Cost-Sensitive Rank Learning From Positive and Unlabeled Data for Visual Saliency Estimation

Jia Li, Yonghong Tian, *Member, IEEE*, Tiejun Huang, *Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*—This paper presents a cost-sensitive rank learning approach for visual saliency estimation. This approach avoids the explicit selection of positive and negative samples, which is often used by existing learning-based visual saliency estimation approaches. Instead, both the positive and unlabeled data are directly integrated into a rank learning framework in a cost-sensitive manner. Compared with existing approaches, the rank learning framework can take the influences of both the local visual attributes and the pair-wise contexts into account simultaneously. Experimental results show that our algorithm outperforms several state-of-the-art approaches remarkably in visual saliency estimation.**

***Index Terms*—Cost-sensitive, positive and unlabeled data, rank learning, visual saliency.**

I. Introduction

From the perspective of signal processing, visual saliency refers to the selection mechanism to pop-out the "important" content from the input visual stimuli. With visual saliency, the limited computational resource can be allocated to the desired targets while the distractors can be ignored. Therefore, the central issue in visual saliency estimation is to distinguish the targets from the distractors using the various visual clues.

Often, visual saliency estimation requires the integration of the bottom-up and top-down factors [1]. In existing works, the bottom-up factor is usually treated as a stimuli-driven component that determines visual saliency by detecting unique or rare visual subsets in a scene. Inspired by the Feature Integration Theory [2], many bottom-up approaches estimated visual saliency by binding the irregularities in different visual attributes. For example, Itti *et al.* [3] presented an approach to estimate image saliency by integrating intensity, color and orientation contrasts. By incorporating motion and flicker contrasts, the same approach was extended to video saliency in [4]. Harel *et al.* [5] represented each scene with a directed graph and adopted a random walker to select the salient locations corresponding to the less visited nodes. In [6], Marat *et al.* presented a biology-inspired model by simulating the filtering mechanism of the retinal cells to estimate spatiotemporal saliency. Similarly, many other approaches detected irregularities in the spatiotemporal domain (e.g., [7]–[9]), in the amplitude spectrum (e.g., [10]) or in the phase spectrum ([11]). These irregularities were then integrated in an ad-hoc manner to locate the salient target. However, such an ad-hoc integration may not always work since the top-down factor also plays a crucial role in visual saliency estimation. Often, the top-down factor can be treated as priors to guide the integration process. For example, Peters and Itti [12] proposed an approach to infer a projection matrix from global scene characteristics to saliency maps. Kienzle *et al.* [13] presented a non-parametric saliency model by using the Support Vector Machine. Navalpakkam and Itti [14] adopted a learning-based algorithm to pop-out the targets and suppress the distractors through maximizing the signal-noise-ratio. Generally speaking, these approaches can achieve promising results but still have some drawbacks. Often, the user data such as eye traces can only provide sparse positive samples. That is, only a few locations in a scene are labeled as positive, while most of other locations in the scene remain unlabeled. These unlabeled data may contain many positive samples so that it is improper to treat all of them as negative samples (e.g., as in [12] and [13]), or randomly select negative samples from them (e.g., as in [13]). Moreover, the influence of pair-wise context (e.g., the competition between targets and distractors [3], [4], the co-occurrence characteristics of various visual stimuli [15]) is not considered in these approaches, which also plays an important role in visual saliency estimation.

To solve these two problems, we propose a cost-sensitive rank learning approach on positive and unlabeled data for visual saliency estimation. In our approach, the influences of local visual attributes and pair-wise contexts are taken into account simultaneously using a pair-wise rank learning framework. Moreover, we avoid the explicit extraction of positive and negative samples by directly integrating both the positive and unlabeled data into the optimization objective in a cost-sensitive manner. Extensive experiments demonstrate that our approach outperforms several state-of-the-art bottom-up (e.g., [3]–[5], [7], [8], [10], [11]) and top-down (e.g., [12]–[14]) approaches in visual saliency estimation. Moreover, both the cost-sensitive integration of positive and unlabeled data and the rank learning framework are proved to be helpful in visual saliency estimation.

