Visual Saliency Detection With Free Energy Theory

Ke Gu, *Student Member, IEEE,* Guangtao Zhai, *Member, IEEE,* Weisi Lin, *Senior Member, IEEE,* Xiaokang Yang, *Senior Member, IEEE,* and Wenjun Zhang, *Fellow, IEEE*

***Abstract*—Visual saliency can be thought of as the product of human brain activity. Most existing models were built upon local features or global features or both. Lately, a so-called free energy principle unifies several brain theories within one framework, and tells where easily surprise human viewers in a visual stimulus through a psychological measure. We believe that this "surprise" should be highly related to visual saliency, and thereby introduce a novel computational Free Energy inspired Saliency detection technique (FES). Our method computes the local entropy of the gap between an input image signal and its predicted counterpart that is reconstructed from the input one with a semi-parametric model. Experimental results prove that our algorithm predicts human fixation points accurately and is superior to classical/state-of-the-art competitors. Our source code will be released at http://www.ntu.edu.sg/home/wslin/Publications.htm and https://sites.google.com/site/guke198701/home.**

***Index Terms*—Saliency detection, free energy, semi-parametric model, linear autoregressive (AR) model, bi-later filtering**

I. Introduction

SALIENCY detection is an active and important research topic in both image processing and computer vision communities. In many applications of graphics, design and human computer interaction, we strongly concern about where human beings look in a scene − where saliency spots are located. Visual saliency can promote the study of quality assessment [1]-[2], object recognition [3]-[4], and computer graphics [5]. So an efficient and effective computational model is eagerly required to detect salient areas in the encountered scene.

More than hundreds of saliency detection models have been proposed during the past 25 years [6], and this number is expected to be increasing quickly. Existing methods are divided into two types according to distinct attentional mechanisms: 1) top-down task-dependent methods; 2) bottom-up stimulus-driven methods. Because top-down approaches require prior knowledge about the visual content, bottom-up approaches that only use information from the visual signal itself have been broadly and deeply researched.

We in this paper concentrate on bottom-up methods. Many techniques in this class were modeled to seek for locations with maximum local saliency and employ biologically motivated local features [7]-[10]. These features, which mainly consist of intensity, edge, texture, color and orientation, are inspired by neural responses in lateral geniculate nucleus and V1 cortex. The benchmark Itti model [7] provides a general architecture for detecting visual saliency. This model works by first subsampling an input image into a Gaussian pyramid, decomposing each pyramid level into various channels for color, intensity and orientation, and then summing and normalizing maps in each channel across scales to yield the final saliency map.

Some other relevant algorithms depend on global features [11]-[15]. The techniques mainly attempt to find regions from a visual signal that implies unique frequencies in transform domains. This renders these algorithms quickly and precisely detect visual "pop-outs" due to global considerations, thus to locate possible salient objects. The classical spectral residual (SR) model [11] was established upon the finding that more high-frequency information than low-frequency one is stored in the residual, and the remaining Fourier amplitude spectrum is used to constitute a saliency map.

Recently, the adoption of only local or global features was found to be somewhat limited. Thus, an increasing number of nowadays studies have been devoted to incorporating both two types of features for saliency detection [16]-[20]. Most of them were developed based on complementary strategies, thereby gaining substantially high performance. In [18], the authors took into account local and global image patch rarities (LG) as two complementary processes to design the saliency detection model. In [19], content-aware saliency detection (CAS) model combines four basic principles of human visual attention, i.e. local low-level considerations, global considerations, visual organization rules, and high-level factors.

