A stochastic model of human visual attention with a dynamic Bayesian network

Akisato Kimura, *Senior Member, IEEE,* Derek Pang, *Student Member, IEEE,* Tatsuto Takeuchi, Kouji Miyazato, Kunio Kashino, *Senior Member, IEEE,* and Junji Yamato, *Senior Member, IEEE.*

✦

Abstract

Recent studies in the field of human vision science suggest that the human responses to the stimuli on a visual display are non-deterministic. People may attend to different locations on the same visual input at the same time. Based on this knowledge, we propose a new stochastic model of visual attention by introducing a dynamic Bayesian network to predict the likelihood of where humans typically focus on a video scene. The proposed model is composed of a dynamic Bayesian network with 4 layers. Our model provides a framework that simulates and combines the visual saliency response and the cognitive state of a person to estimate the most probable attended regions. Sample-based inference with Markov chain Monte-Carlo based particle filter and stream processing with multi-core processors enable us to estimate human visual attention in near real time. Experimental results have demonstrated that our model performs significantly better in predicting human visual attention compared to the previous deterministic models.

Index Terms

Human visual attention, saliency, dynamic Bayesian network, state space model, hidden Markov model, Markov chain Monte-Carlo, particle filter, stream processing.

- *The authors are with NTT Communication Science Laboratories, NTT Corporation, 3-1 Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198 Japan. E-mail: akisato@ieee.org*
- *D. Pang is with Department of Electrical Engineering, Stanford University, Packard 240, 350 Serra Mall, Stanford, CA 94305, USA. He contributed to this work during his internship at NTT Communication Science Laboratories.*
- *K. Miyazato was with Department of Information and Communication Systems Engineering, Okinawa National College of Technology, 905 Henoko, Nago, Okinawa, 905-2192 Japan. He contributed to this work during his internship at NTT Communication Science Laboratories.*
- *Parts of the material in this paper has been presented at IEEE International Conference on Multimedia and Expo (ICME2008), Hannover, Germany, June 2008, and IEEE International Conference on Multimedia and Expo (ICME2009), Cancun, Mexico, June-July 2009.*
- *Manuscript receive March 31 2010.*

arXiv:1004.0085v1 [cs.CV] 1 Apr 2010

Input image   **Saliency map [Itti 1998]**

Fig. 1. An example of a saliency map using Koch-Ullman model

1 INTRODUCTION

Developing a sophisticated object detection and recognition algorithms has been a long distance challenge in computer and robot vision researches. Such algorithms are required in most applications of computational vision, including robotics [1], medical imaging [2], intelligent cars [3], surveillance [4], image segmentation [5], [6] and content-based image retrieval [7]. One of the major challenges in designing generic object detection and recognition systems is to construct methods that are fast and capable of operating on standard computer platforms without any prior knowledge. To that end, pre-selection mechanism would be essential to enable subsequent processing to focus only on relevant data. One promising approach to achieve this mechanism is visual attention: it selects regions in a visual scene that are most likely to contain objects of interest. The field of visual attention is currently the focus of much research for both biological and artificial systems.

