Adaptive Local Context Suppression of Multiple Cues for Salient Visual Attention Detection

Yiqun Hu, Deepu Rajan and Liang-Tien Chia
Center for Multimedia and Network Technology
School of Computer Engineering
Nanyang Technological University, Singapore 639798
{p030070, asdrajan, asltchia}@ntu.edu.sg

Abstract

Visual attention is obtained through determination of contrasts of low level features or attention cues like intensity, color etc. We propose a new texture attention cue that is shown to be more effective for images where the salient object regions and background have similar visual characteristics. Current visual attention models do not consider local contextual information to highlight attention regions. We also propose a feature combination strategy by suppressing saliency based on context information that is effective in determining the true attention region. We compare our approach with other visual attention models using a novel Average Discrimination Ratio measure.

1. Introduction

Visual attention (VA) refers to the mechanism of the human visual system to concentrate on a certain portion of the visual data presented to it. The attention can be based either on salient objects present in an image or on the *a priori* knowledge of a scene and the ultimate goal of capturing VA, e.g. image adaptation. The former is referred to as a *bottom-up* approach while the latter is *top-down*. Computational models for VA attempt to incorporate such bottom-up/top-down processing. Identifying VA regions is useful for image retrieval [9], object recognition [8], image adaptation [1] etc.

Recently, several computational models of visual attention have been proposed. In [5], Itti et al. develop a VA model based on the behavior and neuronal architecture of the early primate visual system using contrasts among low level features like color, intensity and orientation as attention cues. However, a strategy for combining various contrast maps to capture the true attention region (AR) is still a challenge [4, 3]. Different from four post-processes to suppress noisy maps in [4], different contrast maps are selectively combined according to their attention convex hulls using Composite Saliency Indicator (CSI) in [3]. In [6], Ma and Zhang proposed another VA model only using spatial color contrast as the attention cue for simplicity.

In this paper, we (i) propose the use of texture as an *additional* cue for salient region detection and (ii) develop a robust feature combination strategy that suppresses *regions* in contrast maps that do not contribute to the true ARs. The proposed model uses local context information to suppress spurious ARs while simultaneously enhancing the true ARs. We also propose new evaluation criterion, instead of subjective tests as reported in the literature, in order to determine the effectiveness of the VA scheme described here.

2. Texture Attention Cue

Intensity, color and orientation have been used as cues for VA [5]. Here, we describe a method to recover texture information in an image that serves to generate a texture contrast map that facilitates detection of attention regions. Texture is especially useful to capture VA in images containing small objects present in a cluttered background.

An image is divided into blocks, called *texture patches*, each block containing $p \times q$ pixels. By taking the Gabor Wavelet Transform [7] of the image, each texture patch is represented by the mean μ_{sk} and the standard deviation σ_{sk} of the wavelet coefficients, where s and k stand for scale and orientation, respectively. For S scales and K orientations, we then have SK mean maps $MM_{s,k}$ and standard deviation maps $SDM_{s,k}, s = 1 \ldots S, k = 1 \ldots K$. Since our objective is to capture the contrast/variation in texture over the image, we calculate the Average Mean Difference (AMD) and the Average Standard Deviation Difference (ASDD) over a neighborhood of patches, where AMD for a patch centered at (i, j) is given by
$AMD_{s,k}(i, j) =$

$$\frac{1}{N} \sum_{u,v} |MM_{s,k}(i + u, j + v) - MM_{s,k}(i, j)| \quad (1)$$

and ASDD for a similar patch is given by
$ASDD_{s,k}(i, j) =$

0-7803-9332-5/05/$20.00 ©2005 IEEE

$$\frac{1}{N}\sum_{u,v}|SDM_{s,k}(i+u,j+v)-SDM_{s,k}(i,j)| \quad (2)$$

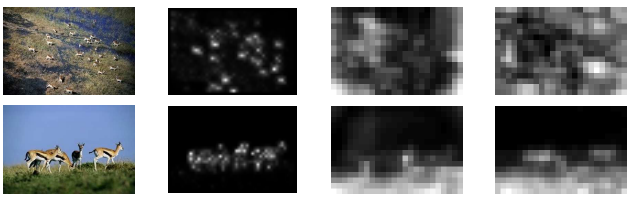
with N being the number of patches in the neighborhood. A measure for texture contrast at a patch (i,j) and at any scale s and orientation k is calculated as