The remainder of this paper is organized as follows. Section II describes the cost-sensitive rank learning approach

Manuscript received February 09, 2010; revised April 06, 2010. Date of publication April 12, 2010; date of current version May 05, 2010. This work was supported by grants from the Chinese National Natural Science Foundation under Contracts 60973055 and 90820003, and by the National Basic Research Program of China under Contract 2009CB320906. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrizio Campisi.

J. Li is with the Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, and also with the Graduate University of CAS, Beijing, China.

Y. Tian, T. Huang and W. Gao are with the National Engineering Laboratory for Video Technology, Peking University, Beijing, China (e-mail: yhtian@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2010.2048049

for visual saliency estimation. Experimental results are shown in Section III and the paper is concluded in Section IV.

II. The Approach

Often, users intend to search the desired targets under the facilitation of experience derived from past similar scenes (i.e., the contextual cueing effect [16], [17]). In this process, a requisite step is to identify the search priority of each location. Such priority is closely related to visual saliency and can be derived from local visual attributes and pair-wise contexts. Therefore, we can formulate visual saliency estimation as a rank learning problem to estimate the searching priority of each location.

Here we first describe how to extract the local visual attributes. In general, a scene can be expressed as a conjugate of information flows from multiple visual feature channels. Among these channels, some are the probable sources of preattentive guidance and visual saliency is related to local contrasts in these preattentive channels. As in [4], we compute the local contrasts from 12 typical preattentive channels in six scales, including intensity (six), color opponencies (12), orientations (24), flickers (six) and motion energies (24). In total, 72 local contrasts are obtained. Here we use a vector $\mathbf{x}_{kn}$ with 72 components to characterize the local visual attributes of the nth location $\mathbf{B}_{kn}$ (e.g., 16×16 block) in the kth scene.

Using these features, we then present our cost-sensitive rank learning approach for visual saliency estimation. In the learning process, an important issue is to train a ranking function $\phi(\mathbf{x})$ using the ground truth saliency g_{km}. For two locations $\mathbf{B}_{km}$ and $\mathbf{B}_{kn}$, $\phi(\mathbf{x}_{km}) > \phi(\mathbf{x}_{kn})$ indicates that $\mathbf{B}_{km}$ ranks higher than $\mathbf{B}_{kn}$ and maintains a higher saliency. However, the user data often contain only sparse positive samples. As shown in Fig. 1(a), the eye traces can only reveal parts of the salient target, while most locations remain unlabeled. To utilize the unlabeled data, we derive g_{km} by considering the visual similarity and the spatial correlation between $\mathbf{B}_{km}$ and the labeled positive samples. Let e_{kn} be the event that $\mathbf{B}_{kn}$ is a labeled positive sample, the visual similarity v_{km} can be calculated as

$$v_{km} = \max_{n\in\{1,\ldots,N\}} \left(\frac{[e_{kn}]_{\mathbf{I}} \cdot \mathbf{x}_{km}^{\mathbf{T}} \mathbf{x}_{kn}}{||\mathbf{x}_{km}|| \cdot ||\mathbf{x}_{kn}||} \right) \tag{1}$$

where $[e_{kn}]_{\mathbf{I}} = 1$ if e_{kn} holds, otherwise $[e_{kn}]_{\mathbf{I}} = 0$. N is the total number of blocks in a scene. As shown in Fig. 1(b), such visual similarities can pop-out the locations that are similar to the positive samples. Moreover, the spatial correlation r_{km} is computed as

$$r_{km} = \max_{n\in\{1,\ldots,N\}} \exp\left(-\left(\frac{[e_{kn}]_{\mathbf{I}} \cdot d_{mn}}{d_k} \right) \right) \tag{2}$$

where d_{mn} is the Euclidean distance between the locations $\mathbf{B}_{km}$ and $\mathbf{B}_{kn}$, while d_k corresponds to the diagonal distance of the kth scene. As shown in Fig. 1(c), such spatial correlations can pop-out the remainder of the salient target. After that, we normalize the visual similarities and the spatial correlations into [0,1] and derive g_{km} by setting:

