2012 10th IAPR International Workshop on Document Analysis Systems

How Salient is Scene Text?

Asif Shahab, Faisal Shafait, Andreas Dengel
German Research Centre for Artificial Intelligence
DFKI, Kaiserslautern, Germany
Email: {FirstName.LastName}@dfki.de

Seiichi Uchida
Kyushu University
Fukuoka, 8190395, Japan
Email: uchida@ait.kyushu-u.ac.jp

***Abstract*—Computational models of visual attention use image features to identify salient locations in an image that are likely to attract human attention. Attention models have been quite effectively used for various object detection tasks. However, their use for scene text detection is under-investigated. As a general observation, scene text often conveys important information and is usually prominent or salient in the scene itself. In this paper, we evaluate four state-of-the-art attention models for their response to scene text. Initial results indicate that saliency maps produced by these attention models can be used for aiding scene text detection algorithms by suppressing non-text regions.**

***Keywords*-character localization;visual attention models; saliency map**

I. Introduction

Detection of text in natural scenes is a challenging problem. Firstly, because of the high dimensionality of colored images and secondly, because of the variety of color, font, size and orientation in which the text can occur in a natural scene. Detection of scene text might be a complex problem for computers but humans seem to detect text without any problems however inconspicuous it is. This is because of our ability to simultaneously process different channels of information and focus our attention on an interesting target, e.g., the text on road sign or the advertising on hoarding in a complex scene.

A lot of research is being done in the vision community on how to accurately model human attention in order to extract regions of interest, so called "salient regions" from an image. Several models of visual attention have been proposed in the literature which can be broadly classified into a) bottom–up, b) top–down methods and c) Bayesian or hybrid model. These models present the result of analysis in the form of a saliency map in which the saliency of a pixel is represented by its gray scale value. Some saliency maps are shown in Figure 1(c)-(g).

Bottom-up models of visual attention [1], [2] use local features in a given image to find image locations which are considerably different from their neighbors. These methods normally work in three steps. 1) *Feature Extraction:* Intensity, color, orientation and motion features are extracted from the image at different scales. 2) *Activation:* saliency is computed by either center–surround [1], or graph-based random walks [2] using multiple features for each of the feature maps. 3) *Normalization and Combination:* saliency maps based on different features are normalized and linearly added to give a master saliency map. These methods are task independent as they do not use any prior information about the object location or shape.

Top-down models of visual attention use prior contextual knowledge of object location and its shape to guide the saliency map. They are task–dependent and based on the fact that search for an object in an image by humans is usually directed and governed by context, e.g., in the task of searching for pedestrians human will focus their attention on bottom of the image near road rather high up in the sky. Recently Torralba et al. [3] proposed a model trained on image features using the collected eye tracking data.

Recently, a hybird model of visual attention is proposed by various researchers, which attempts to model human attention in a Bayesian framework combining the bottom–up saliency model and top–down contextual information of object location and appearance [4] [5] [6]. Such a model gives for each image location the probability of finding the given object. These models usually estimate a probability density of filter responses obtained from local image features (bottom-up saliency) for a given image and combine it with the probability density of object shape (shape prior) and object location (location prior) learned from the training samples, in a Bayesian framework.

Saliency maps and visual attention models have been used in many vision tasks such as scene classification [7], object detection [8], [9] and visual search [5]. However, the use of attention models for the task of text detection is relatively new [6], [10], [11].

Vision studies based on eye tracking experiemnts have shown that faces and text attract human attention as is evident by early fixations on text regions [12]. In this paper we evaluated four different methods of visual attention for the task of text detection in natural scenes. The key contribution of this paper is the comparison of different methods of visual attention and identification of the best

This work was partially funded by the BMBF (German Federal Ministry of Education and Research), project INBEKI (13N10787) and Perspecting (01 IW 08002).

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.42

method that can perform well to separate non–text elements from text regions in early stages of text detection.

II. Models of Visual Attention

A. Itti's Model

Itti et al. [1] proposed a bottom-up model of visual saliency which is neurologically inspired and uses feature integration theory to find salient image locations. Their model divides the given image into different channels namely Colour (C), Intensity (I) and Orientation (O). A dyadic Gaussian pyramid is used which progressively low-pass filter and sub-sample the image from scale 0 (1:1) to scale 8 (1:256) in 8 octaves. Feature vectors are computed using linear "center-surround" operations akin to visual receptive field.