It is human viewers deciding visual saliency, and thus the most valid technique should highly approximate the response of the human brain to visual stimuli. Friston has lately unified some brain theories within the free-energy framework, which indicates that the brain inference process always attempts to infer the meaningful part from a visual stimulus by removing the uncertainty [21]. It is natural that there exists a gap between the real scene and the brain's prediction due to the fact that the internal generative model cannot be universal. It is the gap that makes human viewers "surprise", and thus attracts much more human attention. Therefore, we hypothesize that this gap (i.e. "surprise") highly correlates with the visual saliency. Based on this postulation, this paper designs a new computational Free Energy inspired Saliency detection model (FES). Our work computes the local entropy of the gap between an image and its predicted version reconstructed from the input one by a semi-parametric model, which fuses the parametric autoregressive (AR) operator that can simulate a broad range of natural scenes and the non-parametric bi-lateral filtering that works stably at image edges.

This work was supported in part by the National Science Foundation of China under Grant 61025005, Grant 61371146, Grant 61221001, and Grant 61390514, the Foundation for the Author of National Excellent Doctoral Dissertation of PR China under Grant 201339, and the Shanghai Municipal Commission of Economy and Informatization under Grant 140310.

K. Gu, G. Zhai, X. Yang and W. Zhang are with Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai, 200240, China (email: guke.doctor@gmail.com).

W. Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (email: wslin@ntu.edu.sg).

The rest of this paper is organized as follows. Section II introduces the proposed FES model for saliency detection. In Section III, we compare the performance of our method with classical/state-of-the-art techniques on three popular datasets (Toronto [9], FIFA [8] and MIT [16]). Finally, the whole paper is concluded in Section IV.

II. Saliency Detection Model

The recent free energy principle explains and unifies several significant brain theories in biological and physical sciences [21], which makes a fundamental premise that the cognitive process is determined by an internal generative model in the human brain. According to this model, the brain can predict the encountered scene in a constructive way, which is essentially a probabilistic model that can be separated into a likelihood term and a prior term. Visual perception is then to invert the likelihood term, in order to infer the posterior possibilities of the encountered scene. Typically, there always exists a gap between the given scene and the brain's prediction, because this generative model cannot work effectively everywhere. It is reasonable that the gap between the external input and its generative-model-explainable part is closely related to the "surprise" perceived by the brain, and thereby can be used for visual saliency detection.

It is clear that free energy measures the discrepancy (i.e. the error map) between the input visual signal and its output best explanation which is inferred by the internal generative model. In the error map, larger-value regions are what cannot be well explained by the generative model (i.e. "surprise"), whereas smaller-value pixels are what can be easily described. This error map is obtained through minimizing free energy. Referring to the analysis in [22], the process of free-energy minimization is highly connected to the predictive coding, and it can be finally approximated as the entropy of the residuals between the input image and its predicted one.

Though the AR model is simple and can simulate a wide range of natural scenes [23]-[25], it is sometimes unstable at image edges. Hence, the internal generative model was chosen to be a newly defined semi-parametric model, which combines the parametric AR operator and the non-parametric bi-lateral filtering with a good edge-preserving ability. To specify, the AR operator is expressed by

$$y_i = \Delta^h(y_i)\boldsymbol{a} + e_i \tag{1}$$

where y_i is the value of a pixel at location x_i, $\Delta^h(y_i)$ defines h member neighborhood vector of y_i, $\boldsymbol{a} = (a_1, a_2, ..., a_h)^T$ is a vector of AR parameters, and e_i is a difference term between truth values and predictions. To determine $\boldsymbol{a}$, the linear system can be written in matrix form as

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \|\mathbf{y} - \mathbf{Y}\boldsymbol{a}\|_2 \tag{2}$$

with $\mathbf{y} = (y_1, y_2, ..., y_h)^T$ and $\mathbf{Y}(i,:) = \Delta^h(y_i)$. We can easily solve this linear system using the least square method as $\boldsymbol{a} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{y}$.

