CONTEXTUAL INFORMATION BASED VISUAL SALIENCY MODEL

*Seungchul Ryu, Bumsub Ham and *Kwanghoon Sohn*

Digital Image Media Laboratory (DIML),
School of Electrical and Electronic Engineering, Yonsei University, Seoul, Republic of Korea
E-mail: {ryus01, *khsohn}@yonsei.ac.kr

ABSTRACT

Automatic detection of visual saliency has been considered a very important task because of a wide range of applications such as object detection, image quality assessment, image segmentation, and more. Thanks to active researches in this field, many effective saliency models have been developed. Nevertheless, several challenging problems are still remain unsolved, such as detecting saliency in complex scene and providing high resolution and accurate saliency maps. In order to address such challenging problems, we propose a visual saliency model based on the concept of contextual information. First, we introduce a general framework for detecting saliency of an image using contextual information. Then, the proposed saliency model based on color and shape features is proposed. Quantitative and qualitative comparisons with seven state-of-the-art models on the public database show that the proposed model achieves excellent performance. Especially, the proposed model can provide good performance on challenging images including images with cluttered background and repeating distractors compared to the other models.

Index Terms— Contextual information, Visual saliency model, Color feature, Shape feature.

1. INTRODUCTION

The automatic detection of salient regions in an image is important task for various applications, such as region-of-interest based image compression [1], image segmentation [2], image and video quality assessment [3], object recognition [4], and content-aware image resizing [5]. Over the last decades, considerable efforts have been devoted to develop computational saliency models. Early approaches for detecting salient regions in an image are based on measuring the complexity of a local region under the assumption that a region should be salient if it contains features with many different orientations and intensities [6]. A common critique of the models is that they always assign high saliency scores to highly textured areas regardless of their context.

Center-surround mechanism [7] addresses the problem faced by complexity-based models. As one of the earliest methods, Itti *et al.* [7] proposed a center-surround mechanism using local feature contrast in the color, intensity, and orientation of an image. Thereafter, many center-surround-based saliency models have been proposed. Meur *et al.* [8] proposed the bottom-up visual saliency model using a coherent computational approach. Psychovisual space is used to combine the visual features, and the saliency is computed using a contrast sensitivity function. Gao and Vasconcelos [9] use difference of Gaussians (DoG) filters and Gabor filters, measuring the saliency of a point as the Kullback-Leibler (KL) divergence between the histogram of filter responses at the point and in the surrounding region. Bruce and Tsotsos [10] define bottom-up saliency based on maximum information sampling. In [11], natural statistics based saliency model was proposed in Bayesian framework. In [12], local steering kernels are employed to compute the saliency by contrasting the gradient covariance of a surrounding region. The model presented in [13] uses color and edge orientation to compute saliency map, using a framework similar to the one presented in [12]. A multi-scale center-surround saliency map was proposed in [14].

However, center-surround approach (local approach) may not produce satisfactory results due to the lack of global measurement. Also, these methods usually highlight the object boundary instead of the entire object. This makes such a method be suboptimal for computer vision tasks. To address problems of the local approach, several global information based models were proposed [15, 16, 17, 18, 19]. In [15, 16, 17], Fourier domain saliency models were proposed. These methods are based on finding regions in an image which imply unique frequencies in the Fourier domain. These methods are very fast, but since they are based on global considerations in Fourier domain, they do not detect object boundaries accurately and give high resolution saliency. Graph-representation based visual saliency models were proposed in [18, 19]. A model presented in [18] computes salient regions by determining the frequency of visits to each node in graph-representation. In this model, edge strength is used to represent the dissimilarity between two nodes. In [19], salient regions are computed by modeling the global and local properties of salient regions in graphs. These models can compute the globally optimized salient regions in the graphs.

These models including center-surround-based, Fourier domain, and graph-based methods, suffer from the limitation that cluttered backgrounds may yield higher saliency because such backgrounds possess high global exception in the cases with complex scenes. In order to overcome this problem, global-distinctness based models were proposed [20, 21]. The model presented in [20] computes saliency as the global distinctness of small image segments in feature space. This model needs image segmentation process and the performance heavily relies on the quality of image segmentation, which itself is a challenging problem. In [21], global contrast based saliency model was proposed to produce a full resolution saliency map. Although this model provides full-resolution of saliency map, it provides inaccurate saliency in some cases since only color-rarity is considered in the saliency model.