The center-surround is implemented as a difference between center (fine) and surround (coarse) scale. The centre is the pixel at scale $c \in \{2, 3, 4\}$ and surround is the corresponding pixel at scale $s = c + \sigma$, with $\sigma \in \{3, 4\}$ [1]. They calculate six different maps for Intensity. Similarly, 12 color maps are generated using specialized double-opponent colors such as red–green and blue–yellow. For orientation, they used Gabor filters tuned to 0, 45, 90 and 135 degrees and calculated the response of these filters on intensity values. The filter responses are sub-sampled and 24 orientation maps are obtained for orientation using the center-surround operation.

Since, these feature maps are extracted by different methods and thus have different dynamic ranges. Simply combining these feature maps might result in suppression of weak peaks found in one of the map. Therefore, Itti et al. proposed a number of normalization schemes such as Iterative, Local-Max [13]. In the first step the feature maps for intensity, color and orientation are normalized and linearly added to calculate respective conspicuity maps. These conspicuity maps are further normalized and linearly added to give the final saliency map.

B. Harel's Graph Based Visual Saliency Model

Harel et al. [2] proposed a bottom-up model of visual saliency which uses the same image features as that of Itti's, but defines Markov chains over various image maps and uses the equilibrium distribution over map locations for calculating the activation map (conspicuity map) and saliency maps. They construct a fully connected directed graph joining all the nodes (pixels) of the featuremap and assign weight to the edges proportional to the dissimilarity (log ratio of values) between the nodes and their spatial closeness. They define a Markov process on such a graph and estimate the equilibrium distribution of such a chain. The result is an activation map or conspicuity map derived from pairwise contrast.

These activation or conspicuity maps are later normalized using the same Markovian process, this time constructing the graph from nodes in activation map. The normalized activation maps are later combined to give a final saliency map. We used the Matlab implementation of Harel's method[1]. Some sample saliency maps are shown in Figure 1(e).

C. Torralba's Model

Torralba et al. [4] proposed a hybrid model of visual attention which defines an image saliency in a Bayesian framework. In the Bayesian framework, the probability of finding an object $p(O = 1, X|L, G)$ at a location $X = (x, y)$ given the set of local measurements $L(X)$ and a set of global features G can be expressed by:

$$p(O = 1, X|L, G)$$
$$= \frac{1}{p(L|G)} p(L|O = 1, X, G) p(X|O = 1, G) p(O = 1|G)$$

The first term, $1/p(L|G)$, is the bottom–up saliency factor that represents the inverse of probability of finding local measurements in an image. This term is an integral part of a Bayesian framework and corresponds to the bottom–up saliency computed in Itti's and Harel's model.

The second term, $p(L|O = 1, X, G)$, represents the top-down knowledge of target appearance and how it contributes to the object search [4]. The third term, $p(X|O = 1, G)$, provides the context based information and serves as a Bayesian prior. This factor represents the top–down knowledge of object presence at the given location (location–prior) and can be learned from training samples. The fourth term, $p(O = 1|G)$, represents the probability of finding an object in the scene.

Here we are only interested in the evaluation of saliency for scene text without any prior information about its presence, location or appearance. Thus we chose to use only the bottom up salinecy factor($\frac{1}{p(L|G)}$) which uses local image features to calculate saliency.

Steerable pyramid filters tuned to six orientations and four scales are used to generate local image features as in [4]. Raw RGB channels are fed to the bank of filters to generate a set of (6×4×3=72) features, L, for each image location (x, y). Saliency estimation requires estimating the distribution of local features in the image. We used multivariate Gaussian distribution to estimate the saliency values at each image location as explained in [11], [14]. We also used image intensity as a separate channel and computed the response of steerable pyramid filters. A multivariate Gaussian distribution is estimated which results in saliency maps for intensity. Sample results are shown in Figure 1(c),(d).

