

Dynamic Markov random fields for stochastic modeling of visual attention

Akisato Kimura[†] Derek Pang^{†‡*} Tatsuto Takeuchi[†] Junji Yamato[†] Kunio Kashino[†]

[†] NTT Communication Science Laboratories, NTT Corporation

[‡] School of Engineering Science, Simon Fraser University

Abstract

This report proposes a new stochastic model of visual attention to predict the likelihood of where humans typically focus on a video scene. The proposed model is composed of a dynamic Bayesian network that simulates and combines a person's visual saliency response and eye movement patterns to estimate the most probable regions of attention. Dynamic Markov random field (MRF) models are newly introduced to include spatiotemporal relationships of visual saliency responses. Experimental results have revealed that the proposed model outperforms the previous deterministic model and the stochastic model without dynamic MRF in predicting human visual attention.

1 Introduction

Developing an accurate computational model of human visual attention has been a long standing challenge. Such a model may allow any system to select only relevant information from a complex visual input in numerous artificial vision applications. The first biologically plausible model for explaining the human visual attention system was proposed by Koch and Ullman [4], and later implemented by Itti et al [3]. This model analyzes still images to produce primary visual features such as intensity, color and orientation, which are combined to form a *saliency map* that represents the relevance of visual attention. Although several attempts [9, 2, 5] have been made to improve the Koch-Ullman model, they all suffer from a crucial problem in which the saliency responses are assumed to be deterministic, and therefore they only select a fixed attended location every time for the same input. However, people may attend to different locations on the same visual input at the same time.

To tackle the above problem, we have proposed the first stochastic model [7] of human visual attention

*He contributed to this work during his internship at NTT.

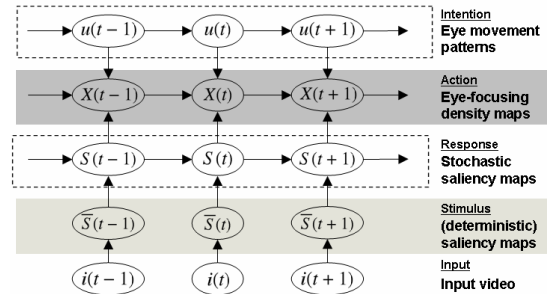


Figure 1. Graphical representation of the stochastic model of visual attention

based on the *signal detection theory* [1]. The theory suggests that elements on a visual display are internally represented as independent Gaussian random variables, and the positions where our eyes may focus are obtained by finding the peak responses through a random process. Based on the paradigm of the signal detection theory, the previous model [7] was composed of a dynamic Bayesian network with four layers (See Figure 1): (1) a *saliency map* that shows the average saliency response at each position of a video frame, (2) a *stochastic saliency map* that converts the saliency map into a natural saliency response through a stochastic process, 3) an *eye movement pattern* that predicts the human viewing patterns, and 4) an *eye focusing density map* that predicts regions humans may attend to. Although the model well simulated the human visual system, a visual saliency response on every position was independently calculated, and therefore any spatial relationships have not been considered yet.

We propose a new stochastic model of visual attention that integrates a dynamic Markov random field (MRF) model into the previous stochastic model. The new model with dynamic MRF is a natural extension of the previous model to describe spatiotemporal relationships of one's visual saliency response. The mean field theory [10] provides an analytical solution for quickly estimating visual saliency responses. Model parameters

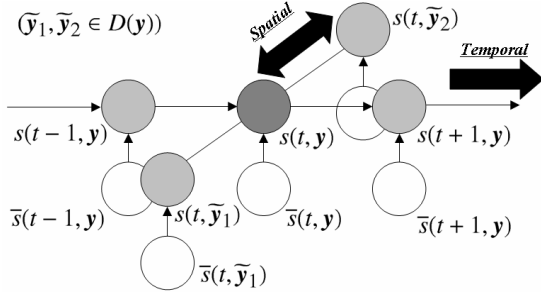


Figure 2. Graphical representation of the dynamic MRF model for estimating stochastic saliency maps.

can be also estimated through the mean field theory and a standard EM algorithm.

