A Corner-based Saliency Model

Wirawit Rueopas, Sangsan Leelhapantu, and Thanarat H Chalidabhongse
Department of Computer Engineering, Faculty of Engineering,
Chulalongkorn University, Bangkok, Thailand.
E-mail: Wirawit.R@student.chula.ac.th, sangsan_momo@hotmail.com, Thanarat.C@chula.ac.th

Abstract— **Biological-based bottom-up saliency models base their computation on low-level features like intensity, color and orientation which input signals are encoded along the retina, lateral geniculate nucleus (LGN), and primary visual cortex (V1) respectively. Visual attention is then evaluated independently on each feature channel and combined later to achieve a final saliency map. In this work, we propose a model that utilizes medium-level features such as corners, line intersections, and line endings to extract possible locations of figures in natural images. Our result on Toronto dataset [1] indicates that the model is competitive among other state-of-the-art models. We also believe these features are essential for object learning and recognition at later stage of visual processing in human.**

Keywords - saliency; visual attention; corner; bottom-up attention

I. INTRODUCTION

Humans receive highly complex and large amount of information every second. Visual attention is a mechanism in human perception that enables us to selectively process and interpret important parts of what we see, thus allows us to act and response efficiently. There are two types of visual attentions: top-down and bottom-up. Top-down visual attention is task-dependent and controlled by organism's goals. For example, we can intentionally search for red, vertical bars or blue circles. In contrast, bottom-up visual attention is a subconscious mechanism that attracts us to conspicuous features. Accordingly, computational model for visual attention can facilitate the processing of visual information in any systems or machines that employ computer vision by suggesting salient locations to process early, reducing the overwhelming amount of visual data to evaluate.

A topographical image that represents the amount of attention paid at each pixel location of corresponding visual scene is called saliency map [2]. Many computational models for bottom-up visual saliency have been proposed using diverse approaches. Biological-based bottom-up models use features from early visual pathway from retina to primary visual cortex such as intensity, colors and orientations. Based on feature integration theory, these features are processed in parallel and combined later [3]. Information theoretic approaches often employ definition of probability on image patches, and promote those with high rarity.

In this work, we propose a model for bottom-up visual saliency using corner features to cue figure locations. Note that the term "figure" comes from the problem of finding objects from the background called figure-ground segregation. The model requires no prior knowledge about objects. Based on assumption that figures usually reside on locations with dense corner information and thus are salient, the model extract tentative corner information in parallel on channels by multiplying responses at different orientations in a pixelwise manner, and then finally combining together to yield a saliency map. Our model is different from many biological-based models that use low-level features like intensity, color and orientation and models that explicitly employ high-level grouping mechanism to extract figures [4], in that the proposed model uses medium-level features like corner, line intersection and line ending, which are found to play an important role in further shape processing [5], to suggest figure locations. The model is evaluated on Toronto Dataset [1]. It performs well on natural images and the results are competitive to other state-of-the-art models. We will also show why representing visual input with corners and line endings opens up possibilities to cover illusion contour and shape processing later on.

The remainder of this paper is organized as follows. Section 2 reviews related works in bottom-up computational models. The proposed model is described in detail in Section 3, which is followed in Section 4 by experimental results. Section 5 discusses on weaknesses and motives of the model. Finally, Section 6 concludes this work and suggests further improvements.

II. RELATED WORKS

One of the classic bottom-up computational model is the work by Itti and Koch [3], which is based on feature integration theory by Treisman [6]. According to the model, stimulus' features are extracted automatically and in parallel, such as intensity, color and orientation feature as used in [3]. Features that are distinct from its surrounding will pop out and others will get suppressed. Another similar approach to Itti's model is the bottom-up saliency calculation part of VOCUS [7] by Frintrop. One difference is that in VOCUS intensity channel is separated into on- and off- center types instead of one channel in Itti's model. This is to manage responses from both dark and light figure equally.

Another important biological-plausible viewpoint is proto-object based computational model, which propose that human

978-1-5090-2033-1/16/$31.00 ©2016 IEEE

978-1-5090-2033-1/16/$31.00 ©2016 IEEE

perceives whole objects, not individual features. Examples of proto-object based models are [8] and [4]. In computational term, proto-object is a region that combines similar pixels together into a coherent region, in which attention is distributed to the whole rather than each pixel in feature-based model. Recent proto-object based model [4] also utilizes border ownership and grouping cells that work together to decide owners of the border of a figure. It also has been shown to exhibit continuity and proximity from Gestalt principles of organization.