$$g_{km} = v_{km} \cdot r_{km}. \tag{3}$$

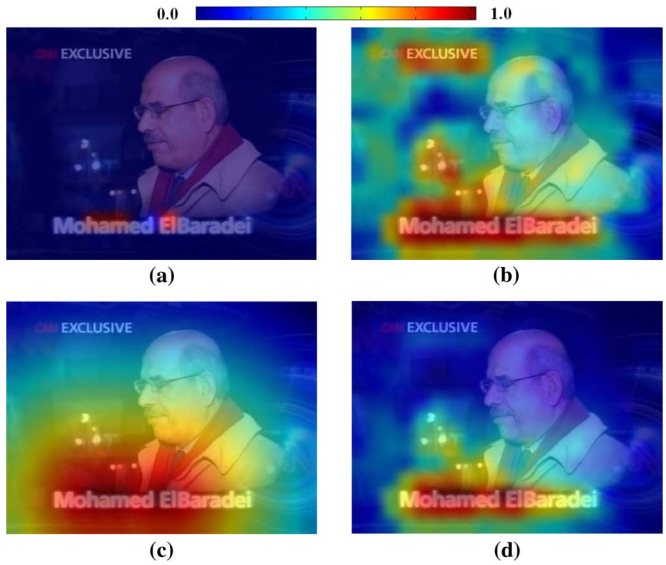

Fig. 1. Generating ground-truth saliency from sparse positive samples. (a) Sparse positive samples (the eye fixations are provided by the MTV dataset in [18]); (b) the visual similarity map; (c) the spatial correlation map; (d) the derived ground-truth saliency map.

As shown in Fig. 1(d), the formulation in (3) will only assign high saliency values to the locations that are adjacent and similar to the labeled positive samples. In the training process, however, it is often difficult to directly determine the label for each sample, especially for the one with medium saliency (e.g., around 0.5). Moreover, visual saliency estimation mainly focuses on distinguishing targets from distractors and the correlations between target pairs or between distractor pairs should be considered with low priority. Therefore, we integrate all the positive and unlabeled data (with estimated ground truth saliency) into a rank learning framework in a cost-sensitive manner. Since visual features can be bound into consciously experienced wholes for visual saliency estimation [2], we adopt a ranking function $\phi(\mathbf{x}) = \omega^{\mathbf{T}}\mathbf{x}$ to combine the input features with linear weights. Given the local visual attributes and ground-truth saliency for each location of the training scenes, the empirical loss can be defined as

$$\mathcal{L}(\omega) = \sum_{k} \sum_{m\neq n}^{N} [g_{km} - g_{kn}]_{+} \cdot [\omega^{\mathbf{T}}\mathbf{x}_{km} \leq \omega^{\mathbf{T}}\mathbf{x}_{kn}]_{\mathbf{I}} \tag{4}$$

where $[x]_{+} = \max(0, x)$. We can see that there will be a loss if the ranking function gives predictions contrary to the ground-truth saliencies (i.e., $\phi(\mathbf{x}_{km}) \leq \phi(\mathbf{x}_{kn})$ when $g_{km} > g_{kn}$). Moreover, the loss function emphasizes the correlations between targets and distractors since the central issue in visual saliency estimation is to distinguish targets from distractors. That is, the cost of erroneously ranking a target-distractor pair (i.e., $g_{km} - g_{kn} \to 1$) is much bigger than that of mistakenly predicting the ranks between target pairs or between distractor pairs (i.e., $g_{km} - g_{kn} \to 0$). Thus, it is cost-sensitive by differentiating target-distractor pairs in our framework.