2 Model description

2.1 Saliency maps

Consider an input video $I = i(1 : T) = \{i(t)\}_{t=1}^T$ of duration T , where $i(t)$ is the t -th frame of the video I . Then, a sequence $\bar{S} = \bar{S}(1 : T) = \{\bar{S}(t)\}_{t=1}^T$ of saliency maps $\bar{S}(t)$ is obtained from the video I by an existing method (e.g. [3, 5]). Each pixel value $\bar{s}(t, \mathbf{y})$ of the saliency map $\bar{S}(t)$ is called the *saliency*. We have to note that I also represents a set of coordinates in the frame. Each saliency represents the strength of visual stimulus on a position.

2.2 Stochastic saliency maps

When estimating a stochastic saliency map $S(t) = \{s(t, \mathbf{y})\}_{\mathbf{y} \in I}$, we introduce a dynamic Gaussian MRF (GMRF) model to include spatiotemporal relationships, which is characterized by the following equations:

$$\begin{aligned}
 p(S(t), \bar{S}(t) | S(t-1)) &= \frac{1}{Z_s} \exp\{-\Phi(S(t), \bar{S}(t) | S(t-1))\}, \\
 \Phi(S(t), \bar{S}(t) | S(t-1)) &= \sum_{\mathbf{y} \in I} \left\{ f_1(s(t, \mathbf{y}) | s(t-1, \mathbf{y})) + f_2(\bar{s}(t, \mathbf{y}) | s(t, \mathbf{y})) \right. \\
 &\quad \left. + \frac{1}{2} \sum_{\tilde{\mathbf{y}} \in D(\mathbf{y})} f_3(s(t, \mathbf{y}) | s(t, \tilde{\mathbf{y}})) \right\}, \quad (1) \\
 f_i(s | \bar{s}) &\propto -\log \mathcal{G}(s; \bar{s}, \sigma_{si}) \quad \forall s, \bar{s}, (i = 1, 2, 3)
 \end{aligned}$$

where $\mathcal{G}(s; \bar{s}, \sigma)$ is the Gaussian density with argument s , mean \bar{s} and variance σ^2 , $D(\mathbf{y})$ is the set of neighbors of \mathbf{y} , and \propto stands for the proportional indicator. We use 3×3 neighboring system as $D(\mathbf{y})$ in the

implementation. The second term in Eq. (1) implies that a saliency map is observed via a Gaussian random process, and the first and third terms respectively exploit the temporal and spatial continuity of saliency responses.

We employ the mean field approximation [10] to estimate the stochastic saliency map quickly. Assume that the density of the stochastic saliency map at time $t-1$ given saliency maps up to time $t-1$ is given as the following Gaussian density:

$$\begin{aligned}
 p(s(t-1, \mathbf{y}) | \bar{S}(1 : t-1)) \\
 = \mathcal{G}(s(t-1, \mathbf{y}); \hat{s}(t-1, \mathbf{y} | t-1), \sigma_s(t-1, \mathbf{y} | t-1)).
 \end{aligned}$$

Then, the density of the stochastic saliency map at time t given saliency maps up to time t is obtained as follows, where $|D|$ is the number of pixels in $D(\mathbf{y})$:

[Estimation step]

$$\begin{aligned}
 p(s(t, \mathbf{y}) | \bar{S}(1 : t-1)) \\
 = \mathcal{G}(s(t, \mathbf{y}); \hat{s}(t, \mathbf{y} | t-1), \sigma_s(t, \mathbf{y} | t-1)),
 \end{aligned}$$

where

$$\hat{s}(t, \mathbf{y} | t-1) = \lim_{l \rightarrow \infty} \hat{s}^{(l)}(t, \mathbf{y} | t-1),$$