In information theoretic approach, Bruce and Tsotsos [1] used Shannon's self-information (logarithm of probability distribution) to define rarity of features. Image patches are extracted from input image and transform into lower dimensional feature vectors of basis functions by using independent component analysis (ICA). Probability distribution function for each feature channel is created for mapping each feature with its corresponding probability. The lower the probability, the higher saliency value at that location.

There are also many other approaches to compute saliency map, usually applying their own terms of feature rarity. Graph based approach [9] construct fully-connected directed graph with definition of dissimilarity to compute saliency map. Hou et al. [10] evaluated properties of natural images in frequency domain. Logarithm of spectrum of image is subtracted from the average spectrums of other images to remove redundancies and the result is a spectral residual that retain novel locations. Borji and Itti [11] calculated local and global image patch rarities separately for each color channels.

A thorough review of computational saliency models can be found in [12].

III. PROPOSED MODEL

Based on the assumption that humans usually pay more attention to regions with dense corner locations since they define contour of figures from ground, the core of our model is tentative corner extraction. Our model consists of three main steps: image representation, corner extraction and channel combination as illustrated in Fig. 1. First in image representation an input image is separated into intensity channel and 4 color opponency channels[3]. The intensity map is separated into 2 edge feature maps: on-center and off-center, to detect both bright and dark figures on its contrary backgrounds, according to [7]. For color channels, high-pass Laplacian filters are applied on each channel to obtain edge information. Laplacian filter is applied directly to image at original size, where edge corner information is most accurate. Next, these edge feature maps are convolved with Gabor filters at 4 different orientations, to express orientational selectivity in primary visual cortex (V1). Then, orientational responses on each channel are combined independently by multiplying these feature maps, to promote only locations with strong possibility for corner that has multiple orientational responses, and implies corner. The tentative corner maps are then suppressed by average of surrounding signals, normalized and finally summed to obtain a saliency map.

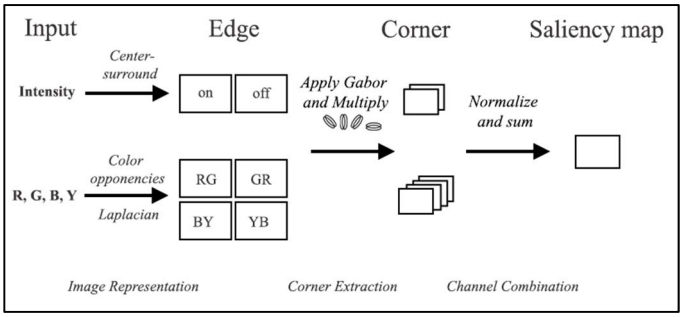

Figure 1. Diagram of a proposed model. Input image is split into R, G, B, Y and intensity channels. Center-surround (for intensity), Color opponency with Laplacian operator (for colors) are used to obtain edge features, yielding six maps. Each is convolved with Gabor filter at four different orientations and multiplied together in pixelwise manner to obtain corner features. These are then normalized and summed together to get a saliency map.

Details of each step are described in the following subsections: image representation, corner extraction and channel combination.

A. Image Representation

An input image is first separated into 4 channels: red (r), green (g), blue (b), and intensity channel (I) which is obtained by

$$I = \frac{r + g + b}{3}. \tag{1}$$

Intensity: On-center and Off-center are extracted using method similar to [7]. First Gaussian pyramids are constructed from intensity image, by downsampling a 5x5 Gaussian filter response of the image. In contrast to [7] where 5-band pyramid is used, our model use only first two bands. Each band is further divided into two intensity feature types: on-center-off-surround (or on-center) and off-center-on-surround (or off-center).