Attention is generally controlled by one or a combination of the two mechanisms: 1) a top-down control that voluntarily chooses the focus of attention in a cognitive and task-dependent manner, and 2) a bottom-up control that reflexively directs the visual focus based on the observed saliency attributes. The first biologically-plausible model for explaining the human attention system was proposed by Koch and Ullman [8], which follows the latter approach. The basic concept underlying this model is the *feature integration theory* developed by Treisman and Gelade [9] which has been one of the most influential theories of human visual attention. According to the feature integration theory, in a first step to visual processing, several primary visual features are processed and represented with separate *feature maps* that are later integrated in a *saliency map* that can be accessed in order to direct attention to the most conspicuous areas. In an example shown in Fig. 1, a red car placed on the right in the frame should be attentive, and therefore people directs one's attention to this area. The Koch-Ullman model has been attracting attention of many researchers, especially after the development of an implementation model
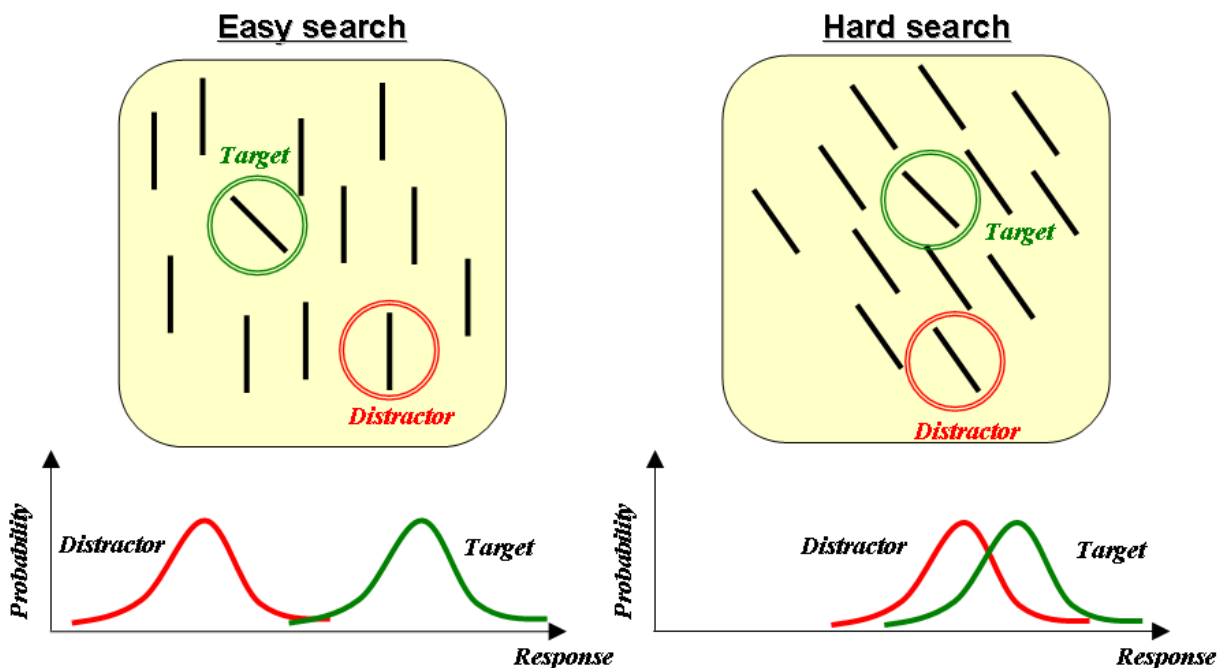
Fig. 2. Visual search response based on the signal detection theory proposed by Eckstein et al. [19], [20].

by Itti, Koch and Niebur [10]. Later, so many attempts have been made to improve the Koch-Ullman model [11], [12], [13], [14], [15] and to extend it to video signals [15], [16], [17], [18].

Although the feature integration theory well explains the early human visual system, a part of the theory includes one crucial problem, namely, people may attend to different locations on the same visual input at the same time. The example shown in Fig. 1 exactly indicates the phenomena: people may pay attention to a blue traffic sign at the center, a white line at the bottom left or others. Previously, this inconsistent visual attention has been considered to be caused by object-based attention, rather than location-based attention [21], which implies that inconsistent visual attention are heavily controlled by higher order processes such as top-down intention, knowledge and preferences. Another typical example can be seen in Fig. 2. Let us consider a search task with a single $45°$ target among a lot of distractors. We can easily understand that the left case is easy and the right case is difficult to find the target. However, based on the feature integration theory, we can immediately identify the target for both easy and hard searches since we always select the location where the response of the detector tuned to the target visual property is greater than at any other locations.