$$\begin{aligned}
 \hat{s}^{(l)}(t, \mathbf{y} | t-1) \\
 = \frac{\sigma_{sp}^2}{\sigma_{s2}^2} \hat{s}(t-1, \mathbf{y} | t-1) + \frac{\sigma_{sp}^2}{\sigma_{s3}^2} \sum_{\tilde{\mathbf{y}} \in D(\mathbf{y})} \hat{s}^{(l-1)}(t, \tilde{\mathbf{y}} | t-1),
 \end{aligned}$$

$$\sigma_s^2(t, \mathbf{y} | t-1) = \sigma_{sp}^2 + \left(\frac{\sigma_{sp}^2}{\sigma_{s3}^2} \right)^2 \sigma_s^2(t-1, \mathbf{y} | t-1),$$

$$\sigma_{sp}^2 = \frac{\sigma_{s2}^2 \sigma_{s3}^2}{|D| \sigma_{s2}^2 + \sigma_{s3}^2} = \left(\frac{1}{\sigma_{s2}^2} + \frac{|D|}{\sigma_{s3}^2} \right)^{-1}.$$

[Update step]

$$p(s(t, \mathbf{y}) | \bar{S}(1 : t)) = \mathcal{G}(s(t, \mathbf{y}); \hat{s}(t, \mathbf{y} | t), \sigma_s(t, \mathbf{y} | t)),$$

where

$$\begin{aligned}
 \hat{s}(t, \mathbf{y} | t) \\
 = \frac{\sigma_s^2(t, \mathbf{y} | t)}{\sigma_s^2(t, \mathbf{y} | t-1)} \hat{s}(t, \mathbf{y} | t-1) + \frac{\sigma_s^2(t, \mathbf{y} | t)}{\sigma_{s1}^2} \bar{s}(t, \mathbf{y}),
 \end{aligned}$$

$$\sigma_s^2(t, \mathbf{y} | t) = \frac{\sigma_{s1}^2 \cdot \sigma_s^2(t, \mathbf{y} | t-1)}{\sigma_{s1}^2 + \sigma_s^2(t, \mathbf{y} | t-1)},$$

2.3 Eye focusing density maps

The method for integrating the stochastic saliency map $S(t)$ and eye movement pattern $u(t)$ is almost the same as the previous stochastic model [7]. Namely, we introduce the following relationship to estimate the eye focusing position $\mathbf{x}(t)$:

$$p(\mathbf{x}(t), u(t) | p(S(t)), \mathbf{x}(t-1), u(t-1))$$

$$= \frac{1}{Z} p(\mathbf{x}(t)|p(S(t))) \cdot p(u(t)|u(t-1)) \cdot p(\mathbf{x}(t)|\mathbf{x}(t-1), u(t)), \quad (2)$$

where Z is a normalizing constant and $p(S(t))$ is the density of the stochastic saliency map at time t .

$$\begin{aligned} p(S(t)) &\stackrel{\text{def.}}{=} \{p(s(t, \mathbf{y}))\}_{\mathbf{y} \in I}, \\ p(s(t, \mathbf{y})) &\stackrel{\text{def.}}{=} p(s(t, \mathbf{y})|\bar{S}(1:t)). \end{aligned}$$

The first term of Eq. (2) represents the fact that the eye focusing position is selected based on the signal detection theory, where the position at which the stochastic saliency takes the maximum is determined as the eye focusing position.

$$\begin{aligned} p(\mathbf{x}(t)|p(S(t))) &= \int_{-\infty}^{\infty} p(s(t, \mathbf{x}(t)) = s) \prod_{\tilde{\mathbf{x}} \neq \mathbf{x}(t)} P(s(t, \tilde{\mathbf{x}}) \leq s) ds, \end{aligned}$$

where $P(s(t, \tilde{\mathbf{y}}) \leq s)$ is the distribution function (i.e. the cumulative density) that corresponds to $p(s(t, \tilde{\mathbf{y}}))$.