Center is determined directly from pixels at each location, and surround is computed by averaging all pixels in radius. On-center and off-center features given a scale (s) and a surround radius (rd) are obtained by:

$$Cs_{on,s,rd}(x,y) = \left| I_{s,rd}(x,y) - \frac{1}{N} \sum_{i,j \in [-rd,rd]} I_{s,rd}(x+i, y+j) \right|$$
$$Cs_{off,s,rd}(x,y) = \left| \frac{1}{N} \sum_{i,j \in [-rd,rd]} I_{s,rd}(x+i, y+j) - I_{s,rd}(x,y) \right| \tag{2}$$

where I is the intensity image, s is level of Gaussian Pyramid, rd is radius of surrounding region, N is total number of pixels in surrounding area which is $(2 * rd + 1)^2$ in this case, and the operator $\lfloor \rfloor$ is half-wave rectifier which is defined as:

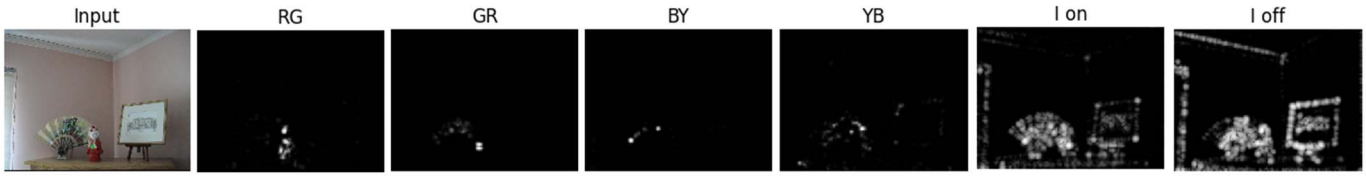

Figure 2. Figural Cues for each channel, obtained from (8).

$\lfloor x \rfloor = max(0, x)$. Since we assume baseline of any response starts at 0, negative values are ignored, in other words, neuron does not fire below this threshold.

This yields 4 maps for each intensity type (on and off), scale s = {1,2}, surround radius rd = {3,7}, so a total 8 maps are computed. For each type, maps at coarser scale are resized up to finer scale by linear interpolation and all are summed together pixelwise, yielding one on-center (I_{on}) and one off-center (I_{off}) map.

Color: In Lateral geniculate nucleus (LGN), color signals are also encoded by On- and Off-Center cells as intensity, but in two particular pairs of colors which is called color opponency [3]. This is due to different wirings of three cone types in the retina each tuned for different wavelengths projected to LGN [13]. The opponency system consists of red-green (RG), green-red (GR), blue-yellow (BY) and yellow-blue (YB). For example, red-green denotes on for red at center, but off for green at surround. These channels are generated according to:

$$\begin{aligned} R &= \left\lfloor r - \frac{g+b}{2} \right\rfloor \\ G &= \left\lfloor g - \frac{r+b}{2} \right\rfloor \\ B &= \left\lfloor b - \frac{r+g}{2} \right\rfloor \\ Y &= \left\lfloor \frac{r+g}{2} - \frac{|r-g|}{2} - b \right\rfloor, \end{aligned} \tag{4}$$

then opponency maps are computed as follows:

$$\begin{aligned} RG &= \lfloor R - G \rfloor \\ GR &= \lfloor G - R \rfloor \\ BY &= \lfloor B - Y \rfloor \\ YB &= \lfloor Y - B \rfloor. \end{aligned} \tag{5}$$

Laplacian filter is then applied independently to extract edge information for each color channel.

In total, there are 6 maps to represent edge information: two for intensity ($I_{on}$and I_{off}), and four for color opponencies (RG,GR,BY and YB).

B. Corner Extraction

At this stage, corner, line intersection and line ending locations are extracted by multiplying orientational feature maps together, as they usually respond to multiple orientations.