In this paper, contextual information based saliency model is proposed based on the fact that human visual system attends the most informative points in an image to reduce the uncertainty about what we are looking at. The proposed model provides high-resolution of accurate saliency map through all images from easy to challenging images. The remainder of this paper is organized as follows. In

(*): corresponding author

978-1-4799-2341-0/13/$31.00 ©2013 IEEE

section 2, first, a general framework for the contextual information based saliency is introduced. Then, the proposed saliency model is proposed, in which color and shape features are employed to compute the contextual information. In Section 3, experimental results and applications of the proposed model are presented. Lastly, Section 4 concludes this paper with some future works.

2. THE PROPOSED SALIENCY DETECTION MODEL

2.1. Problem formulation

Let the binary random variable r_i denote whether a pixel position $\mathbf{x}_i = [x_i, y_i]^T$ in an input image $\mathbf{I}$ is salient or not as follows:

$$r_i = \left\{ \begin{array}{cc} 1, & \mathbf{x}_i \text{ is salient} \\ 0, & \text{otherwise}, \end{array} \right. \tag{1}$$

where $i = \{1\,, \,...\,, \,N\}$, and N is the number of pixels in the image. Then, the detection of saliency S_i is formulated as computing a posterior probability $\Pr(r_i = 1\,|\mathbf{F}_i\,)$ at pixel position $\mathbf{x}_i$ as follows:

$$S_i = \Pr(r_i = 1|\mathbf{F}_{\,i}), \tag{2}$$

where the feature matrix, $\mathbf{F}_i = [\mathbf{f}^1{}_i, ..., \mathbf{f}^K{}_i]$ at pixel position $\mathbf{x}_i$ contains a set of feature vectors $(\mathbf{f}^k)$. K is the number of employed feature vectors.

2.2. Contextual information based saliency detection framework

As described above, the saliency detection task is to find an assignment function Φ such that

$$S_i = \Phi(\mathbf{F}_i), \tag{3}$$

where S_i is the saliency computed from feature matrix $\mathbf{F}_i$, in which higher value indicates higher salient location.

Human visual system during natural and active vision center the salient point on the most informative points in an image in order to reduce the uncertainty about what we are looking at [22]. In this paper, based on this phenomenon, we measure saliency at a pixel in terms of how much it contains information. Note that the most informative point highly depends on the observer's knowledge-background, not only contents of the image. We only consider information derived from contents of the image, i.e., *contextual information*, since the observers knowledge-background is impractical to be considered into a general saliency detection model.

The information of a pixel is determined by uncertainty of features. In other words, The most informative points is the pixels containing the rarest features in an image. We let $c^k{}_i$ denotes the contextual information for d $\times$ 1 feature vector $\mathbf{f}^k{}_i$ at pixel position $\mathbf{x}_i$, and $c^k{}_i$ is defined as follows:

$$c^k{}_i = \frac{1}{e^{\lambda^k \cdot p^k(\mathbf{f}^k{}_i)}}, \tag{4}$$

where $p^k(\mathbf{u}) = \Pr(\mathbf{f}^k = \mathbf{u})$ is probability distribution for k^{th} feature $\mathbf{f}^k$, and λ is controlling parameter. Then, contextual information $c^k{}_i$ are pooled up to compute the saliency S_i using Minkowski-mean with weighting parameters $\omega^k \in (0\,,\,1)$, $\sum\limits_k \omega^k = 1$ as follows:

$$S_i = \left(\sum_k \omega^k \cdot \left(c^k{}_i \right)^\rho \right)^{1/\rho}, \tag{5}$$

where ρ is Minkowski parameter.

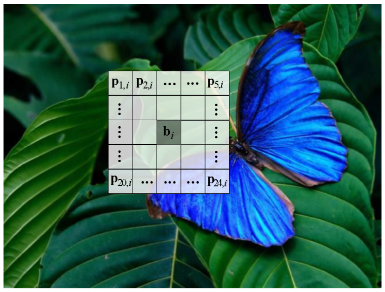


Fig. 1. Descriptions of local binary patch-pattern.

2.3. Implementation of The Proposed Saliency Model

In this paper, we employed two features for computing contextual information: color feature $\mathbf{f}^1$ and shape feature $\mathbf{f}^2$ which are defined in CIE L*a*b* color space $\mathbf{I} = (\mathbf{I_L}, \mathbf{I_a}, \mathbf{I_b})$. For each pixel position $\mathbf{x}_i$, the color feature $\mathbf{f}^1{}_i$ is defined as pixel-wise quantized chroma and computed as follows:

$$\mathbf{f}^1{}_i = \{f^1{}_{i,1}\,,\, f^1{}_{i,2}\} = \left\{ \left[\frac{\mathbf{I}_a(\mathbf{x}_i)}{\Theta} \right], \left[\frac{\mathbf{I}_b(\mathbf{x}_i)}{\Theta} \right] \right\}, \tag{6}$$

where $[\cdot]$ and Θ represent Gauss operator and quantization step, respectively. Quantization is employed for better discriminative power of color information.