The second and third terms of Eq. (2) suggest that the degree of eye movement is driven by the eye movement pattern. We introduce two typical eye movement patterns [8]: 1) The passive state $u(t) = 0$, in which the person's attention stays around one particular position, and 2) the active state $u(t) = 1$, where the person's attention actively moves around in a scene. The transitional probability $p(u(t)|u(t-1))$ is characterized by a 2×2 matrix $\Phi = \{\phi_{(i,j)}\}_{(i,j)}$. Given the eye movement pattern $u(t)$, the probability of eye movements is obtained as follows:

$$\begin{aligned} p(\mathbf{x}(t)|\mathbf{x}(t-1), u(t)) &= \mathcal{L}(\mathbf{x}(t); \mathbf{x}(t-1), \gamma_{x,u(t)}, \sigma_{x,u(t)}), \end{aligned}$$

where $\mathcal{L}(\mathbf{x}; \bar{\mathbf{x}}, \gamma, \sigma)$ is a shifted Gaussian density with argument \mathbf{x} , average $\bar{\mathbf{x}}$, indent γ , and variance σ^2 s.t.

$$\mathcal{L}(\mathbf{x}; \bar{\mathbf{x}}, \gamma, \sigma) \stackrel{\text{def.}}{=} \frac{1}{Z_L} \exp \left\{ -\frac{(\|\mathbf{x} - \bar{\mathbf{x}}\| - \gamma)^2}{2\sigma^2} \right\},$$

Z_L is a normalizing constant, and $\gamma_{x0} < \gamma_{x1}$.

Since it is impractical to calculate Eq. (2), we utilize Monte-Carlo sampling based on the rejection sampling strategy. Each pair of samples from $\tilde{X}(t) = \{\tilde{\mathbf{x}}_n(t)\}_{n=1}^N$ and $\tilde{U}(t) = \{\tilde{u}_n(t)\}_{n=1}^N$ is updated to generate a new sample in $\tilde{X}(t+1)$ and $\tilde{U}(t+1)$ according to (2), where N is the number of samples. The distribution of samples $\tilde{X}(t)$ can then be represented as the eye position map $X(t)$.

3 Parameter estimation

To derive maximum likelihood (ML) model parameters, we divide parameter estimation into two stages.

The first stage estimates the model parameters for computing stochastic saliency maps, and the second stage for estimating eye focusing density maps. The second stage is almost the same as the previous model [7], and therefore we focus only on the first stage.

The first stage derives parameters $\theta_s = (\sigma_{s1}, \sigma_{s2}, \sigma_{s3})$ for computing stochastic saliency maps by using the EM algorithm. The observation is a sequence $\bar{S} = \bar{S}(1:T)$ of saliency maps and the hidden variable is a sequence $S = S(1:T)$ of stochastic saliency maps.

[E step]

The k -th E step updates the density of the stochastic saliency map given the saliency map with the previously estimated parameter $\theta_{s,k-1}$ by making use of Kalman smoother and the mean field approximation. Assume that the density $p(s(t+1, \mathbf{y})|\bar{S}; \theta_{s,k-1})$ of the stochastic saliency $s(t+1, \mathbf{y})$ at position \mathbf{y} and time $t+1$ is given by the following Gaussian density:

$$\begin{aligned} p(s(t+1, \mathbf{y})|\bar{S}) &= \mathcal{G}(s(t+1, \mathbf{y}); \hat{s}(t+1, \mathbf{y}|T), \sigma_s(t+1, \mathbf{y}|T)). \end{aligned}$$