On the other hand, another theory to understanding visual search and attention has been of interest, called the *signal detection theory* [19], [20]. According to this theory, the elements in a visual display are internally represented as independent random variables. Again let us consider the search task shown in

Fig. 2. The response of a detector tuned to the target orientation is represented as a Gaussian density. The response of the same detector to the distractor is also a Gaussian density with lower mean value. For a $45°$ target and vertical distractors, these densities barely overlap, which implies that we can immediately detect the target. On the other hand, in the case of hard search, the target density is identical to the easy search case, but the distractor density is shifted rightward, so that the two densities corresponding to the target and distractor overlap. This implies that the probability we focus on the distractors becomes high and therefore it takes much time to detect the target.

With the paradigm of the signal detection theory, we proposes a new stochastic model of visual attention. With this model, we can automatically predict the likelihood of where humans typically focus on a visual input. The proposed model is composed of a dynamic Bayesian network with four layers: (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 human response through a Gaussian state space model based on the finding of the signal detection theory, 3) an *eye movement pattern* that controls the degree of "overt shifts of attentions" (shifts with saccadic eye movements) through a hidden Markov model (HMM), and 4) an *eye focusing density map* that predicts positions that people probably pay attention to based on the stochastic saliency map and eye movement patterns. When describing the Bayesian network of visual attention, the principle of the signal detection theory is introduced, namely, the position where values of the stochastic saliency map takes the maximum is the eye focusing positions. The proposed model also provides a framework simulating top-down cognitive states of a person at the layer of eye movement patterns. The introduction of eye movement patterns as hidden states of HMMs enables us to describe the mechanism of eye focusing and eye movement naturally.

The paper is organized as follows: Section 2 discusses several related researches that focuses on modeling of human visual attention by using probabilistic techniques or concepts. Section 3 describes the proposed stochastic model of visual attention. Section 4 presents the methods for finding maximum likelihood (ML) estimates of the model parameters based on the Expectation-Maximization (EM) framework. Section 5 discusses some evaluation results. Finally Section 6 summarizes the report and discusses future work.

2 RELATED WORK

Several previous researches focused on modeling of human visual attention by using some kind of probabilistic techniques or concepts. Itti and Baldi [16] investigated a Bayesian approach to detecting surprising events in video signals. Their approach models a surprise by Kullback-Leibler divergence between the prior and posterior distributions of fundamental features. Avraham and Lindenbaum [22] utilized a graphical model approximation to extend their static saliency model based on self similarities. Boccignone [4] introduced a nonparametric Bayesian framework to achieve object-based visual attention.
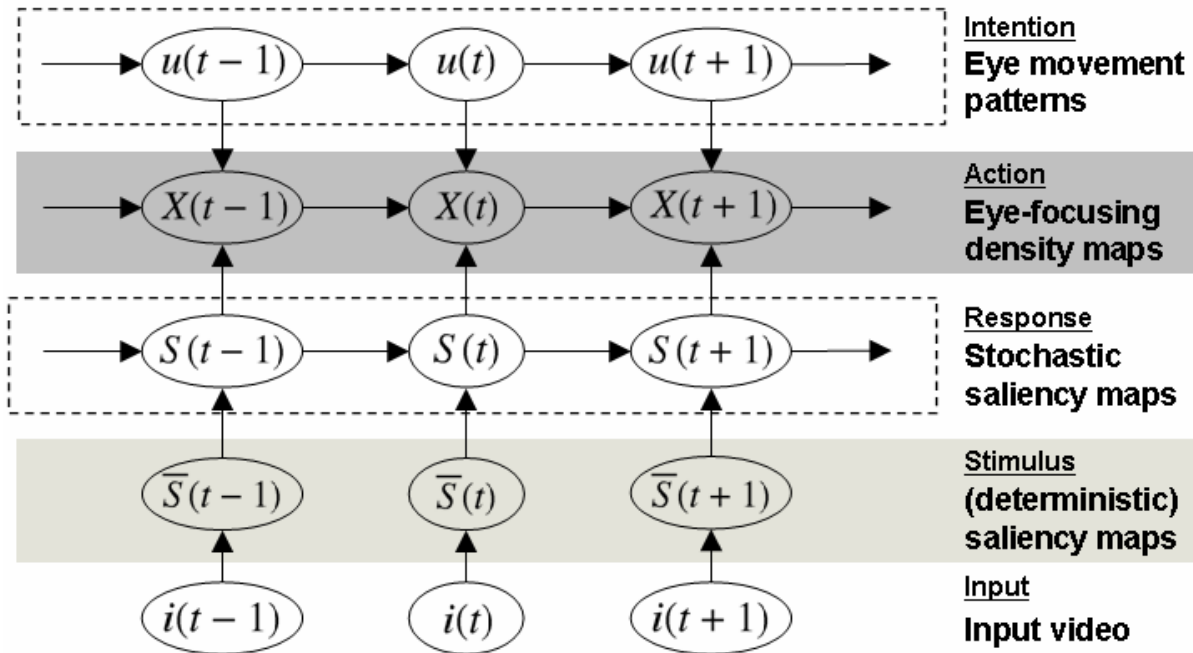
Fig. 3. Graphical representation of the proposed stochastic model of human visual attention, where arrows indicate stochastic dependencies.