Generally speaking, minimizing the loss function (4) will not only pop-out the target using the local visual attributes, but also suppress the distractors by considering the influences of pair-

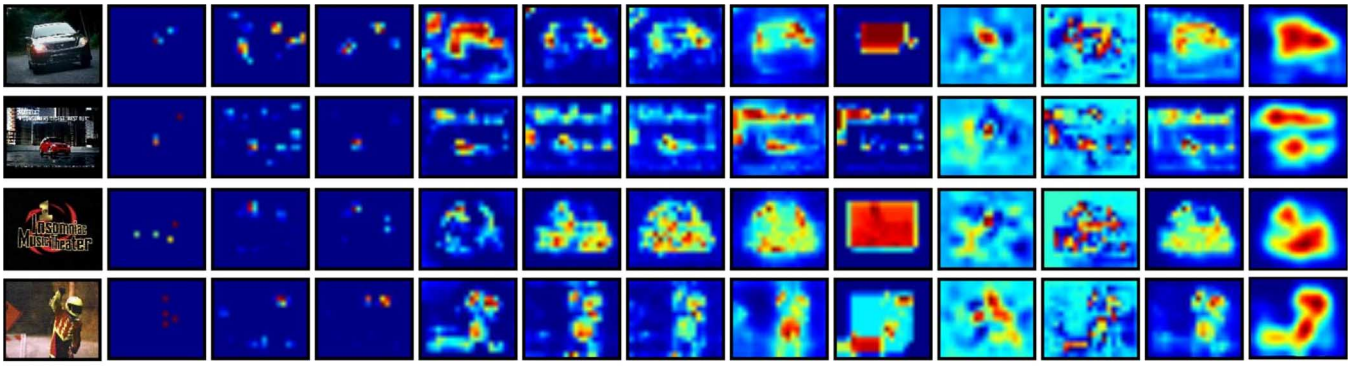

Fig. 2. Some representative results. Note that here the eye density maps are not convolved with a Gaussian kernel, which is a popular method to recover more positive samples for the evaluation. (a) Original frames; (b) eye fixation maps; (c) Itti98 [3]; (d) Itti01 [4]; (e) Itti05 [7]; (f) Hou07 [10]; (g) Guo08 [11]; (h) Harel07 [5]; (i) Zhai06 [8]; (j) Peters07 [12]; (k) Kienzle07 [13]; (l) Navalpakkam07 [14]; (m) our approach.

wise correlations. However, it is difficult to directly minimize such a binary loss function. Toward this end, a feasible solution is to find an upper bound of (4) and minimize the upper bound instead. From the definition of the loss function, we can see that:

$$[\omega^{\mathbf{T}}\mathbf{x}_{km} \leq \omega^{\mathbf{T}}\mathbf{x}_{kn}]_{\mathbf{I}} \leq e^{\omega^{\mathbf{T}}(\mathbf{x}_{kn}-\mathbf{x}_{km})}. \tag{5}$$

By incorporating (5) into (4), the optimization objective can be rewritten as

$$\min_{\omega} \sum_{k} \sum_{m \neq n}^{N} [g_{km} - g_{kn}]_{+} \cdot e^{\omega^{\mathbf{T}}(\mathbf{x}_{kn}-\mathbf{x}_{km})}. \tag{6}$$

Note that (6) contains only exponential terms with linear positive weights. Thus the objective function is convex and the global optimum can be reached using gradient-based method:

$$\triangle\omega \propto \sum_{k} \sum_{m \neq n}^{N} [g_{km} - g_{kn}]_{+} \cdot (\mathbf{x}_{kn} - \mathbf{x}_{km}) \cdot e^{\omega^{\mathbf{T}}(\mathbf{x}_{kn}-\mathbf{x}_{km})}. \tag{7}$$

With the derived ranking function, a rank $c_{km} \in \{1, \ldots, N\}$ is assigned to each block $\mathbf{B}_{km}$. To get the visual saliency map, we empirically turn this rank into a saliency value $((N - c_{km})/N)^{\beta}$, where $\beta > 0$ is a constant to pop-out the probable salient targets. For larger β (empirically set to 3 in this study), the locations other than the most salient location can be suppressed more effectively. Moreover, we convolve the saliency map with a Gaussian kernel $(\sigma = 5)$ to ensure that the entire salient object can be detected. For convenience, the saliency values are normalized into [0,1].