$$TC_{s,k}(i,j) = AMD_{s,k}(i,j) \times ASDD_{s,k}(i,j) \quad (3)$$

while the final texture contrast at patch (i,j) is obtained as

$$TC(i,j) = \sum_{s}\sum_{k} TC_{s,k}(i,j) \quad (4)$$

The above texture cue captures an AR even if other cues like intensity and color fail. There are several situations in which the AR is captured using texture as an aid, e.g. texture foreground in non-textured background, non-textured foreground in texture background, regular texture in a random textured background, random texture in regular textured background etc. Figure 1 shows two examples where texture is useful as a cue to detect salient objects that capture VA. The texture contrast map obtained using the proposed algorithm is shown in Figure 1 (b). Also, we show that intensity and color contrast maps generated by the algorithm described in [5] fail to indicate the ARs in Figure 1 (c) and (d). Note how the texture map is able to enhance the salient objects.



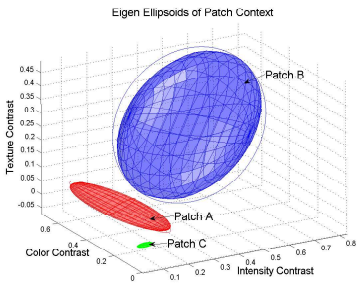
(a) (b) (c) (d)

Figure 1: (a) Original image (b) Texture contrast map (c) Intensity contrast map and (d) Color contrast map using Itti's model [5]

3. Feature combination through Context Suppression

The combination of low level feature contrasts like intensity, color and texture to yield a final saliency map that includes only the true ARs is a hard problem. While some approaches suggest a simple linear combination method [5], others suggest some post-processes [4] or a weighted combination of contrast maps based on their convex hulls [3]. However, such methods alter the contribution of an entire contrast map (e.g.intensity), without considering local contextual information that might be crucial to determining final saliency. For example, if the true AR is highlighted in the intensity map along with spurious ARs, the weight of the map as a whole is reduced instead of suppressing only the outliers. Motivated by the mechanism of non-classical receptive field inhibition for contour detection [2], we argue that local context information is essential to discriminate between true and spurious ARs. The influence of context can be brought about by suppressing the saliency of a region whose neighborhood has similar contrast, through the process of surround suppression. In Figure 2(a), for example, the ARs caused by the background bushes and the grass need to be suppressed to capture the true AR of the antelopes as perceived by the human visual system. Here, we propose a local context suppression strategy to adaptively combine multiple attention cues like intensity, color and texture. Consider an image divided into blocks, each

(a)



(b)

Figure 2: (a) Original image and (b) Eigen ellipsoids for each of the patches in (a)

containing $p \times q$ pixels. We call each block an *Attention Patch*. The variation or contrast of a particular feature at a particular patch centered at (i,j) with respect to its neighborhood is calculated as

$$FV(i,j) = \frac{1}{N}\sum_{u,v}|MF(i,j)) - MF(i+u,j+v)| \quad (5)$$

where $MF(i,j)$ is the mean of the feature in patch (i,j) and N is the number of patches in its neighborhood. The contrasts at patch (i,j) for n features/attention cues $\{FV_1(i,j), FV_2(i,j), ...FV_n(i,j)\}$ are normalized to lie between $[0,1]$. For those features that themselves have components, e.g. color has hue and saturation, we sum up the contrasts for individual components to get the contrast for the feature. Each patch is now represented by the n dimensional feature contrast vector which is compared with other feature contrast vectors in its neighborhood and its

contrast measure (equation (5)) is suppressed if the patch and its neighbors are 'similar'. Similarity is estimated by the variance of data along eigen vectors of an $n \times n$ covariance matrix, called the Attention Cue Covariance Matrix (ACCM). The ACCM is formed from the feature contrast vectors at a patch (i, j) and its neighborhood. The eigen values of ACCM represent the extent of similarity or dissimilarity among the attention cues. A large (small) eigen value indicates large (small) variance along the direction of its corresponding eigen vector, which in turn implies higher (lower) discriminating power. It is shown that regions having such higher discriminating power correspond to the true ARs. Consider the image shown in Figure 2(a) and a 3 dimensional feature contrast vector of intensity, color and texture. We examine 3 patches as potentially true ARs: patch A is from the background bushes, patch C is from the grass and patch B is from the true AR of the antelopes. Figure 2(b) shows eigen ellipsoids corresponding to each patch. For each patch, the ellipsoid is centered at the mean of the feature contrast vectors over the neighborhood of the patch. Each axis of the ellipsoids points towards the eigen vectors of the $ACCM$ while their semi-radii are proportional to the eigen values. As seen from the figure, patch C has small variance along all the axes of the ellipsoid while patch A has small variance along two of its axes. However, patch B has large variance along all the 3 axes indicating higher discriminating ability with respect to its neighborhood and should therefore belong to a true AR, which indeed it does. Thus, it is required that the contribution to the saliency map of patches A and C should be suppressed while that of patch B should be enhanced. The suppression factor (SF) for patch (i, j) is obtained as $\tau(i, j) = \prod_{u=1}^{p} \bar{\lambda}_u$ where the $\bar{\lambda}$'s are sorted in ascending order and the parameter p controls the degree of suppression. The saliency value $S(i, j)$ for patch (i, j) is obtained in two steps: first the multiple attention cues or the contrast maps are linearly combined and the result is modulated by the SF as