D. Zhang's Fast Saliency Model

Zhang et al. [15] also proposed a hybrid model of visual attention, which attempts to calculate human attention in a

[1] http://www.klab.caltech.edu/~harel/share/gbvs.php

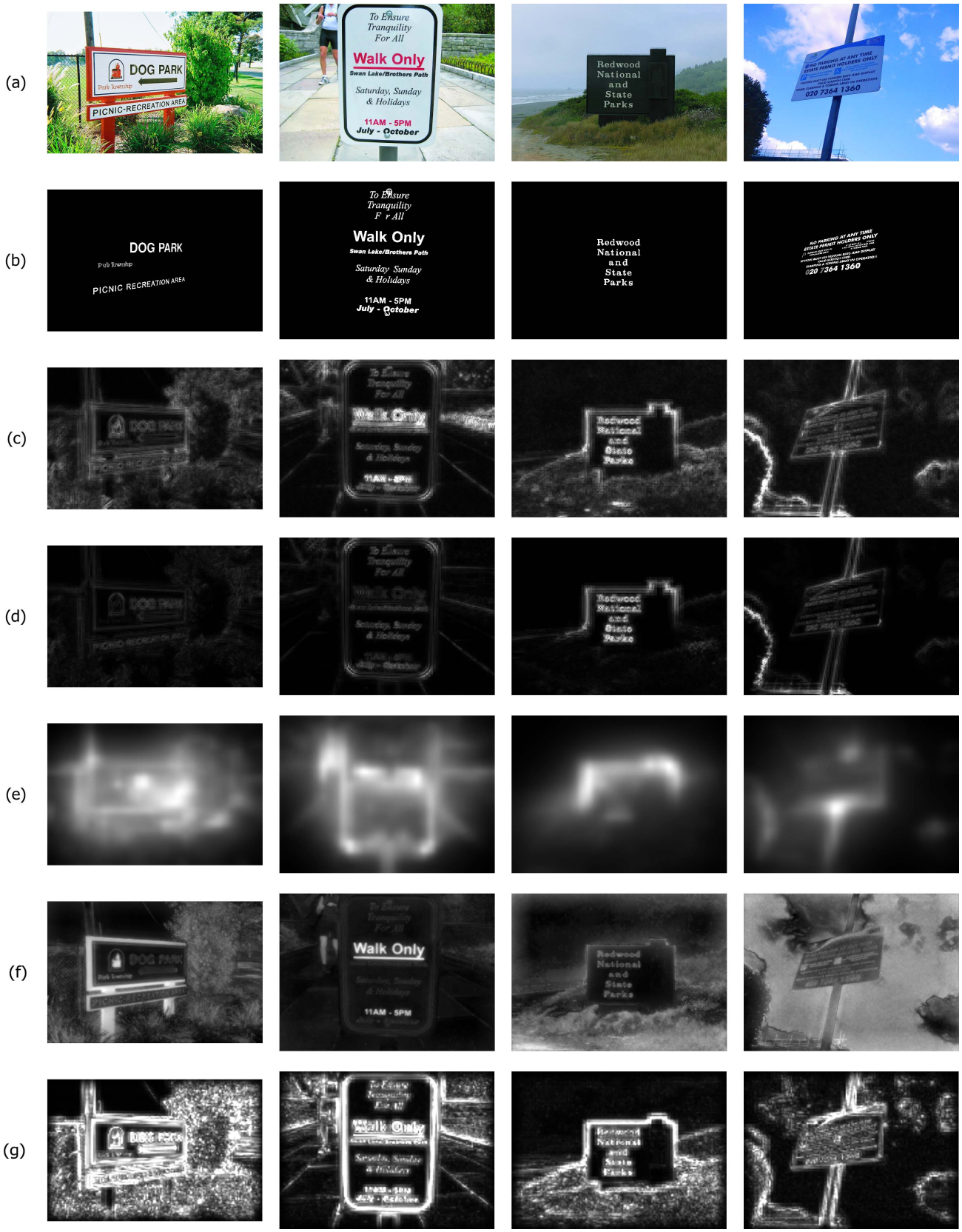


Figure 1: (a) Sample Images, (b) Ground Truth, (c) Torralba's saliency map (Color), (d) Torralba's saliency map (Intensity), (e) Harel's GBVS, (f) Zhang's Fast Saliency, (g) Itti's saliency map (N2P2CI)

Bayesian framework. Their formalism is similar to that of Torralba's [4]. However, they proposed using difference of Gaussian filters (DoG) for the calculation of local saliency. They used image intensity as input and the filter responses are estimated by a mutlivariate generalized Gaussian distribution. Zhang's method was later optimized by Nich et al. [16] for robot vision by the use of difference of box filters (DoB) and estimating a Laplacian distribution of unit variance. We used their C++ implementation of fast saliency[2] [16]. Sample results are shown in Figure 1(f).

