An Inverted Saliency Model for Display Enhancement

Sara L. Su SARASU@MIT.EDU
Frédo Durand FREDO@MIT.EDU
Maneesh Agrawala MANEESH@MICROSOFT.COM
MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge MA, 02139

1. Introduction

Visual attention is crucial to a subject's ability to retrieve information from complex visual stimuli. The order of saccades and fixations as a subject scans a scene determines the mental representation formed from it. It has been shown that image comprehension is significantly improved when a subject is provided with visual cues to guide the gaze path to semantically important regions (Bétrancourt & Tversky, In press). This problem is particularly important in image analysis scenarios involving medical, geological, or remote sensing data where important features might not appear to the untrained eye.

Studies have found human visual attention to be driven by neuronal tuning for particular features in the field of view (Itti & Koch, 2001). The bottom-up view of visual attention theorizes that a scene is first processed in parallel at low resolution. *Salient* low-level features may "pop out" at this level, causing the eyes to move to focus on an initial region of interest for further inspection at high resolution (Rosenholtz, 1999). Pop-out may be triggered by contrasts in intensity, color, or orientation, or by features such as edges or corners.

Based on a biologically plausible model of early visual processes (Koch & Ullman, 1985), Itti et al. have proposed a computational model that attempts to simulate feature-specific neuronal responses to an input image with center-surround filters (Itti et al., 1998). Salient regions found during the pre-attentive stage are identified as outliers in the filtered responses.

We use this model to develop image processing algorithms for redirecting attention to specific regions in a scene by imposing saliency contrast constraints between cognitively significant regions and the background. Simply identifying a region to emphasize is not enough, as there exists an infinite number of transformations leading to this goal. The challenge is to find a transformed image close in pre-attentive appearance to the original. In this paper, we describe a technique for altering saliency by changing the frequency distribution of an image and discuss early results.

2. Applications

Saliency alteration tools would be useful in a number of scenarios. An interactive system would aid educators who wish to create more comprehensible images but who lack the artistic skill. An automatic pictorial emphasis tool is an alternative to the labor-intensive photo editing methods that are the norm today. The multi-scale image processing techniques we are developing subtly change saliency of image regions without introducing objectionable visual artifacts.

We target an educational scenario in which a user indicates which regions of an image should be salient to an audience, or a semi-automated scenario in which a machine-vision system analyzes the image and tags regions of interest. To generate the results shown in this paper, a subject simply drew polylines around regions of interest. These were used to create binary masks indicating regions in which to increase and decrease saliency.

The development of saliency-alteration tools is also an important validation step for the Itti et al. model which has been previously lacking. Because our work relies on the assumption that this model captures the bottom-up mechanisms of human visual attention, we will need to empirically validate it using eye-tracking (Duchowski, 2003). If the model is accurate, eye-tracking experiments should indicate different gaze patterns and search performances before and after application of the emphasis tool. Our findings could also aid the design of more effective user interfaces by measuring effectiveness of spatial layout and complexity.

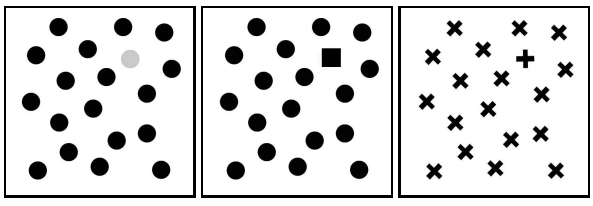

Figure 1. Three simple images containing salient, low-level features triggering pop-out phenomena.

3. Approach

We are developing tools to emphasize or de-emphasize regions of a natural image. Informally, this corresponds to inverting the Itti et al. computational model of saliency. This model can be viewed as a mapping from an image to a spatial saliency map. Given a target saliency map, we wish to find the image satisfying these constraints. Inversion is non-trivial because the model contains a number of non-linearities and is not a one-to-one mapping. However, we can define image transformations that increase or reduce the output of the various stages of the model, thereby modifying the output saliency. This is similar in spirit to gradient descent optimization. We view this as a special case optimization with the goal being to reach a feasible set that satisfies user-specified constraints.

Recall that the model defines salient regions as outliers from the local feature distribution. In other words, these are regions in which feature distribution changes dramatically. Therefore, to decrease saliency of a region, we must reduce variation in its feature distribution. We must use invertible low-level features to facilitate final image reconstruction. With this goal in mind, we have explored an image decomposition built on the *steerable pyramid*, a multi-scale, multi-orientation image decomposition in the frequency domain using an overcomplete wavelet transform (Freeman & Adelson, 1991). This representation has the desirable properties of near-perfect image reconstruction and encoded orientation information about the original image. We discuss only the basics of steerable pyramids here and refer the reader to the original paper by Freeman and Adelson for further detail.

