier and the number of correctly classified fixated and non-fixated points can be determined. Selection of a large number of such thresholds yields an entire curve (an ROC curve), and the area under the curve a performance measure as a predictor of salient visual content. Figure 4 demonstrates the performance of the various magnitude spectrum based saliency operators in the form of ROC curves. Areas under each of the curves are (Types 1-5): 0.5843, 0.5165, 0.6662, 0.6076, 0.6876. The proposed operators are also compared with the model of Itti and Koch [4] which yields an ROC score of 0.6654. Operator 5 shows especially strong performance and overall is the strongest classifier based on ROC area. Also note that it is especially strong in the range in which false positives are unacceptable.

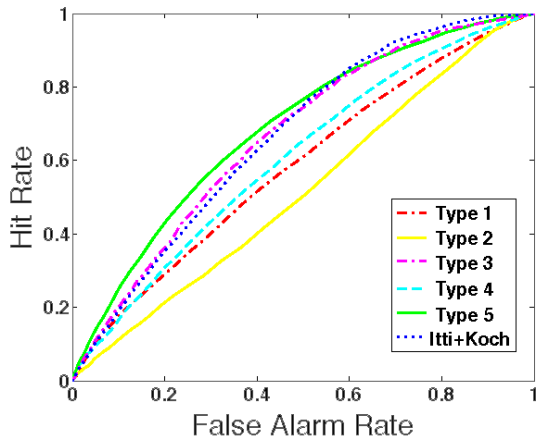

Fig. 4. ROC curves for the various operators.

Interestingly, the set of saliency measures considered gives rise to a wide array of performance measures ranging from effectively chance with operator 2, to decent predictors of fixations for operators 3 and 5. It is interesting to note that types 1 and 2 are very poor predictors of fixations by themselves, but that their difference is quite a good classifier. This is perhaps intuitive since distance from the *average* spectrum might be expected to predict salient content, while proximity to the characteristic spectrum of fixated content might be expected to predict salient content. As such type 3 in some sense is a better measure of the quantity that type 2 seeks, since it characterizes the distinction between fixated neighborhoods, and average neighborhoods. The results at this stage are very promising, as type 3 and type 5 operators perform better than that of Itti and Koch and also any of the operators considered in [1], which includes a variety of different measures based on contrast, edges and color at multiple scales.

The ambition of this work lies in 1. trying to characterize the difference between fixated and non-fixated regions both based solely on local statistics, and also on context and 2. To consider the efficacy of local magnitude spectra as saliency operators in light of the results described in [5]. With respect to this latter goal, it is important to note that the results presented here include only an intuitive selection of admittedly *ad hoc* operators based on the magnitude spectra. The intention of this effort is largely to establish whether this type of analysis (based on local magnitude spectra) may prove effective in producing a high quality characterization of saliency given further consideration. In this regard, the results are very encouraging, yielding performance greater than a wide selection of feature measures previously considered. One last point is that it is likely that both low-level local statistics and context play an important role in selection, and as such, some metric that combines the intuition behind both operators 3 and 5, or bootstrapping based on these two and/or other metrics might perform especially well.

6. CONCLUSION

A novel means of characterizing content within a local neighborhood was proposed within the context of considering content associated with the deployment of overt attention. Within this representation, the spectral profile differentiating fixated from non-fixated patches was considered revealing relatively less low spatial frequency content in fixated patches with a greater emphasis on mid range spatial frequencies, and a slight bias for horizontally and vertically oriented content.

Given prior work demonstrating the characteristic magnitude spectra associated with different contexts, and objects, we considered the potential for such spectra to distinguish between content that warrants fixation from human observers versus that which does not. Results based on a selection of intuitive choices for saliency metrics reveals significant promise for the possibility of local magnitude spectra based saliency operators.

7. REFERENCES

[1] Tatler, B.W., Baddeley, R.J., and Gilchrist, I. D., Visual correlates of fixation selection: effects of scale and time. Vision Research, 45(5), 643–659, 2005.

[2] Parkhurst, D., Law, K., and Niebur, E., Modelling the role of salience in the allocation of visual selective attention. Vision Research, 42 (1), 107–123, 2002.

[3] Bruce, N., Tsotsos, J.K., Saliency Based on Information Maximization. Advances in Neural Information Processing Systems, 18, 155–162, 2006.

[4] Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions PAMI, 20(11), 1254–1259, 1998.

[5] Torralba, A., Oliva, A., Statistics of natural image categories. Network: computation in neural systems, 14, 391–412, 2003.

[6] Bruce, N., Tsotsos, J.K., A Statistical Basis for Visual Field Anisotropies. Neurocomputing, 69:10-12, 1301-1304, 2006.