In V1, simple cells are tuned preferentially for different orientations and their receptive fields can be approximated with Gabor filter [14]. Each map from the previous step is convolved with Gabor filters at 4 orientations. Impulse responses of 2D Gabor filters are products of plane sinusoidal waves (cosine for real part and sine for imaginary part) and 2D Gaussian functions. Energy mechanism [15] is applied to yield an amplitude map (A) for each orientation:

$$A_{\Psi,\theta}(x,y) = \sqrt{G_{\Psi,c,\theta}(x,y)^2 + G_{\Psi,s,\theta}(x,y)^2}\,. \tag{6}$$

$G_{\Psi,c,\theta}$ and $G_{\Psi,s,\theta}$ are Gabor filter responses of input for real and imaginary parts respectively for input channel Ψ, and $\theta \in \{0°, 45°, 90°, 135°\}$. So four orientational feature maps are computed with the edge map from each channel. These four maps are then multiplied together directly to obtain corner feature (Crn):

$$Crn_{\Psi}(x,y) = \prod_{\theta \in \{0°,45°,90°,135°\}} A_{\Psi,\theta}(x,y)\,. \tag{7}$$

However, this usually results in a highly sparse map for corner locations, since edge information are not often clear in natural images. This is the reason why the Gabor filter cannot efficiently extract orientational information at cluttered scene where figures have no clear boundary. To promote more figural cues in a saliency map, natural logarithm function is applied on the response to obtain figural cue feature (Fc):

$$Fc_{\Psi}(x,y) = \ln(Crn_{\Psi}(x) + 1)\,. \tag{8}$$

It can also be viewed as adjusting the range of orientational responses with logarithm function and summing them up directly. Note that this process is used in natural images and is also the core process of our model. To visualize the original tentative corners locations in images, Crn is used instead of Fc.

With this, six figural cue feature maps are obtained, one for each channel. Examples of Fc response for each channel are shown in Fig. 2.

C. Channel Combination

Since figural cue responses for any channels are comparable, the maps are normalized into the same range of [0,1]. However, the figural cue responses in the same map are not equally important. Individual clusters of the response usually indicate main figure locations and are more salient. To promote individual cluster of figural cues, the response is subtracted from the average of surrounding pixels and clipped negative values with the half-wave rectifier:

$$Fc'\psi(x,y) = \left| Fc\psi(x,y) - \frac{1}{N} \sum_{i,j\in[-\rho w,\rho w]} Fc\psi(x+i,y+j) \right|. \qquad (9)$$

N is number of pixels in the rectangular surrounding area with horizontal and vertical radius of ρw. Then the Fc' response map for each channel is adjusted its significance by dividing with the square root of number of local peaks as in [7] to bias the channel with few local maximums. Here ρw is one-tenth of the input image width and local peaks are calculated with minimum distance 3.

All six normalized figural cue maps of all corresponding channels: on and off intensity maps (I_{on} and I_{off}) and four color opponency maps (RG, GR, BY and YB) are combined together by pixelwise addition, downsampled by linear interpolation and post-processed by convolving with Gaussian filter. Examples of saliency maps generated are shown alongside other models in Fig. 6.

IV. Experimental Results

The proposed model was tested on Toronto dataset [1] which contains 120 color images with resolution of 511 x 681, mostly indoor and outdoor environments. Random images are shown and eye fixations are recorded with viewing time 3 seconds on 20 subjects to obtain the ground truth for human fixation density.

For algorithm evaluation, we use AUC-shuffled [16]. AUC is the area under Receiver Operating Characteristics (ROC) curve used to measure performance of binary classifier, in this case, a saliency map. The original AUC [1] takes recorded eye fixations of human as positive set and random fixations (at uniform distribution) as negative set. First thresholding is applied on the saliency map to make a binary image. Then, fixations that fall in the "on" area are counted as true positive and those that fall in the "off" area as false positive. All possible thresholds are computed and ROC graph is plotted from true-positives and false-positives. Area under this curve is then computed to obtain AUC score. However, AUC and previous used scores such as Correlation Coefficient (CC) [17] and KL Distance [18] suffer from center bias. Center bias refers to how viewer usually makes eye fixations at center of the image. So the saliency map with only Gaussian blob at the center yields high KL value regardless of image content [16]. This allows any algorithms to gain increase in performance by masking saliency maps with Gaussian blob at the center. To workaround this bias, AUC-shuffled [16] instead collects eye fixations from other images (not containing fixations from the current image) and to be used as negative set, in contrast with AUC that uses uniform distribution over all the image. Since possible human fixations are used, it implies that center bias is also in the negative set and helps adjust the false positive calculation.

However, AUC-shuffled is sensitive to final blurring on a saliency map [19]. To evaluate the overall performance of a model, final saliency maps are convolved with Gaussian kernel at various standard deviations (STD). Average of AUC-shuffled scores of 120 images is then plotted as function of standard deviations of the Gaussian kernel.