III. Evaluation

A. Dataset

We used the scenery image dataset prepared by Uchida et al. [10]. Using Google Image Search, top 300 photo images (each of which containing some characters and has around 640×480) were first collected. The keywords used in the search were "park" and "sign". Some sample images from the dataset are shown in Figure 1(a). For each image a ground–truth (i.e. character and non–character labels) is attached to each pixel manually. Note that small characters have ambiguous boundary and thus their ground–truth became inevitably rough (like a bounding box). Ground truth for some of the images from dataset is shown in Figure 1(b).

We could not use the ICDAR Robust Reading Competition dataset because of inavailability of the pixel level ground truth. However, we are working towards preparing an accurate pixel level ground truth of the ICDAR dataset for future comparative evaluations.

B. Evaluation Protocol

We first calculate the saliency map S by all of these methods and their different parameter combinations for each of the image I in the dataset. We then apply a size threshold t_n of top $n\%$ pixels where $n \in [0-100]$ in step of 5%. The threshold t_n is estimated by making a histogram of gray level (256 bins) for the saliency map and finding the gray value which contains atleast top $n\%$ pixels of the image size. Given a ground–truth image I_{GT} with number of text pixels, G_T and number of non-text (background) pixels G_B, we apply a range of threshold, t_n on saliency map and calculate for each value of the threshold,

1) total number of salient pixels, $\mid S_S \mid$
2) total number of non–salient pixels, $\mid S_{NS} \mid$
3) number of salient pixels that overlap with the ground–truth text region, $\mid S_T \mid$
4) number of salient pixels that overlap with the ground–truth non-text(background) region, $\mid S_B \mid$

We define, for each of the threshold value, the following performance metrics.

[2]http://mplab.ucsd.edu/~nick/NMPT/bib_page.html

$$FAR = \frac{\mid S_B \mid}{\mid G_B \mid} \quad , \quad FRR = \frac{\mid G_T \mid - \mid S_T \mid}{\mid G_T \mid}$$

We show the performance of an algorithm by receiver operator characteristic (ROC) curves. False acceptance rate (FAR) and false rejection rate (FRR) are plotted on x and y axis respectively for the range of threshold values as shown in Figure 2. The dashed line crossing the origin in the plot shows equal error rate. The curve closest to the origin represent the best performing algorithm as it has the lowest equal error rate.

IV. Results and Discussion

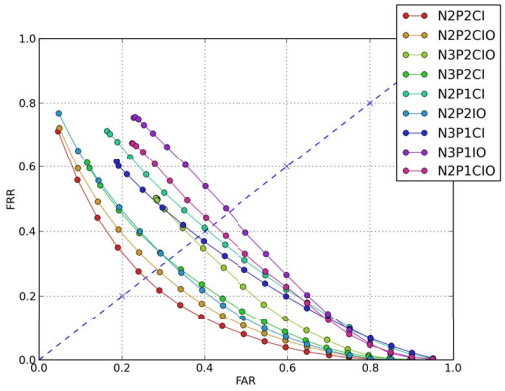
We used Matlab implementation of Itti's method [3]. Itti's method has a number of free parameters as described in section II-A each for, 1) Information Channel, Color (C), Intensity (I) and Orientation (O), 2) Normalization method, Iterative (N1), no-normalization (N2) and local-max (N3), 3) Pyramid type for sub–sampling, dyadic Gaussian (P1), and sqrt2 (P2).

We experimentally evaluated all different parameter settings for Itti's method in order to find the best combination for scene text detection. The iterative normalization scheme (N2) produces very selective saliency maps as it tries to predict the most salient location in an image and are not suitable for text detection. Saliency maps for the best performing parameter combination of Itti's method are shown in Figure 1(g). ROC curves for a few of the parameter settings are shown in Figure 2(a) that shows the range of performance we can achieve with Itti's method. Each ROC curve is plotted for different parameter combinations. The parameter combination, N2P2CI, which corresponds to using sqrt2 pyramid (P2) for spatial scales and using the Color (C) and Intensity (I) conspicuity maps without any normalization (N2), peforms best for text detection with equal error rate of 0.25. This is reasonable, since color and intensity produces the values in the same range and thus normalization will have very little effect. Similarly, the worst performing parameter combination is N3P1IO, which uses the local–max normalization (N3) and only intensity and orientation maps.