III. Experimental Results

In this section, we evaluate the feasibility of our cost-sensitive rank learning approach for visual saliency estimation. In the experiments, we adopt one video dataset proposed in [18] with eye traces of eight subjects. The dataset consists of over 46 000 video frames in 50 video clips (25 min), which mainly contain scenes such as "outdoors day&night," "crowds," "TV news," "sports," "commercials," and "video games." When watching each video clip, the eye traces of four to six subjects were recorded using a 240 HZ eye-tracker. In the experiment, we randomly divide the dataset into training/validation/test sets for ten times. For each division, we adopt ten state-of-the-art approaches for comparison. In general, these approaches can be grouped into two categories.

TABLE I
Performance of Various Approaches in Visual Saliency Estimation

Algorithm		AUC	Improvement (%)
BU	Itti98 [3]	0.571±0.008	35.6
	Itti01 [4]	0.564±0.007	37.2
	Itti05 [7]	0.643±0.011	20.4
	Hou07 [10]	0.680±0.009	13.8
	Guo08 [11]	0.697±0.008	11.0
	Harel07 [5]	0.590±0.009	31.2
	Zhai06 [8]	0.633±0.010	22.3
TD	Peters07 [12]	0.630±0.016	22.9
	Kienzle07 [13]	0.533±0.017	45.2
	Navalpakkam07 [14]	0.701±0.013	10.4
	Our	**0.774**±0.011	

1) **Bottom-Up (BU) approaches for saliency estimation**, including Itti98 [3], Itti01 [4], Itti05 [7], Zhai06 [8], Harel07 [5], Hou07 [10] and Guo08 [11].
2) **Top-Down (TD) approaches for saliency estimation**, including Peters07 [12], Kienzle07 [13] and Navalpakkam07 [14].

Among these approaches, Itti98 [3], Itti01 [4], Itti05 [7] and Navalpakkam07 [14] adopt the same local visual features as in our approach. For fair comparison, Kienzle07 [13] also trained the SVM classifier using these features, other than the local intensities. In the experiments, the parameters for the learning-based approaches are optimized on the validation set.

In the experiments, we use the Area Under the ROC Curve $(\mathbf{AUC})$ to evaluate the overall performance. In the evaluation, salient locations are selected from the estimated saliency maps using different thresholds. These locations are then validated using the ground-truth saliency maps (approximated by the eye-density maps as in [3], [6], [12]–[14], etc.) and the ROC curve is plotted as the *false positive rate* vs. *true positive rate*. Perfect prediction corresponds to $\mathbf{AUC}$ score of 1, while random prediction gives a $\mathbf{AUC}$ score of 0.5. The $\mathbf{AUC}$ scores of different approaches are shown in Table I and some representative results are given in Fig. 2.

From Table I, we can see that our approach outperforms all the other ten approaches. As shown in Fig. 2(c)–(d), Itti98 [3] and Itti01 [4] only maintain the most salient locations using the "winner-take-all" competition. Thus they yield low $\mathbf{AUC}$

scores since the competition may not only suppress the distractors but also inhibit the targets. Other bottom-up approaches perform much better by integrating the irregularities in the spatiotemporal domain (e.g., Itti05 [7], Zhai06 [8] and Harel07 [5]), in the amplitude spectrum (e.g., Hou07 [10]) or in the phase spectrum (e.g., Guo08 [11]). However, the experience on past scenes can often provide effective clues for suppressing the distractors, which are not considered by these approaches. Thus the saliency maps generated by these approaches often contain much noise.

Compared with the bottom-up approaches, the top-down approaches [12]–[14] do not demonstrate much improvement due to their inappropriate strategies on sampling negative training samples. For example, Peters07 [12] and Navalpakkam07 [14] treated all the locations that have received no fixations as negative samples (i.e., distractors), while Kienzle07 [13] generated the negative samples by using the same coordinates of eye fixations, but on different scenes. Therefore, many positive samples in the unlabeled data will be assigned to wrong labels, which will greatly degrade the overall performance. As shown in Fig. 2(j)–(l), these approaches can only recover parts of the targets, while the other parts are often suppressed as distractors.