The bi-lateral filtering is a classical non-linear filtering and is easy to construct and compute [26]. We define the bi-lateral filtering to be

$$y_i = \Delta^h(y_i)\boldsymbol{b} + e_i' \tag{3}$$

where $\boldsymbol{b} = (b_1, b_2, ..., b_h)^T$ is a vector of bi-lateral filtering coefficients, and e_i' is an error term. The vector $\boldsymbol{b}$ is controlled by two factors (i. the spatial Euclidean distance between x_i and x_j; ii. the photometric distance between y_i and y_j) and it is defined by

$$\begin{aligned} b_j &= B(x_i, x_j; y_i, y_j) \\ &= \exp\left\{\frac{-\|x_i - x_j\|^2}{2\sigma_x^2} + \frac{-(y_i - y_j)^2}{2\sigma_y^2}\right\} \end{aligned} \tag{4}$$

where σ_x and σ_y are fixed numbers for adjusting the relative importance of the spatial and photometric distances.

We finally estimate the error map $\tilde{\mathbf{s}}$ between the input visual signal and the output best explanation by integrating the AR model and bi-lateral filtering. The pixel value $\tilde{y}_i$ at the location x_i in $\tilde{\mathbf{s}}$ is computed by

$$\tilde{y}_i = \frac{\Delta^h(y_i)\hat{\boldsymbol{a}} + t\Delta^h(y_i)\boldsymbol{b}}{1+t} \tag{5}$$

where t is a positive constant to alter the relative importance of above two components. Then visual saliency, which is thought of as "pop-outs" in each of small patches, can be detected by measuring the local entropy of the error map.

Here we summarize details of the proposed saliency model as follows: First, an input color image is resized to a coarse 63×47 pixel representation similar to the scheme used in [13]; Second, the error map in each color channel is estimated using the semi-parametric model before computing the local entropy; Third, the saliency map is formed to be the weighted sum across three local entropy maps in different color channels, which has been filtered by Gaussian kernel and normalized as follows:

$$S = \sum_{i=\{L,A,B\}} w_i \mathbb{N}(\mathbb{G}(\tilde{\mathbf{s}}_i)) \tag{6}$$

where $\tilde{\mathbf{s}}_L$, $\tilde{\mathbf{s}}_A$ and $\tilde{\mathbf{s}}_B$ stand for the local entropy maps in L, A and B channels. w_L, w_A and w_B are fixed positive weights for

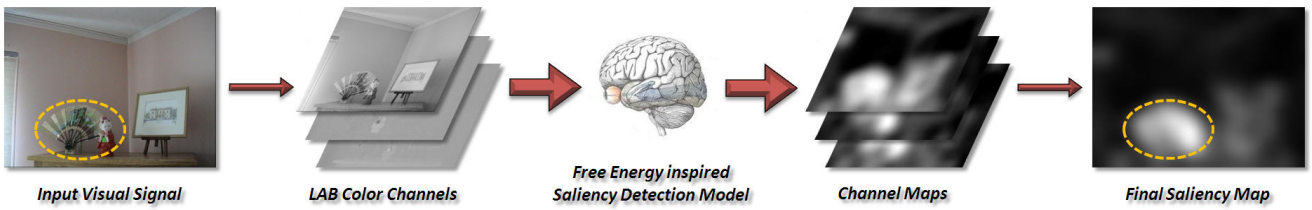

Fig. 1: An illustration of the proposed FES model. The input color image is decomposed into three LAB channels. A saliency map is computed for each channel independently, and the final saliency map is the weighted sum across three.

(a) (b) (c)

(d) (e) (f)

Fig. 2: Saliency map construction: For a visual stimulus in (a), (b)-(d) show the saliency maps computed in L, A and B channels. (e) shows the final combined saliency map. (f) shows the fixation map.

adjusting the relative importance across different colors. $\mathbb{G}(\cdot)$ denotes a Gaussian smoothing function and $\mathbb{N}(\cdot)$ is a map normalization operator. A brief flowchart of our algorithm is given in Fig. 1.

III. Experiments and Analysis

An example for illustrating the proposed model is shown in Fig. 2. For the sample image in Fig. 2(a), Figs. 2(b)-(d) present the saliency maps that are computed in L, A and B color channels. Figs. 2(e)-(f) show the final combined saliency map and the associated fixation map. As exemplified, our FES algorithm accurately predicts the human fixations.