Other six saliency models (Itti [3], GBVS [9], AIM [1], SUN [16], SUN-small, Signature [19]) are compared with our model denoted as Cor (8), using publicly available codes at the respective authors' website. Examples of saliency maps from each model are shown in Fig. 6. Note that SUN-small used input images in smaller size than SUN (one-tenth of original image size). The relationship of standard deviations of final blurring to average of AUC-shuffled scores on 120 images is shown in Fig. 4. Mean and optimal STD scores for each model are also shown in Table I. Our method (denoted as Cor) used (8) for figural cues and this is our main model. Adjusting responses by logarithm allow more corners to be seen. Gaussian filtering at final stage also helps to cover areas adjacent to corner response that implies figure locations, as shown in Fig. 4. Our model significantly increased in performance initially as more STD of Gaussian blur is used and reached the peak at STD of 0.06 of the image width.

Overall, our saliency model based on figural cue can suggest tentative figure locations and the model is very competitive among other approaches.

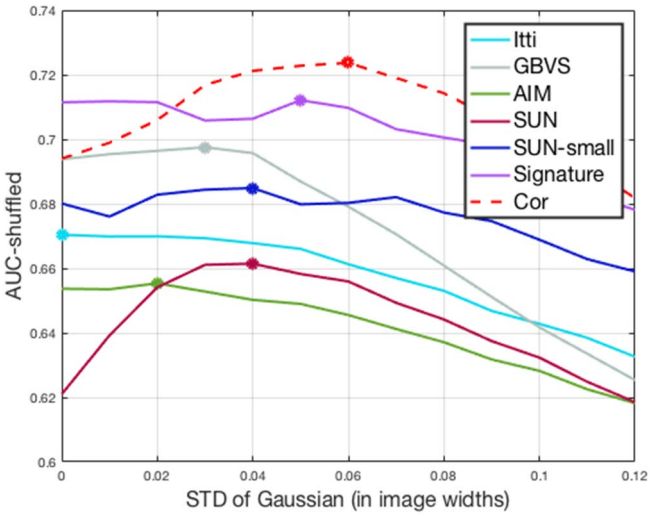

Figure 3. The plot of average of AUC-shuffled scores of 120 images on Toronto dataset as a function of standard deviations of Gaussian kernel (in image widths) used in final blurring of saliency map generated. The dots represent optimal STD values for each model.

TABLE I. 0PERFORMANCE OF 7 ALGORITHMS ON TORONTO DATASET

Algorithm	AUC-shuffled (mean STD)	AUC-shuffled (optimal STD)
Itti et al. [3]	0.6573	0.6704
GBVS [9]	0.6714	0.6975
AIM [1]	0.6415	0.6554
SUN [16]	0.6429	0.6615
SUN-small	0.6764	0.6849
Signature [19]	0.7016	0.7121
Cor (8)	**0.7074**	**0.7237**

V. DISCUSSION

Firstly, in the proposed model, we generate saliency map from tentative corner features. It must be emphasized that our simple corner feature extraction method is not accurate on natural images and gives highly sparse responses. Despite these drawbacks, it is considered in our work mainly because it builds directly on the edge features in the previous stage and because of its ease of implementation. Moreover, we can strengthen the response by logarithm function and apply final blurring to cover more figures that usually reside in corner locations.

Secondly, saliency models based on Feature Integration Theory ([3],[7]) compute intensity, colors and orientation features in parallel, and finally distinct feature maps are promoted by apply weighting and normalization functions. In contrast, our model does not follow the theory and thus do not exhibit similar results to psychological tests such as orientational "pop out" as in other models. Our model operates mainly on corner features to probe for possible figure locations.

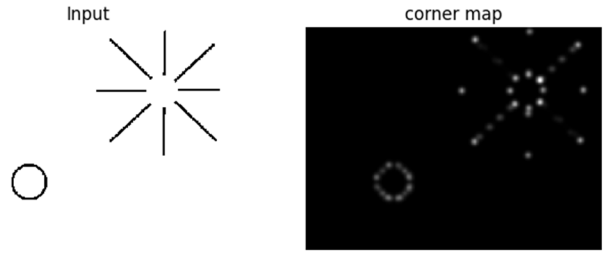

Figure 4. A corner feature map that can be processed further to perceive the upper-right figure as a circular shape with illusory contour. The lower-left circle is included for comparison.