Then, the density $p(s(t, \mathbf{y})|\bar{S}; \theta_{s,k-1})$ of the stochastic saliency $s(t, \mathbf{y})$ at time t is obtained by the following recurrence relation:

$$p(s(t, \mathbf{y})|\bar{S}) = \mathcal{G}(s(t, \mathbf{y}); \hat{s}(t, \mathbf{y}|T), \sigma_s(t, \mathbf{y}|T)),$$

$$\hat{s}_k(t, \mathbf{y}|T) = \lim_{l \rightarrow \infty} \hat{s}_k^{(l)}(t, \mathbf{y}|T),$$

$$\begin{aligned} \hat{s}_k^{(l)}(t, \mathbf{y}|T) &= \frac{\sigma_{sq,k}^2(t, \mathbf{y}|t)}{\sigma_{s2}^2} \hat{s}_k(t+1, \mathbf{y}|T) + \frac{\sigma_{sq,k}^2(t, \mathbf{y}|t)}{\sigma_{s,k}^2(t, \mathbf{y}|t)} \hat{s}_k(t, \mathbf{y}|t) \\ &\quad + \frac{\sigma_{sq,k}^2(t, \mathbf{y}|t)}{\sigma_{s3}^2} \sum_{\tilde{\mathbf{y}} \in D(\mathbf{y})} \hat{s}_k^{(l-1)}(t, \tilde{\mathbf{y}}|T), \end{aligned}$$

$$\sigma_{s,k}^2(t, \mathbf{y}|T) = \sigma_{sq,k}^2(t, \mathbf{y}|t) + \frac{\sigma_{sq,k}^4(t, \mathbf{y}|t)}{\sigma_{s2,k}^4} \sigma_{s,k}^2(t+1, \mathbf{y}|T),$$

$$\sigma_{sq,k}^2(t, \mathbf{y}|t) = \frac{\sigma_{sp,k}^2 \sigma_{s,k}^2(t, \mathbf{y}|t)}{\sigma_{sp,k}^2 + \sigma_{s,k}^2(t, \mathbf{y}|t)},$$

where $\hat{s}_k(t, \mathbf{y}|t)$, $\sigma_{s,k}^2(t, \mathbf{y}|t)$ and $\sigma_{sp,k}^2$ can be obtained by the equations shown in Section 2.2 with the parameter $\theta_{s,k}$.

[M step]

The k -th M step updates the parameter θ_s to maximize the expected log of the density $p(\bar{S}, S; \theta_s)$. We can derive a new parameter $\theta_{s,k}$ from the result of the E step by taking the derivatives of the log likelihood in terms of θ_s and setting to 0.

$$\sigma_{s1,k+1}^2 = \frac{1}{T} \frac{1}{|I|} \sum_{t=1}^T \sum_{\mathbf{y} \in I}$$

$$\begin{aligned}
& \{(\bar{s}(t, \mathbf{y}) - \hat{s}_k(t, \mathbf{y}|T))^2 + \sigma_{s,k}^2(t, \mathbf{y}|T)\}, \\
\sigma_{s2,k+1}^2 &= \frac{1}{T-1} \frac{1}{|I|} \sum_{t=1}^{T-1} \sum_{\mathbf{y} \in I} \\
& \left[(\hat{s}_k(t+1, \mathbf{y}|T) - \hat{s}_k(t, \mathbf{y}|T))^2 \right. \\
& \left. + \sigma_{s,k}^2(t, \mathbf{y}|T) + \frac{\sigma_{s2,k}^2 - \sigma_{s,k}^2(t, \mathbf{y}|t)}{\sigma_{s2,k}^2 + \sigma_{s,k}^2(t, \mathbf{y}|t)} \sigma_{s,k}^2(t+1, \mathbf{y}|T) \right], \\
\sigma_{s3,k+1}^2 &= \frac{1}{T} \frac{1}{|I|} \frac{1}{|D|} \sum_{t=1}^T \sum_{\mathbf{y} \in I} \sum_{\tilde{\mathbf{y}} \in D(\mathbf{y})} \left[\sigma_{s,k}^2(t, \mathbf{y}|T) \right. \\
& \left. + \sigma_{s,k}^2(t, \tilde{\mathbf{y}}|T) + \hat{s}_k(t, \mathbf{y}|T)^2 + \hat{s}_k(t, \tilde{\mathbf{y}}|T)^2 \right. \\
& \left. - 2 \left(\hat{s}_k(t, \mathbf{y}|T) - \frac{\sigma_{sq,k}^2(t, \mathbf{y}|t)}{\sigma_{s3,k}^2} \hat{s}_k(t, \tilde{\mathbf{y}}|T) \right) \right. \\
& \left. \left(\hat{s}_k(t, \tilde{\mathbf{y}}|T) - \frac{\sigma_{sq,k}^2(t, \tilde{\mathbf{y}}|t)}{\sigma_{s3,k}^2} \hat{s}_k(t, \mathbf{y}|T) \right) \right],
\end{aligned}$$