To build a steerable pyramid, an image is first filtered into lowpass and highpass subbands. The lowpass subband is further filtered into a set of oriented bandpass subbands and one lowpass subband. The new lowpass subband is then subsampled by a factor of 2 in the x- and y-directions, and the process is repeated to generated the multi-scale steerable pyramid. The strength of coefficients at each scale corresponds to the local spatial frequencies in the original image.

Because each subband is itself an image, we can use it as the input to the steerable pyramid algorithm. This *recursive steerable pyramid* effectively encodes the variation in local feature distribution in the original image. Using this decomposition, we can suppress and promote select frequencies with the effect of altering spatial variation of texture.

In early experiments, we have completely removed variations in local feature distribution (i.e. setting coefficients constant), in a sense creating an image in which texture is globally uniform. Results are shown in Figure 2 for a natural image taken by an amateur photographer. The middle image shows the result of reconstructing the original image's recursive pyramid after altering it to reduce texture variation in the scene background. We applied a high-pass filter on feature response after non-linearity and completely removed the low-frequency variation of feature strength. Note that we do not blur the background but instead reduce texture variation. Therefore, although saliency is reduced, information is not lost. In other words, while the image remains sharp, the *texture boundaries* have been blurred. In practice, we built the image's steerable pyramid, applied an absolute-value non-linearity followed by a high-pass filter with a low cutoff. The image was reconstructed by reinjecting the signs of the original image to resolve the ambiguity due to the absolute value.

The bottom image in Figure 2 shows the result of exaggerating existing variations in local feature distribution, effectively increasing global texture variation to more sharply define region boundaries.

4. Current progress and future work

Early results are encouraging, but our approach needs to be refined. First, we need to process not only contrast, but all *feature channels* captured by the saliency model, including orientation and color. Second, removing variations globally may be too extreme; we may only want to reduce the local variations in feature value distribution. Finally, in Figure 2, we have treated all scales equally for the steerable pyramid, including coarse scales (up to 1/10th of the image). We need to study the respective influence of different scales, which will provide important information about the frequency tuning of bottom-up visual attention. We believe that running the saliency model backwards will afford crucial insights about human visual attention.

4.1 Validation

Our goal is to develop image processing techniques that subtly alter saliency without introducing objectionable visual artifacts. Thus far, we have relied on visual inspection to evaluate the success of the techniques discussed in this paper. We plan to empirically validate our results using eye-tracking, comparing subjects' gaze patterns and fixations for modified and unmodified images. To confirm that the techniques yield quantitatively plausible natural images, we plan to compare the spectral signatures of modified and unmodified images (Olshausen & Field, 1996). It would also be interesting to consider the change in contrast energy across texture boundaries in the before and after images. Specifically, the texture discrimination metric proposed by Rosenholtz could be used to provide a quantitative measure of how much texture boundaries are "blurred" or "sharpened" (Rosenholtz, 2000).