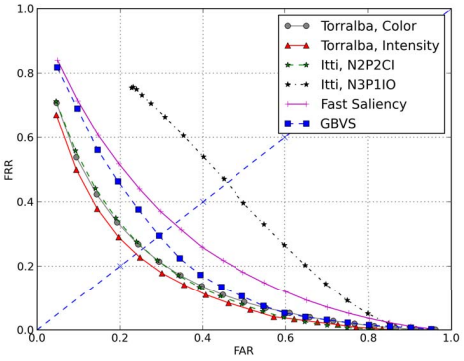
We also used two variations of Torralba's method as described in section II-C. The resulting saliency map for Color and Intensity are shown for some of the images in Figure 1(c),(d). Similarly the saliency maps for Zhang's fast saliency model and Harel's graph based visual saliency method are shown in Figure 1(f) and Figure 1(e).

The comparison of these methods is shown in the plot of Figure 2(b). Torralba's saliency map using the intensity channel clearly performs the best with equal error rate of 0.23. The performance of Itti's best parameter combination (N2P2CI) is comparable to that of Torralba's saliency

[3]http://www.saliencytoolbox.net/



(a) ROC curve for different parameter settings of Itti's method



(b) ROC curve comparing different methods

Figure 2: Evaluation results of visual attention models

maps obtained by using the color information. Zhang's fast saliency model obtained using the image intensity is less suitable for text detection. It is able to capture text information; however it is also very sensitive to slight variations in intensity resulting in generation of many false positives (salient background regions) as can be seen from saliency maps in Figure 1. The worst performing Itti's parameter combination (N3P1IO) is only shown here for reference.

The evaluation results clearly show that Torralba's saliency model can be effectively used in the initial stages of text detection. It is to be noted that we only modelled saliency estimation based on local features for each of the methods in order to be fair. However, Torralba's model can be improved by using the visual appearance based shape–prior and the location–prior.

V. Conclusion

In this paper we have evaluated four state-of-the-art models of visual attention for the task of scene text detection. The goal of our evalution is to see which models of visual attention are best suited for the task of text detection in natural scenes. The experimenal results showed that Torralba's model performed best for separation of text elements from non-text elements (background). We also identified the parameter combination for Itti's method which can be used for the task of text detection. The results clearly show that attention–based models can be used in early stages of scene text detection.

References

[1] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[2] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press, 2007, pp. 545–552.

[3] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Int. Conf. on Computer Vision*, Kyoto, Japan, 2009, pp. 2106–2113.

[4] A. Torralba, M. S. Castelhano, A. Oliva, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search," *Psychological Review*, vol. 113, no. 4, pp. 766–786, 2006.

[5] L. Elazary and L. Itti, "A Bayesian model for efficient visual search and recognition," *Vision Research*, vol. 50, no. 14, pp. 1338–1352, 2010.

[6] Q. Sun, Y. Lu, and S. Sun, "A visual attention based approach to text extraction," in *Int. Conf. on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 3991–3995.

[7] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 11, pp. 300–312, February 2007.

[8] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition - a gentle way," in *2nd Workshop on Biologically Motivated Computer Vision*. Tuebingen: Springer, 2002, pp. 472–479.

[9] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, pp. 353–367, 2011.

[10] S. Uchida, Y. Shigeyoshi, Y. Kunishige, and F. Yaokai, "A key point-based aproach toward scenery character detection," in *Int. Conf. on Document Analysis and Recognition*, Beijing,China, 2011, pp. 819–823.

[11] A. Shahab, F. Shafait, and A. Dengel, "Bayesian Approach to Photo Time-stamp Recognition," in *Int. Conf. on Document Analysis and Recognition*, Beijing,China, 2011, pp. 819–823.

[12] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Journal of Vision*, vol. 9, no. 12, pp. 1–15, 2010.

[13] L. Itti, "Models of bottom-up and top-down visual attention," , Pasadena, California, Jan 2000.

[14] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Int. Conf. on Image Processing*, Barcelona, Catalonia, 2003, pp. 253–256.

[15] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 1–20, 2008.

[16] N. J. Butko, L. Zhang, G. W. Cottrell, and J. R. Movellan, "Visual salinecy model for robot cameras," in *Int. Conf. on Robotics and Automation*, Pasadena, California, 2008, pp. 2398–2403.