We measure the proposed FES saliency detection model on three popular databases: Toronto dataset [9], FIFA dataset [8], and MIT dataset [16]. A total of nine methods, including three classical Itti [7], AIM [9], Judd [16], and six state-of-the-art QDCT [12], SigSal [13], HFT [14], LG [18], CAS [19], AWS [10], are used for performance comparison.

First, we qualitatively compare the proposed method with nine algorithms. We present several representative images in the Toronto dataset and associated saliency maps in Fig. 3. For each sample image in the first row, the images in the second and third rows show the corresponding fixation maps and the FES-based saliency maps. The images in the fourth to twelfth rows exhibit the saliency maps computed by the testing techniques. Through simulating the brain process in saliency detection, our approach is found to predict human fixation points precisely as compared to classical/state-of-the-art competitors.

Second, we quantitatively compare the proposed approach with existing relevant models. We compute the shuffled ROC Area Under the Curve (sAUC) score for each image to evaluate the consistency between a particular saliency map and a group of fixations. The authors of [27] have pointed out that the strong center-bias exists in human fixations and it may affect the performance indices of saliency detection techniques. To reduce such bias, we follow the procedure proposed in [27]. The positive sample set in an image includes the fixation points of all subjects on that one, whereas the negative sample

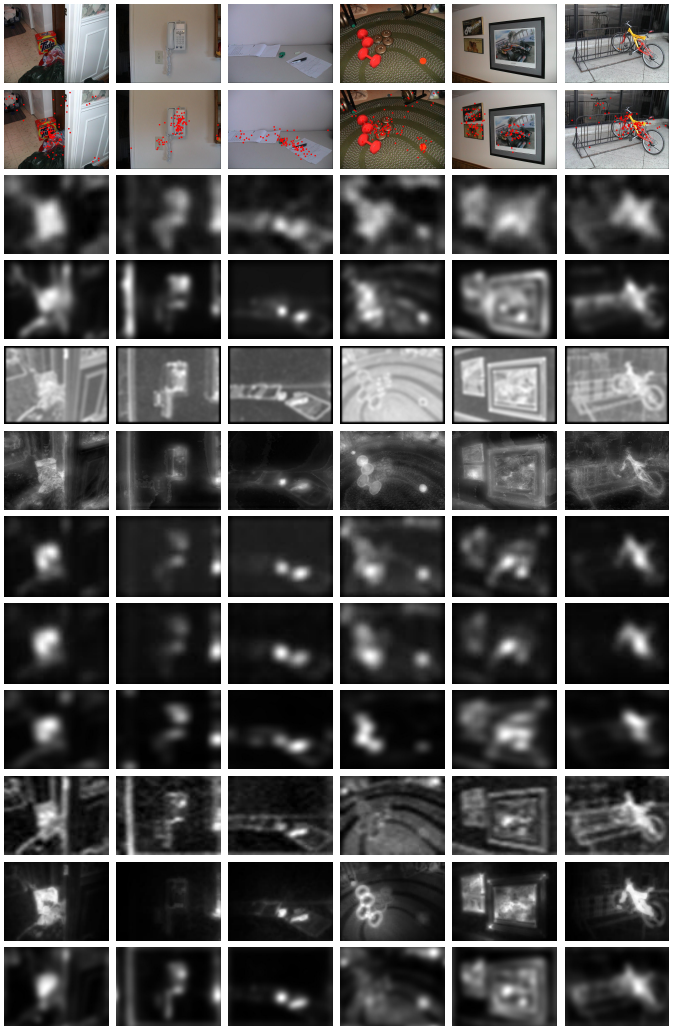

Fig. 3: Comparison of saliency detection models: The first to second rows show the representative images in the Toronto dataset [9] and the corresponding fixation maps. The images in the third to twelfth rows show the saliency maps computed by the proposed FES model, Itti [7], AIM [9], Judd [16], QDCT [12], SigSal [13], HFT [14], LG [18], CAS [19], and AWS [10].