4 Evaluation

4.1 Collecting eye tracking data

For the purpose of parameter estimation and model evaluation, we collected samples of eye focusing positions from six human subjects. Each subject viewed 13 different video clips. The first three video clips are taken from the "Movie Task" video demonstration distributed by VisCog Productions, Inc., and each of the remaining 10 clips comprises a sequence of five to six different natural scenes. Each video was from 30 to 90 seconds long, and the video clip size was 640 x 480 pixels. Each subject's right eye position was recorded at 30 Hz with an eye tracking device [6] based on corneal reflection. We gave no specific instructions to the subject during the experiment.

4.2 Evaluation metric

To quantify how well a model generally predicts actual human eye focusing positions, we used the normalized scanpath saliency (NSS) [8]. Let $R_n(t)$ be a set of all pixels in the circular region centered on the eye focusing position of test subject n with a radius of 30 pixels. Then, the NSS value at time t is defined as

$$NSS(t) = \frac{1}{N_s} \sum_{n=1}^{N_s} \frac{1}{\sigma(p(\mathbf{x}))} \left\{ \max_{\mathbf{x}(t) \in R_n(t)} p(\mathbf{x}(t)) - \bar{p}(\mathbf{x}) \right\},$$

where N_s is the total number of subjects, $\bar{p}(\mathbf{x})$ and $\sigma(p(\mathbf{x}))$ are the mean and the variance of pixel values of the model's output, respectively. $NSS(t) = 1$ indicates that the subjects' eye positions fall on a region whose

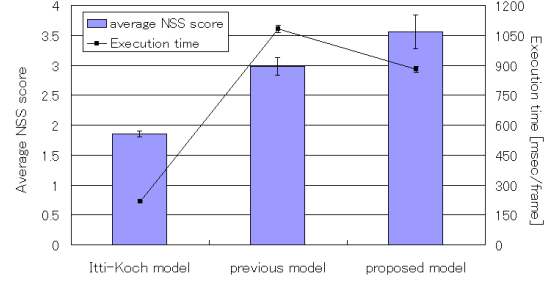


Figure 3. Experimental results

predicted density is one standard deviation above average. Meanwhile, $NSS(t) \leq 0$ indicates that the model performs no better than picking a random position on the map.

4.3 Results

We evaluated the performance of our new model by comparing it with the Itti-Koch model [3] and the previous stochastic model [7]. The video clips are divided into three data sets, and only one data set was retained for evaluation each time with the remaining sets being used as the training data. All the algorithms were implemented with a standard C++ platform (Microsoft Visual C++ .NET, no optimization), and the evaluation were carried out on a standard PC (Intel Core 2 Duo E6850 3.0GHz, 3.0GB RAM).

Figure 3 shows the average NSS scores and the average execution times per frame of all the video clips. Figure 4 shows samples of eye focusing density maps estimated by our new model, where the first and third rows are snapshots taken from an input video, the second and bottom rows are the corresponding eye focusing density maps. The average NSS result indicates that our new model performs about 2 times better than the Itti-Koch model, and 20% better than even the previous stochastic model. The execution time result indicates that our new model needed less calculation cost than the previous stochastic model despite the increase of calculation cost when estimating stochastic saliency maps. This is because eye focusing densities derived from our new model were concentrated into a few small regions, and therefore rejection sampling did not need much time. Although our new model increased the execution time compared with the Itti-Koch model, our model is excellent with parallel computing which can accelerate almost all calculations of our model.