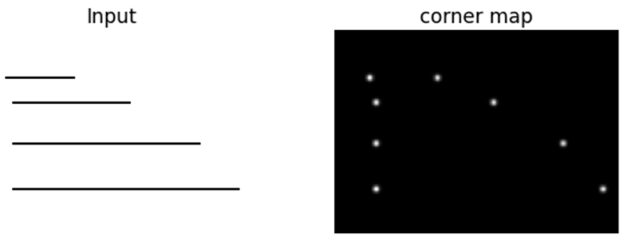

Figure 5. A corner feature map for lines at different lengths. The corner features can imply the length of a line and remove redundancy at intermediate intervals.

Last but not least, the main reason we consider corner and line intersection as features for saliency modeling is because of its possibility to generalize and be used in higher-order visual processing. After the intensity, colors and orientation features extraction from the retina to LGN and V1, extraction of features of intermediate complexity like corner, line intersection and contour is certainly an important next step in intermediate shape processing in V4 [5]. Possibly with the aid of long-range horizontal excitation mechanism in V1[20], some of human visual phenomenon like illusory contours can be achieved.

This phenomenon is emphasized in Fig. 4 and 5. Corner features are extracted from input images into corner maps using (7). Each channel is normalized into range [0,1] and sum up directly.

Fig. 4 shows that corner and line ending feature responses are generalized to cover both circular shape on the lower left, and implicit circular shape on the upper right that can be detected later on with illusory contour mechanism. In this way, connecting dots along contour can help uncover illusory shape perceived by humans. However, prior knowledge about shapes is still needed to connect these dots correctly.

Similarly, the type of shape might be line. Fig. 5 shows corner responses for lines of variable lengths. The intermediate visual information along the lines may not be equally important as a pair of corner responses at the beginning and the end, since they directly determine the length of the line. Note that this was not meant to ignore those visual responses at intermediates completely since endpoints alone cannot sufficiently determine shape. Both orientation and corner information must be integrated to yield accurate shape representation for further visual processing.

VI. CONCLUSION

We proposed a bottom-up saliency model based on corner features to suggest possible figure locations in a natural image. The proposed model is simple to implement but very competitive among other saliency models. Further processing can improve the model either by using better corner extraction algorithm to efficiently extract corners in natural scenes, or employing a more sophisticated method to predict figure locations near those corners. Moreover, traditional biological saliency models operate on three common low-level features: intensity, color and orientation channels. A middle-level feature like corner might be the fourth important link to connect saliency computation from low-level to higher-level features like shape.