Gao, Mahadevan and Vasconcelos [15], [23] developed a decision-theoretic approach attention model for object detection.

The main contribution of our stochastic model against the above previous researches is the introduction of a unified stochastic model that integrates "covert shifts of attention" (shifts of attentions without saccadic eye movements) driven by bottom-up saliency with "overt shifts of attention" (shifts of attention with saccadic eye movements) driven by eye movement patterns by using a dynamic Bayesian network. Our proposed model also provides a framework that simulates and combines the bottom-up visual saliency response and the top-down cognitive state of a person to estimate probable attended regions, if eye movement patterns can deal with more sophisticated top-down information. How to integrate such kinds of top-down information is one of the most important future researches.

3 STOCHASTIC VISUAL ATTENTION MODEL

3.1 Overview

Figs. 3 and 4 illustrates the graphical representation of the proposed visual attention model. The proposed model consists of four layers: (deterministic) saliency maps, stochastic saliency maps, eye focusing positions and eye movement patterns. Before describing the model of the proposed visual attention model, let us introduce several notations and definitions.
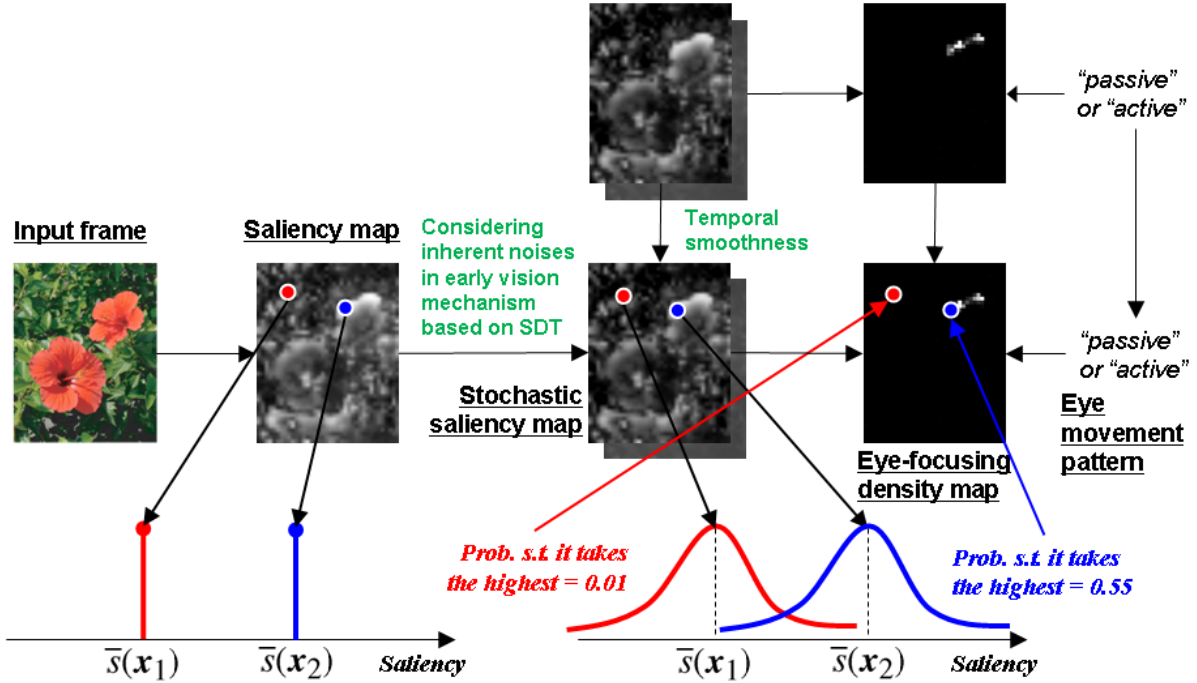
Fig. 4. Procedure of the proposed model