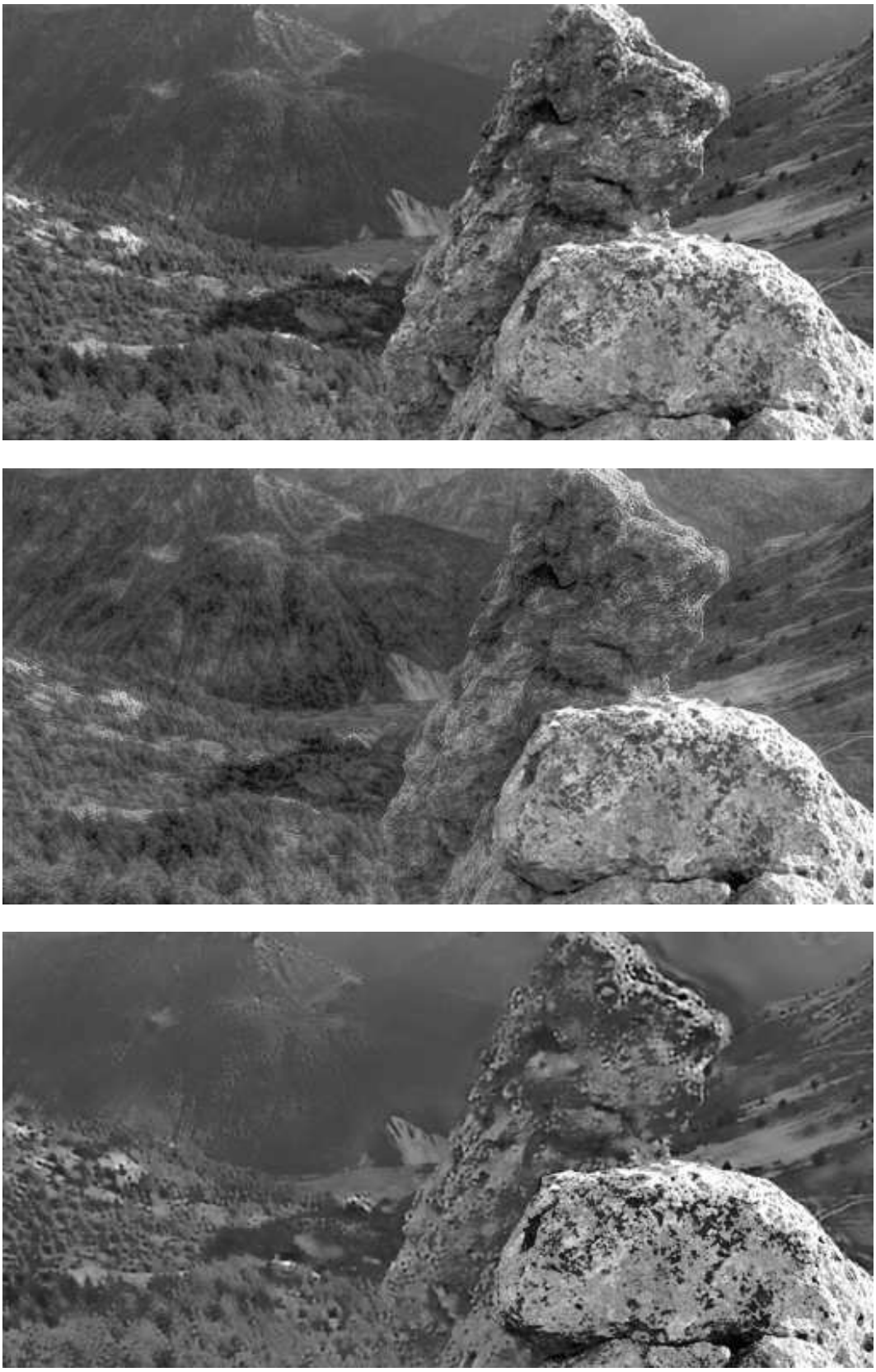

Figure 2. (Top) Photograph taken by an amateur photographer. (Middle) Saliency of the rightmost rock has been increased by decreasing saliency of the background, effectively "blurring" texture boundaries. (Bottom) Global texture variation has been increased, "sharpening" texture boundaries. The binary mask used to segment regions in the original image is shown in Figure 3. Original photograph ©Janne Sinkkonen.

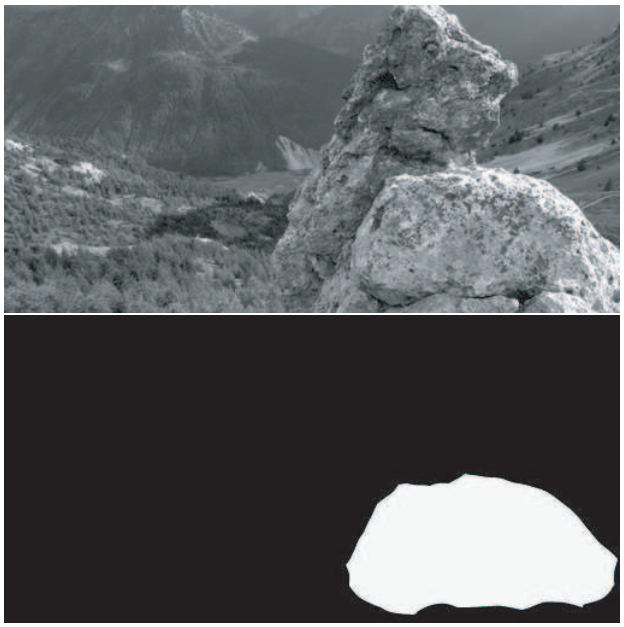

Figure 3. Binary mask used to segment image regions in Figure 2.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. SLS was supported by an NSF Graduate Research Fellowship.

References

Bétrancourt, M., & Tversky, B. (In press). Simple animations for organizing diagrams.

Duchowski, A. T. (2003). *Eye tracking methodology: Theory and practice*. Springer-Verlag.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*, 891–906.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience, 2*, 194–203.

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

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuity. *Human Neurobiology, 4*, 279–283.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607–609.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research, 39*, 3157–3163.

Rosenholtz, R. (2000). *Proceedings of European Conference on Computer Vision* (pp. 197–211).