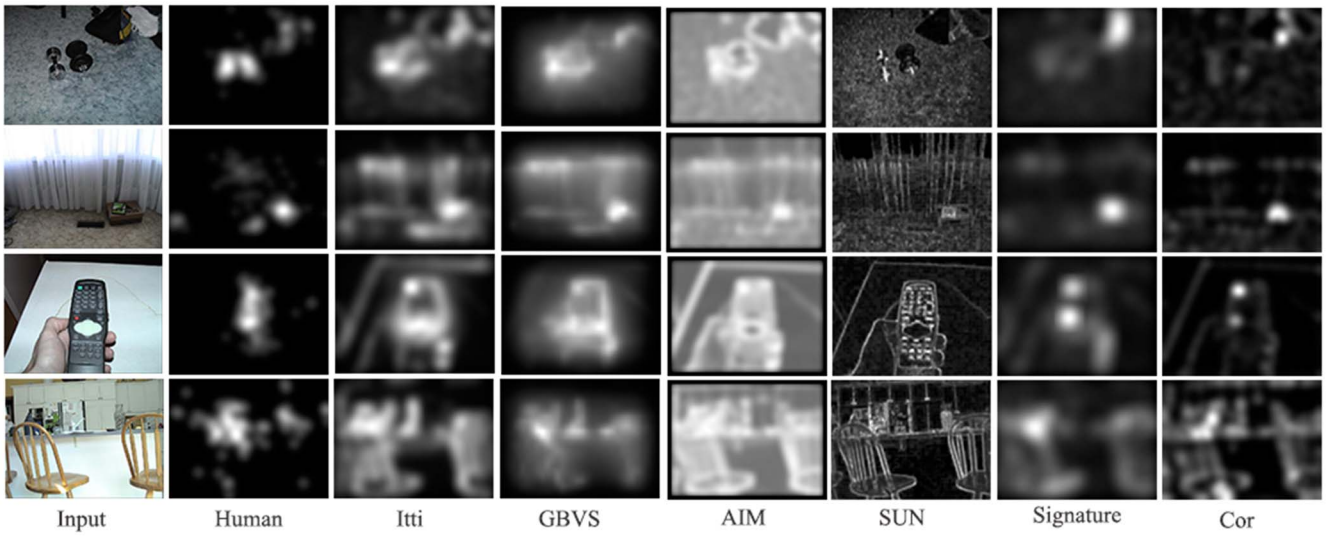

Figure 6. Comparison of saliency maps generated from 6 models. The first and second column is inputs and human fixation density maps respectively from the Toronto dataset. The proposed model is at the last column (denoted as Cor).

REFERENCES

[1] J. K. T. Neil D. B. Bruce, "Saliency Based on Information Maximization.," *Adv Neural Inf Process Syst*, vol. 18, 2005.

[2] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," in *Matters of Intelligence*, L. M. Vaina, Ed. Springer Netherlands, 1987, pp. 115–141.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 11, pp. 1254–1259, 1998.

[4] A. F. Russell, S. Mihalaş, R. von der Heydt, E. Niebur, and R. Etienne-Cummings, "A model of proto-object based saliency," *Vision Res.*, vol. 94, pp. 1–15, Jan. 2014.

[5] A. Pasupathy and C. E. Connor, "Responses to contour features in macaque area V4," *J. Neurophysiol.*, vol. 82, no. 5, pp. 2490–2502, Nov. 1999.

[6] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[7] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, vol. 3899. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[8] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw. Off. J. Int. Neural Netw. Soc.*, vol. 19, no. 9, pp. 1395–1407, Nov. 2006.

[9] "Graph-Based Visual Saliency - 3095-graph-based-visual-saliency.pdf." [Online]. Available: https://papers.nips.cc/paper/3095-graph-based-visual-saliency.pdf. [Accessed: 20-Mar-2016].

[10] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, pp. 1–8.

[11] A. Borji, "Exploiting Local and Global Patch Rarities for Saliency Detection," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2012, pp. 478–485.

[12] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[13] L. M. Hurvich and D. Jameson, "An opponent-process theory of color vision," *Psychol. Rev.*, vol. 64, Part 1, no. 6, pp. 384–404, Nov. 1957.

[14] "Jones, J. P. & Palmer, L. A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J. Neurophysiol. 58, 1233-1258." [Online]. Available: https://www.researchgate.net/publication/19719670_Jones_J_P_Palmer_L_A_An_evaluation_of_the_two-dimensional_Gabor_filter_model_of_simple_receptive_fields_in_cat_striate_cortex_J_Neurophysiol_58_1233-1258. [Accessed: 20-Mar-2016].

[15] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Am. A*, vol. 2, no. 2, pp. 284–299, Feb. 1985.

[16] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 32.1–20, 2008.

[17] N. Ouerhani, R. von Wartburg, H. Hugli, and R. Muri, "Empirical Validation of the Saliency-based Model of Visual Attention," *ELCVIA Electron. Lett. Comput. Vis. Image Anal.*, vol. 3, no. 1, pp. 13–24, Dec. 2003.

[18] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Res.*, vol. 49, no. 10, pp. 1295–1306, Jun. 2009.

[19] X. Hou, J. Harel, and C. Koch, "Image Signature: Highlighting Sparse Salient Regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, 2012.

[20] W. H. Bosking, Y. Zhang, B. Schofield, and D. Fitzpatrick, "Orientation Selectivity and the Arrangement of Horizontal Connections in Tree Shrew Striate Cortex," *J. Neurosci.*, vol. 17, no. 6, pp. 2112–2127, Mar. 1997.