$I = i(1 : T) = \{i(t)\}_{t=1}^{T}$ denotes an input video, where $i(t)$ is the t-th frame of the video I and T is the duration (i.e. the total number of frames) of the video I. The symbol I also denotes a set of coordinates in the frame. For example, a position $\boldsymbol{y}$ in a frame is represented as $\boldsymbol{y} \in I$.

$\overline{S} = \overline{S}(1 : T) = \{\overline{S}(t)\}_{t=1}^{T}$ denotes a *saliency video* which comprises a sequence of *saliency maps* $\overline{S}(t)$ obtained from the input video I. Each saliency map is denoted as $\overline{S}(t) = \{\overline{s}(t, \boldsymbol{y})\}_{\boldsymbol{y} \in I}$, where $\overline{s}(t, \boldsymbol{y})$ is called *saliency* which is the pixel value at the position $\boldsymbol{y} \in I$. Each saliency represents the strength of visual stimulus on the corresponding position of a frame with the real value between 0 and 1.

$S = S(1 : T) = \{S(t)\}_{t=1}^{T}$ denotes a *stochastic saliency video* which comprises a sequence of *stochastic saliency maps* $S(t)$ obtained from the input video I. Each stochastic saliency map is denoted as $S(t) = \{s(t, \boldsymbol{y})\}_{\boldsymbol{y} \in I}$, where $s(t, \boldsymbol{y})$ is called *stochastic saliency* which is the pixel value at the position $\boldsymbol{y} \in I$. Each stochastic saliency corresponds to saliency response perceived through a certain kind of random processes.

$U = u(1 : T) = \{u(t)\}_{t=1}^{T}$ denotes a sequence of *eye movement patterns* each of which represents a pattern of eye movements, A previous research [24] implies that there are two typical patterns [1] of eye movements when one is simply watching a video: 1) Passive state $u(t) = 0$ in which one tends to stay

---

1. Peters and Itti [24] prepared the other pattern, interactive state, which can be seen when playing video games, driving a car or browsing webs. We will omit the interactive state since our setting in this paper does not include any interactions

around one particular position to continuously capture important visual information, and 2) active state $u(t) = 1$ in which one actively moves around and searches various visual information on the scene. Eye movement patterns may reflect purposes or intentions of human eye movements.

$X = X(1 : T) = \{\boldsymbol{x}(t)\}_{t=1}^{T}$ denotes a sequence of eye focusing positions. The proposed model estimates the eye focusing position by integrating the bottom-up information (stochastic saliency maps) and the top-down information (eye movement patterns). A map that represents a density of eye focusing positions is called an *eye focusing density map*.

Only the saliency maps are observed, and therefore eye focusing positions should be estimated under the situation where other layers (stochastic saliency maps and eye movement patterns) are hidden.

In what follows, we denote a probability density function (PDF) of an x as $p(x)$, a conditional PDF of an x given y as $p(x|y)$, and a PDF of x with a parameter θ as $p(x; \theta)$.

The rest of this section describes the detail of the proposed stochastic model and the method for estimating eye focusing positions only from input videos.