Compared with these approaches, our approach can accurately locate the salient target while suppressing the distractors. Particularly, our approach can pop-out the entire salient object, other than the most salient points. As shown in the second row of Fig. 2(m), our approach can pop-out the whole caption line and the car from a complex scene. Generally speaking, the success of our approach is mainly due to two reasons. Firstly, the positive and unlabeled data is directly exploited by the optimization objective in a cost-sensitive manner, which provides effective clues for locating the entire targets. Secondly and more importantly, the rank learning framework focuses not only on the local visual attributes but also on the pair-wise correlations. In the learning process, the local visual attributes assist to pop-out the targets, while pair-wise correlations can help to suppress the distractors.

We also design an experiment to illustrate why the cost-sensitive formulation is necessary. In one experiment, we set a threshold T_D to select only the sample pairs whose saliency differences are larger than T_D into the training process. In this process, the selected sample pairs are treated with equal weights (i.e., the term $[g_{km} - g_{kn}]_+$ in (4) is fixed to 1). Firstly, we set $T_D \approx 1$ and the $\mathbf{AUC}$ score reaches 0.743. This indicates that the rank learning framework itself can improve the overall performance, even only using the most reliable samples. Then we gradually decrease T_D and the $\mathbf{AUC}$ reaches its maximal (0.775) when $T_D \approx 0.6$, and then decrease to 0.766 when $T_D = 0$. The reason is that the correlations between targets and distractors (i.e., sample pairs with high saliency differences) can provide effective clues for selecting the discriminative features for the ranking function. However, correlations between targets or between distractors (i.e., sample pairs with low saliency differences) may bias the selection process to the wrong visual attributes, thus decreasing the $\mathbf{AUC}$ score. This implies that penalizing correlations between targets and between distractors in a cost-sensitive manner is helpful in visual saliency estimation.

IV. Conclusion

In this paper, we propose a novel approach for visual saliency estimation. Our contributions are two-fold: firstly, the task of visual saliency estimation is formulated as a rank learning problem on positive and unlabeled data for the first time. The rank learning framework can consider the influences of both the local visual attributes and pair-wise contexts simultaneously. Secondly, we propose an approach to directly integrate the positive and unlabeled data into the optimization objective in a cost-sensitive manner. This helps to detect the entire targets while suppressing the distractors by focusing on the target-distractor correlations. From the experimental results, our approach outperforms several state-of-the-art bottom-up and top-down approaches. In the future work, we will extend the rank learning framework to user-targeted visual attention modeling. Moreover, we will incorporate other clues such as global scene characteristics into the rank learning framework to further improve the performance.

References

[1] L. Itti, G. Rees, and J. Tsotsos, *Neurobiology of Attention*. San Diego, CA: Elsevier, 2005.

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

[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.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[4] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.

[5] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Adv. Neural Inform. Process. Syst.*, 2007, pp. 545–552.

[6] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guerin-Dugue, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 231–243, 2009.

[7] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 631–637.

[8] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

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

[10] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[11] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[12] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[13] W. Kienzle, F. A. Wichmann, B. Scholkopf, and M. O. Franz, "A nonparametric approach to bottom-up visual saliency," in *Adv. Neural Inform. Process. Syst.*, 2007, pp. 689–696.

[14] V. Navalpakkam and L. Itti, "Search goal tunes visual features optimally," *Neuron*, vol. 53, pp. 605–617, 2007.

[15] D. Parikh, C. Zitnick, and T. Chen, "Determining patch saliency using low-level context," in *Eur. Conf. Computer Vision*, 2008.

[16] M. M. Chun, "Contextual guidance of visual attention," *Neurobiol. Attention*, pp. 246–250, 2005.

[17] A. Torralba, "Contextual influences on saliency," *Neurobiol. Attention*, pp. 586–592, 2005.

[18] L. Itti, "Crcns data sharing: Eye movements during free-viewing of natural videos," in *Collaborative Research in Computational Neuroscience Annu. Meeting*, 2008.