For each pixel position $\mathbf{x}_i$, the shape feature $\mathbf{f}^2$ is defined as local binary patch-pattern inspired by local binary pattern [23] and computed as follows:

$$\begin{array}{l} \mathbf{f}^2 = \{f^2{}_{1,i}\,, \,...,\, f^2{}_{24,i}\} \\ \quad = \left\{ \frac{1}{M} \cdot \left|\mathbf{b}_i - \mathbf{p}_{1,i}\right| > \delta, ..., \frac{1}{M} \cdot |\mathbf{b}_i - \mathbf{p}_{24,i}| > \delta \right\} \end{array}, \tag{7}$$

where $\mathbf{b}_i$ and $\mathbf{p}_{k,i}$ are current block and k^{th} patch at pixel position $\mathbf{x}_i$, respectively, which are illustrated in Fig. 1. $|\cdot|$ represents L_2 norm. M and δ represent the size of a patch and threshold, respectively.

Then, from f^1 and f^2, contextual information for color feature c^1 and shape feature c^2 are computed using (4). Based on these two contextual information c^1 and c^2, the saliency S_i at pixel position $\mathbf{x}_i$ is computed as follows similarly to (5):

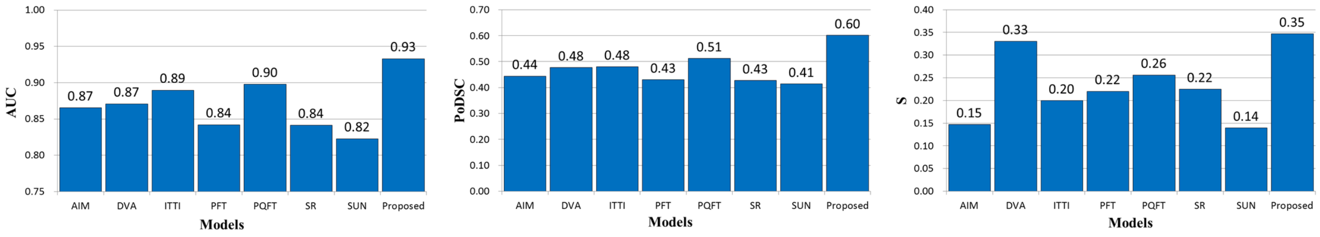
$$S_i = \left(\omega^1 \cdot (c^1{}_i)^\rho + \omega^2 \cdot (c^2{}_i)^\rho \right)^{1/\rho}. \tag{8}$$

The parameters empirically used in this paper are as follows: $\Theta = 2$, $M = 15$, $\delta = 80$, $\omega^1 = \omega^2 = 0.5$, $\lambda^1 = \lambda^2 = 5000$, and $\rho = 3$. A Matlab implementation of the proposed saliency model is available at http://diml.yonsei.ac.kr/∼sryu/cis/.

3. EXPERIMENTAL RESULTS

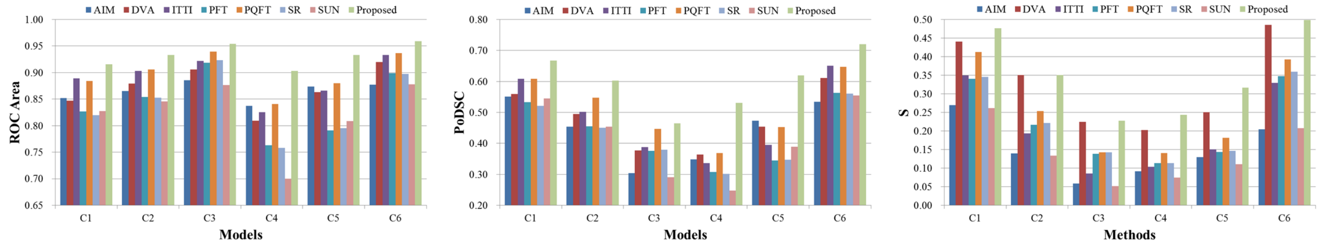
3.1. Database and evaluation measures

We evaluate the proposed saliency model on the well-constructed database provided in [24]. The database contains 235 images divided into 6 categories: C1) 50 images with large salient regions, C2) 80 images intermediate salient regions, C3) 60 images with small salient regions, C4) 15 images with cluttered background, C5) 15 images with repeating distractors, and C6) 15 images with both large and small salient regions. This database provides both region ground



(a) (b) (c)

Fig. 2. Quantitative comparisons for all images in terms of (a) AUC, (b) PoDSC, and (c) S



(a) (b) (c)

Fig. 3. Quantitative comparisons for 6 categories, C1: large salient region, C2: intermediate salient region, C3: small salient region, C4: cluttered background, C5: repeating distractors, and C6: both large and small salient regions, in terms of (a) AUC, (b) PoDSC, and (c) S.

truth (human labeled by 19 nave subjects) and fixation ground truth (by eye tracker Tobii T60, 21 subjects).