set includes the union of all fixation points across the entire images from the same dataset−except for the positive samples. Each saliency map generated by the algorithm is thresholded and then used as a binary classifier to separate the positive samples from negative samples. At a particular threshold level Thr, the true positive rate is the proportion of the positive samples that fall into the positive (white) region of the binary saliency map. The false positive rate is computed similarly by using the negative sample set. Sweeping over thresholds yields an ROC curve, of which the area beneath provides a good index of judging how the saliency map can accurately predict where fixations occurred on an image. Note that the chance level is 0.5 while the perfect prediction is 1.0. An example is provided in Fig. 4, in which the blue curve shows the ROC curve computed using our FES technique on the first sample image in Fig. 3, while the red diagonal dash line indicates the chance level. The area under the blue curve is 0.7775, much higher than the chance level of 0.5.

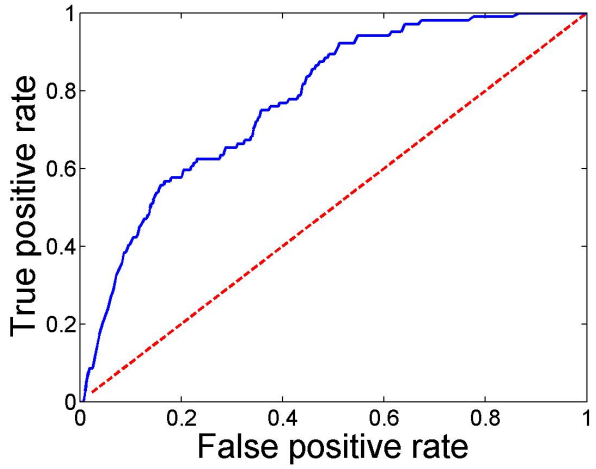

Fig. 4: The blue curve shows the ROC curve of our FES algorithm on the first sample image in Fig. 3, with the red reference dash line indicating the chance level of 0.5. The area under the blue curve is up to 0.7775.

Furthermore, we report the sAUC scores of the nine testing saliency detection models and the proposed FES technique on three eye tracking datasets as well as their direct averages in Table I. In the first classical Toronto dataset, our approach achieves the top performance across all testing algorithms. In the second FIFA dataset dedicated to the face detection, the proposed FES model is noticeably superior to other methods, which indicates our model is good at the face detection and it may help to facilitate the scientific research and practical application of face detection and recognition. In the third MIT dataset, the proposed approach still obtains the second place, a little less than the recent AWS algorithm, which is possibly due to the fact that there are quite a few images with various resolutions or corrupted by more or less motion blur. We also compute the average performance of each testing model and prove the effectiveness of our FES method.

Finally, it deserves to emphasize two points. Firstly, our model is built upon a reasonable hypothesis that the "surprise" in an encountered scene largely draws the visual attention, and our model works by using a semi-parametric method to simulate the free energy based brain theory and thus to model the aforementioned process of perceiving "surprise". This is why the proposed FES model works so effectively. Secondly, the used semi-parametric model combines the traditional AR operator with bi-lateral filtering. Instead, recently developed methods, e.g. guided image filter (GIF) [28] and its advanced weighted GIF [29] of the better edge-preserving ability, might lead to higher performance for visual saliency detection.

IV. Conclusion

In this paper we have put forward a novel computational Free Energy inspired Saliency detection technique (FES). This model is motivated by a recently revealed human brain theory, and it works by searching for the "surprise" between an input visual stimulus and its predicted version that is reconstructed from the input signal by using a semi-parametric model. The concept of surprise provides a natural ground and connection to saliency modeling, since visual saliency is about difference, contrast, pop-outs and unusual/unexpected/unpredictable happenings. By both qualitative and quantitative comparisons, our FES algorithm is shown to predict human fixation points accurately and outperform nine classical/state-of-the-art saliency detection models on sAUC scores using three benchmark eye tracking datasets. Along this line of research, several issues will be further explored in saliency detection, such as adaptive weighting for various color channels and adaptive scaling for different visual scenes.