$$S(i,j) = \tau(i,j) \times \sum_{u=1}^{k} FV_u(i,j) \tag{6}$$

Figure 3 shows the steps leading to the final saliency map of the image shown in Figure 2(a). The intensity and color contrast maps are obtained using equation (5) and the texture contrast map is obtained as described in Section 2. The linear combination of the contrast maps is implemented simply as their sum and the SF for each patch is displayed as an image with darker regions representing high SF and brighter regions representing low SF. The product of the combined map and the SF yields the final saliency map which contains the true AR. Note that both the color and texture contrast maps have been able to indicate the true AR to some extent. However, the former has more spurious ARs than the latter. Moreover, using the proposed SF, these spurious regions have been successfully removed.

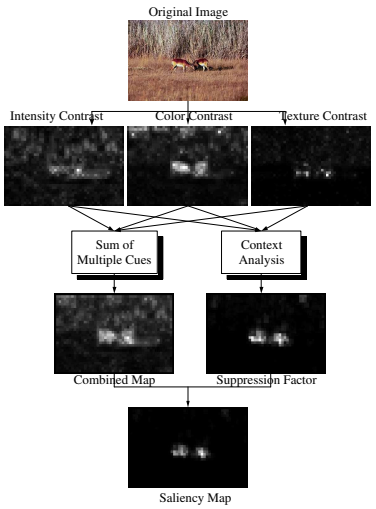


Figure 3: The proposed local context suppression process

4. Experiment Results

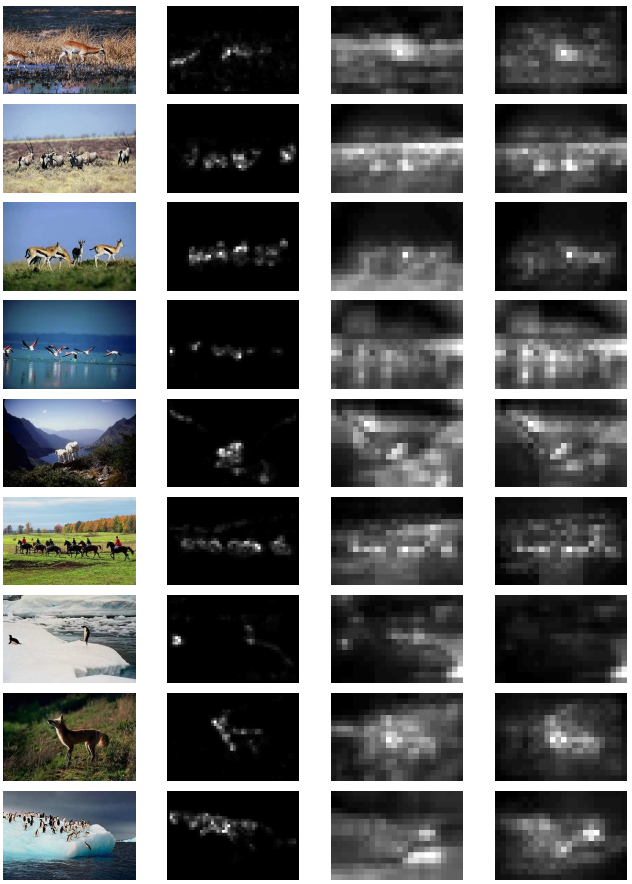
We demonstrate the efficacy of the texture attention cue and new feature combination scheme on images selected from the Corel Photo Library. The images were selected in such a way as to ensure that they contained 'true ARs' as perceived by the human visual system. However, we do realize that a 'true AR' is a subjective phenomenon. As mentioned earlier, we consider 3 cues, the contrasts of intensity ($I = (R+G+B)/3$), color (hue and saturation in the HSV space) and the proposed texture contrast. The size of an Attention Patch was 8×8. To determine the texture map, the Gabor Wavelet Transform used 4 scales and 6 orientations. The parameter p to control the SF was chosen as 2. Figure 4 (a) and (b), respectively, show the original images and the visual ARs detected by the proposed method. We compare the saliency map with those obtained using the algorithms in [5] and [3] whose results are shown in Figure 4 (c) and (d) respectively. It is evident that the proposed method is able to capture the visual AR better than other two methods.