3.2 Saliency maps

We used Itti-Koch saliency model [10] shown in Fig. 5 to extract (deterministic) saliency maps. Our implementation includes twelve feature channels sensitive to color contrast (red/green and blue/yellow), temporal luminance flicker, luminance contrast, four orientations ($0°$, $45°$, $90°$ and $135°$), and two oriented motion energies (horizontal and vertical). These features detect spatial outliers in image space using a center-surround architecture. Center and surround scales are obtained from dyadic pyramids with 9 scales, from scale 0 (the original image) to scale 8 (the image reduced by a factor of $2^8 = 256$ in both the horizontal and vertical dimensions). Six center-surround difference maps are then computed as point-wise differences across pyramid scales, for combinations of three center scales ($c = \{2, 3, 4\}$) and two center-surround scale differences ($s = \{3, 4\}$). Each feature map is additionally endowed with internal dynamics that provide a strong spatial within-feature and within-scale competition for activity, followed by within-feature, across-scale competition. In this way, initially noisy feature maps can be reduced to sparse representations of only outlier locations which stand out from their surroundings. All feature maps finally contribute to a unique saliency map representing the conspicuity of each location in the visual field. The saliency map is adjusted with a centrally-weighted 'retinal' filter, putting a higher emphasizes on the saliency values around the center of the video.

3.3 Stochastic saliency maps

When estimating a *stochastic saliency map* $S(t) = \{s(t, \boldsymbol{y})\}_{\boldsymbol{y} \in I}$, we introduce a pixel-wise state space model characterized by the following two relationships:

$$p(\overline{s}(t, \boldsymbol{y})|s(t, \boldsymbol{y})) = \mathcal{N}(s(t, \boldsymbol{y}), \sigma_{s1}),$$
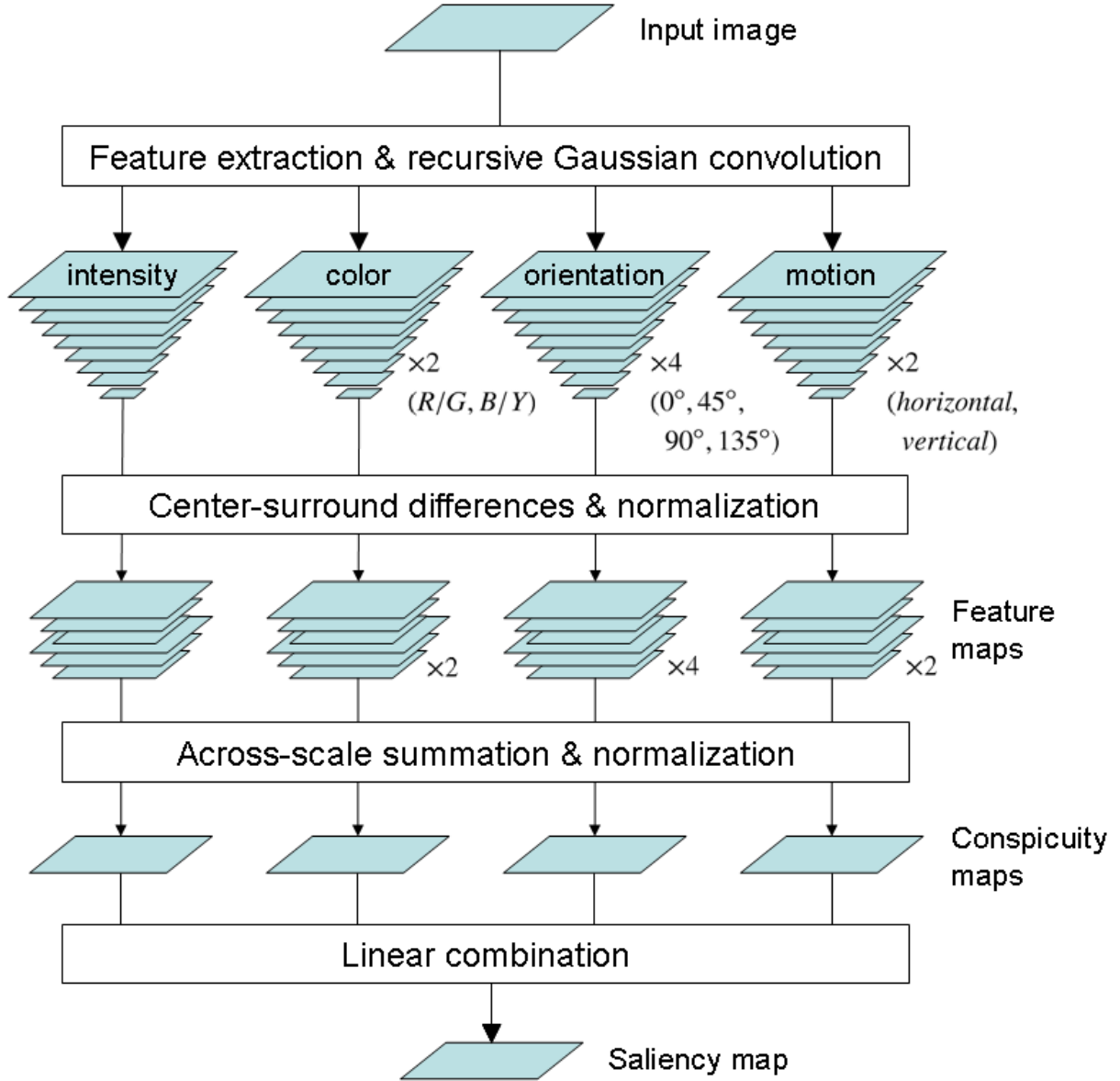
Fig. 5. Saliency map extraction using Itti et al.'s model