TABLE I: Shuffled-AUC results of the proposed FES model and nine testing classical/state-of-the-art saliency detection techniques.

Models	Toronto [9]	FIFA [8]	MIT [16]	Average
Itti [7]	0.6570	0.6724	0.6424	0.6573
AIM [9]	0.6816	0.7233	0.6674	0.6908
Judd [16]	0.6836	0.7089	0.6587	0.6837
QDCT [12]	0.7162	0.7278	0.6732	0.7057
SigSal [13]	0.7047	0.7278	0.6681	0.7002
HFT [14]	0.6896	0.6991	0.6514	0.6800
LG [18]	0.6883	0.6736	0.6717	0.6779
CAS [19]	0.6919	0.7114	0.6684	0.6906
AWS [10]	0.7116	0.7084	0.6916	0.7039
FES (Pro.)	0.7195	0.7539	0.6871	0.7202

References

[1] K. Gu, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "Automatic contrast enhancement technology with saliency preservation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, 2015.

[2] K. Gu, G. Zhai, W. Lin, and M. Liu, "The analysis of image contrast: From quality assessment to automatic enhancement," *IEEE Trans. Cybernetics*, vol. 45, 2015.

[3] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325-3337, Jun. 2015.

[4] J. Han, C. Chen, L. Shao, X. Hu, J. Han, and T. Liu, "Learning computational models of video memorability from fMRI brain imaging," *IEEE Trans. Cybernetics*, 2015, to appear.

[5] L. Dong, W. Lin, Y. Fang, S. Wu, and H. S. Seah, "Saliency detection in computer rendered images based on object-level contrast," *J. Vis. Commun. Image Represent.*, vol. 24, no. 1, pp. 27-38, 2014.

[6] 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, 2013.

[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. 20, no. 11, pp. 1254-1259, Nov. 1998.

[8] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *in Proc. Adv. Neural Inf. Process. Syst.*, vol. 20, 2008.

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

[10] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image and Vision Computing*, vol. 30, no. 1, pp. 51-64, 2012.

[11] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," *in Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, Jun. 2007.

[12] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," *in Proc. European Conf. Computer Vision*, pp. 116-129, 2012.

[13] 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, Jan. 2012.

[14] 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.*, vol. 35, no. 4, pp. 996-1010, Apr. 2013.

[15] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," *in Proc. IEEE Int. Conf. Computer Vision*, pp. 153-160, Dec. 2013.
[16] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," *in Proc. IEEE Int. Conf. Computer Vision*, pp. 2106-2113, 2009.
[17] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, and C.-W. Lin, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 187-198, Feb. 2012.
[18] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," *in Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 478-485, Jun. 2012.
[19] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915-1926, Oct. 2012.
[20] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, 2015, to appear.
[21] K. Friston, "The free-energy principle: A unified brain theory?" *Nature Rev. Neuroscience*, vol. 11, pp. 127-138, 2010.
[22] H. Attias, "A variational bayesian framework for graphical models," *in Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 209-215, 2000.
[23] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 41-52, Jan. 2012.
[24] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Hybrid no-reference quality metric for singly and multiply distorted images," *IEEE Trans. Broadcasting*, vol. 60, no. 3, pp. 555-567, Sept. 2014.
[25] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50-63, Jan. 2015.
[26] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *in Proc. IEEE Int. Conf. Computer Vision*, pp. 836-846, Jan. 1998.
[27] B. Tatler, R. Baddeley, and I. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision Research*, vol. 45, no. 5, pp. 643-659, Mar. 2005.
[28] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397-1409, Jun. 2013.
[29] Z. Li, J. Zheng, Z. Zhu, W. Yao, and S. Wu, "Weighted guided image filtering," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 120-129, Jan. 2015.