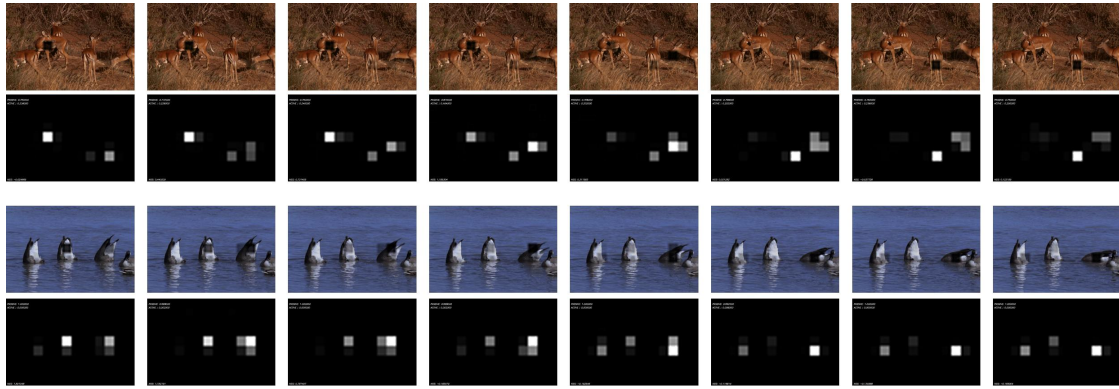


Figure 4. Snapshots of results

5 Conclusion

We have presented a new stochastic model of human visual attention that introduces a dynamic MRF model to include spatiotemporal relationships of visual saliency responses. We predict the likelihood of regions of human attention with 1) the probability of having the maximum saliency response at a given region based on the signal detection theory and 2) the probability of the eye movement based on the predicted cognitive state. Experiments have revealed that our model offers a better eye-gazing prediction than previous models. Future work includes introduction of parallel computing and integration of the proposed model into applications such as object recognition and video retrieval.

Acknowledgement

The authors thank Dr. Hirokazu Kameoka of NTT Communication Science Laboratories for his valuable discussions and helpful comments, which led to improvements of this work. The first author contributed to this work during his internship at NTT Communication Science Laboratories. The authors also thank Dr. Yoshinobu Tonomura, Dr. Naonori Ueda, Dr. Hiroshi Sawada, Dr. Kenji Nakazawa and Dr. Eisaku Maeda of NTT Communication Science Laboratories for their help to the internship.

References

- [1] M. P. Eckstein, J. P. Thomas, J. Palmer, and S. S. Shimozaki. A signal detection model predicts effects of set size on visual search accuracy for feature, conjunction, triple conjunction and disjunction displays. *Perception and Psychophysics*, 62:425–451, 2000.
- [2] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 631–637, June 2005.
- [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.*, 20(11):1254–1259, November 1998.
- [4] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.
- [5] C. Leung, A. Kimura, T. Takeuchi, and K. Kashino. A computational model of saliency depletion/recovery phenomena for the salient region extraction of videos. In *Proc. International Conference on Multimedia and Expo (ICME)*, pages 300–303, July 2007.
- [6] T. Ohno, N. Mukawa, and A. Yoshikawa. FreeGaze: A gaze tracking system for everyday gaze interaction. In *Proc. Symposium on ETRA*, pages 125–132, 2002.
- [7] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino. A stochastic model of selective visual attention with a dynamic bayesian network. In *Proc. International Conference on Multimedia and Expo (ICME)*, pages 1076–1079, June 2008.
- [8] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.
- [9] C. M. Privitera and L. W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(9):970–982, 2000.
- [10] J. Zhang. The mean field theory in EM procedures for Markov random fields. *IEEE Trans. Signal Process.*, 40(10):2570–2583, 1992.