Measuring the area under the receiver operating characteristics curve (AUC) [24] is the most widely used method to evaluate the performance of saliency models. However, according to [25] the inherent limitation of AUC is that it only depends on the ordering of the values, always providing high AUC as long as the hit rates are high regardless of the false alarm rate. While the AUC analysis is still useful, it is insufficient to evaluate the performance of saliency models.

For more comprehensive evaluation, we also use two similarity measures presented in [26] and [27]. The similarity measure presented in [26], referred here as S, computes how similar two distributions, while the similarity measure in [27], referred as the peak value of dice similarity coefficient (PoDSC), measure the overlap between the thresholded saliency map and the ground truth. They penalize false positive and thus can complement the limitation of AUC.

3.2. Evaluation results

To evaluate the performance of the proposed saliency model, AUC, S, and PoDSC are measured on the database [24] with region ground truth. Seven sate-of-the-art models (actively cited models) were employed to conduct the comparisons: AIM [10], DVA [28], ITTI [7], PFT [17], PQFT [17], SR [16], and SUN [11].

The quantitative results including AUC, S, and PoDSC for each model are presented in Fig. 2. The results indicate that the proposed model shows better performance than the other models in terms of all measures. For more detailed analysis on different categories, we also analyze the performance for each category as shown in Fig. 3. The results show that the proposed model presents the highest performance through all the categories. Especially, for challenging test sets (C4 and C5), the proposed model shows good performance, while the other models fail to predict saliency.

In Fig. 4, examples of qualitative comparison for each category are presented. Due to space limitations, we are unable to show all of the qualitative results. The results are available at http://diml.yonsei.ac.kr/∼sryu/cis/. shown in Fig. 4, it is clear that the proposed model achieves the best performance with providing full-resolution and accurate saliency maps. For challenging images with cluttered backgrounds and repeating distractors (fourth and fifth rows in Fig. 4), the proposed model achieves excellent performance compared to the other models that are very sensitive to background and repeating objects, which is also supported by the quantitative results.

4. CONCLUSION AND FUTURE WORKS

In this paper, the visual saliency model is proposed based on the contextual information. The experimental results show that the proposed model outperforms seven state-of-the-art models, especially for the challenging images with cluttered backgrounds and repeating distractors. Furthermore, the proposed saliency model provides high resolution of accurate saliency maps, while the other models present inaccurate object boundary and low resolution saliency maps. Nevertheless, a saliency map computed by the proposed model still contains some outliers at complex backgrounds. Thus, our future works include the outlier rejection scheme for high quality saliency map. The other direction of future works include the extension of the proposed saliency model into video domain and three-dimensional domain.

5. REFERENCES

[1] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[2] J. Han, K. N. Ngan, M. Li, and H. Zhang, "Unsupervised