$$p(s(t, \boldsymbol{y})|s(t-1, \boldsymbol{y})) \quad = \quad \mathcal{N}(s(t-1, \boldsymbol{y}), \sigma_{s2}),$$

where $\mathcal{N}(\bar{s}, \sigma)$ is the Gaussian PDF with mean $\bar{s}$ and variance σ^2. The first equation in the above model implies that a saliency map is observed through a Gaussian random process, and the second equation exploits the temporal characteristics of the human visual system. For brevity, only in this section we will omit the position $\boldsymbol{y}$ where explicit expression is unnecessary, e.g. $s(t)$ instead of $s(t, \boldsymbol{y})$.

We employ a Kalman filter to recursively compute the stochastic saliency map. Assume that the density at each position on the stochastic saliency map $s(t-1)$ at time $t-1$ given saliency maps $\bar{s}(1:t-1)$ up

to time $t-1$ is given as the following Gaussian PDF:

$$p(s(t-1)|\bar{s}(1:t-1))$$
$$= \mathcal{N}(\widehat{s}(t-1|t-1), \sigma_s(t-1|t-1)).$$

where the position $\boldsymbol{y}$ is omitted for simplicity. Then, the density $p(s(t)|\bar{s}(1:t))$ of the stochastic saliency map at time t is updated by the following recurrence relations with the saliency maps $\bar{s}(1:t)$ up to time t:

[Estimation step]

$$p(s(t)|\bar{s}(1:t-1)) = \mathcal{N}(\widehat{s}(t|t-1), \sigma_s(t|t-1)),$$

where

$$\widehat{s}(t|t-1) = \widehat{s}(t-1|t-1),$$
$$\sigma_s^2(t|t-1) = \sigma_{s2}^2 + \sigma_s^2(t-1|t-1).$$

[Update step]

$$p(s(t)|\bar{s}(1:t)) = \mathcal{N}(\widehat{s}(t|t), \sigma_s(t|t)),$$

where

$$\widehat{s}(t|t) = \tag{1}$$
$$\frac{\sigma_{s1}^2}{\sigma_{s1}^2 + \sigma_s^2(t|t-1)}\widehat{s}(t|t-1) + \frac{\sigma_s^2(t|t-1)}{\sigma_{s1}^2 + \sigma_s^2(t|t-1)}\bar{s}(t),$$
$$\sigma_s^2(t|t) = \frac{\sigma_{s1}^2 \cdot \sigma_s^2(t|t-1)}{\sigma_{s1}^2 + \sigma_s^2(t|t-1)}, \tag{2}$$