For an objective evaluation, we manually segmented out the salient object(s) and defined an Average Discrimination Ratio (ADR) as

$$ADR = \frac{(\sum_{(i,j)\in\varphi} S(i,j))/|\varphi|}{(\sum_{(i,j)\in\varphi} S(i,j))/|\varphi| + (\sum_{(i,j)\in\psi} S(i,j))/|\psi|} \tag{7}$$

where $|\varphi|$ and $|\psi|$ represent the cardinality of the set of pixels belonging to the salient region and the non salient regions, respectively. Since the true ARs must contain high saliency, the higher the ADR, better is the detection of the true AR. We added White Gaussian Noise to the images and compare the ADR's obtained using our method with those in [5] and [3]. Figure 5 shows the plot of ADR with increasing variance of noise for each of the methods. The

proposed method has significantly higher ADR of above 0.9 compared to the others, implying that the saliency value of the non salient regions is close to zero. It is interesting to note that all the methods perform consistently with increase in noise variance.



(a) (b) (c) (d)

Figure 4: Experiment Results (a) Original Image; Saliency Maps using (b) proposed method (c) Itti's Model [5] (d) CSI [3]

5. Conclusions

In this paper, we propose a useful attention cue of texture information and design a general context suppression model inspired by the mechanism of non-classical receptive field inhibition of the primate visual cortex for attention detection. This model analyzes the context contrast variance adaptively to find the potential similarity of context contrast and uses this factor to highlight the true attention regions and suppress spurious ARs simultaneously. To evaluate the performance of the proposed method, we introduce a criteria to test its discriminating power between salient object regions and background. In the future work, we will integrate scale-space information to our model.

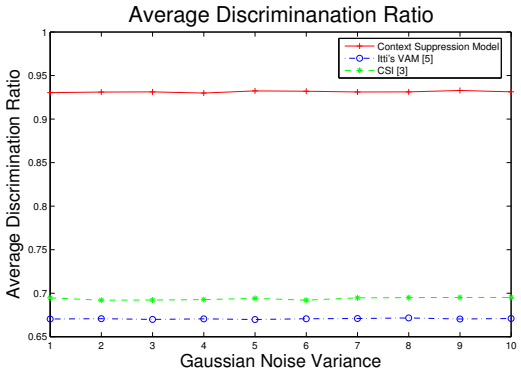


Figure 5: Average Discrimination Ratio

References

[1] 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. *ACM Multimedia Systems Journal*, 9(4):353–364, 2003.

[2] C. Grigorescu, N. Petkov, and M.A. Westenberg. Contour and boundary detection improved by surround suppression of texture edges. *Journal of Image and Vision Computing*, 22(8):583–679, 2004.

[3] Y. Hu, X. Xie, W.Y. Ma, L.T. Chia, and D. Rajan. Salient region detection using weighted feature maps based on the human visual attention model. In *Proceedings of the Fifth IEEE Pacific-Rim Conference on Multimedia*, Tokyo Waterfront City, Japan, November 2004.

[4] L. Itti and C. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99)*, volume 3644, pages 473–482, San Jose, CA, 1999.

[5] 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*, 20(11):1254–1259, 1998.

[6] Y.F. Ma and H.J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, Berkeley, CA, USA, 2003. ACM Press.

[7] B. S. Manjunath and W.Y. Ma. Texture feature for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[8] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition - a gentle way. *Lecture Notes in Computer Science*, 25(25):472–279, 2002.

[9] X.J. Wang, W.Y. Ma, and X. Li. Data-driven approach for bridging the cognitive gap in image retrieval. In *Proceedings of 2004 IEEE International conference on Multimedia and Expo*, Taipei, 2004.