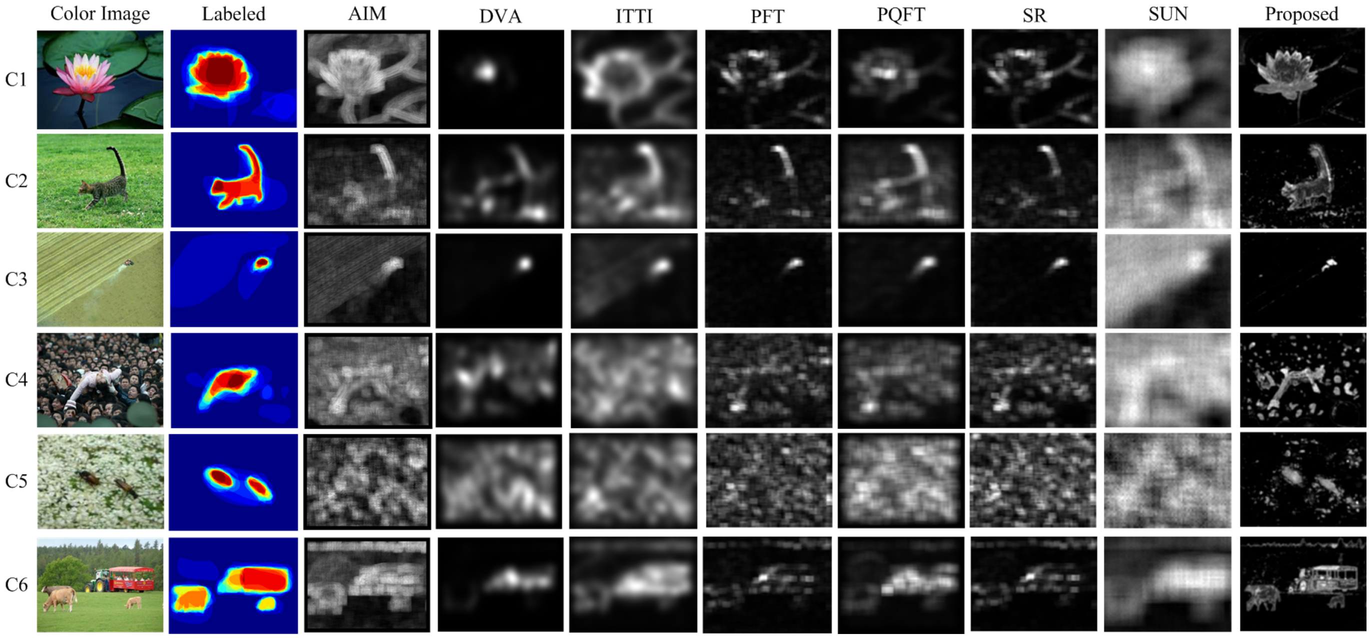


Fig. 4. Examples of qualitative comparison for each category

extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.

[3] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Top. Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.

[4] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Comput. Vis. Image Underst.*, vol. 100, no. 1-2, pp. 41–63, Oct.-Nov. 2005.

[5] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 1–9, Jul. 2007.

[6] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Comput. Vis.*, vol. 45, no. 2, pp. 83–105, Nov. 2001.

[7] 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. 30, no. 11, pp. 1254–1259, Nov. 1998.

[8] O. Le Meur, P. Callet, and D. Barba, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[9] D. Gao, V. Mahadevan, and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *Proc. IEEE Int. Conf. Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007.

[10] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vision*, vol. 9, no. 3, pp. 1–24, Mar. 2009.

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

[12] H. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," in *Proc. IEEE Int. Conf. Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007.

[13] W. Kim, C. Jung, and C. Kim, "Saliency detection: A self-ordinal resemblance approach," in *Proc. IEEE Int. Conf. Pattern Recogn.*, Suntec City, Singapore, Jul. 2010.

[14] R. Huang, N. Sang, L. Liu, and Q. Tang, "Saliency based on multi-scale ratio of dissimilarity," in *Proc. IEEE Int. Conf. Pattern Recogn.*, Istanbul, Turkey, Aug. 2010.

[15] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vision and Pattern Recogn.*, Miami, FL, USA, Jun. 2009.

[16] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vision and Pattern Recogn.*, Minneapolis, MN, USA, Jun. 2007.

[17] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Conf. Comput. Vision and Pattern Recogn.*, Anchorage, AK, USA, Jun. 2008.

[18] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Advances in Neural Information Processing Systems*, Lake Tahoe, NV, USA, Dec. 2006.

[19] V. Gopalakrishnan, H. Yiqun, and D. Rajan, "Random walks on graphs for salient object detection in images," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3232–3242, Dec. 2010.

[20] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 693–708, Apr. 2010.

[21] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE*

Conf. Comput. Vision and Pattern Recogn., Providence, RI, USA, Jun. 2011.

[22] T. S. Lee and S. X. Yu, "An information-theoretic framework for understanding saccadic eye movements," in *Proc. Advances in Neural Information Processing Systems*, Denver, CO, USA, Nov. 2000.

[23] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texutre classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[24] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, In press.

[25] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vision*, vol. 11, no. 3, pp. 1–15, Mar. 2011.

[26] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," *MIT Computer Science and Artificial Intelligence Laboratory Technical Report*, Jan. 2012.

[27] T. Veit, J. Tarel, P. Nicolle, and P. Charbonnier, "Evaluation of road marking feature extraction," in *Proc. IEEE Conf. Intell. Trasport. Syst.*, Beijing, China, Oct. 2008.

[28] X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments," in *Proc. Advacnes in Neural Information Processing Systems*, Vancouver, Canada, Dec. 2009.