Remark 1: The above model implies that model parameters $(\sigma_{s1}, \sigma_{s2})$ of every Gaussian random variable is independent from the frame index t and the position $\boldsymbol{y}$. We can easily extend the model to consider adaptive model parameters depending on the frame index and the position. In this case, model parameters can be updated via on-line learning with adaptive Kalman filters (e.g. [25], [26].) *(Remark 1 ends.)*

3.4 Estimating eye motions

3.4.1 Overview

By incorporating the stochastic saliency map $S(t) = \{s(t, \boldsymbol{y})\}_{\boldsymbol{y} \in I}$ and the *eye movement pattern* $u(t)$, we introduce the following transition PDF to estimate the *eye focusing position* $\boldsymbol{x}(t)$ such that

$$p(\boldsymbol{x}(t), u(t)|p(S(t)), \boldsymbol{x}(t-1), u(t-1))$$
$$\propto p(\boldsymbol{x}(t)|p(S(t)))$$
$$\cdot p(u(t)|u(t-1)) \cdot p(\boldsymbol{x}(t)|\boldsymbol{x}(t-1), u(t)), \tag{3}$$

where the PDF of the stochastic saliency map at time t is represented as $p(S(t))$ for simplicity, namely

$$p(S(t)) \;=\; \{p(s(t, \boldsymbol{y}))\}_{\boldsymbol{y} \in I},$$

$$p(s(t, \boldsymbol{y})) \;=\; p(s(t, \boldsymbol{y}) | \overline{s}(1:t, \boldsymbol{y})) \quad \forall \boldsymbol{y} \in I.$$

The stochastic saliency map $S(t)$ controls "covert shifts of attention" through the PDF $p(\boldsymbol{x}(t) | p(S(t)))$ [2]. On the other hand, the eye movement pattern $u(t)$ controls the degree of "overt shifts of attention". In what follows, we call a pair $z(t) = (\boldsymbol{x}(t), u(t))$, consisting of an eye focusing position and an eye movement pattern, as the *eye focusing state $z(t)$* for brevity. The following PDF of eye focusing positions $\boldsymbol{x}(t)$ given a PDF $p(S(1:t))$ of stochastic saliency maps up to time t characterizes an eye focusing density map at time t:

$$p(\boldsymbol{x}(t) | p(S(1:t))) = \sum_{u(t)=0,1} p(z(t) | p(S(1:t))), \tag{4}$$

$$p(z(t) | p(S(1:t))) = \int_{z(t-1)} p(z(t-1) | p(S(1:t-1))) \\ \cdot p(z(t) | p(S(t)), z(t-1)) dz(t-1). \tag{5}$$

Since the formula for computing Eq. (4) cannot be derived, we introduce a technique inspired by a particle filter with Markov chain Monte-Carlo (MCMC) sampling instead. The PDF of eye focusing states shown in Eq. (5) can be approximated by samples of eye focusing states $\{z_n(t)\}_{n=1}^{N}$ and the associated weights $\{w_n(t)\}_{n=1}^{N}$ as

$$p(z(t) | p(S(1:t))) \;\approx\; \sum_{n=1}^{N} w_n(t) \cdot \delta(z(t), z_n(t)), \tag{6}$$

where N is the number of samples, and $\delta(\cdot, \cdot)$ represents Kronecker delta.

Fig. 6 shows the procedure for estimating eye focusing density maps, which can be separated into three steps: 1) generating samples from the PDFs $p(u(t)|u(t-1))$ and $p(\boldsymbol{x}(t)|\boldsymbol{x}(t-1), u(t))$ derived from an eye movement pattern, 2) weighting samples with the PDF $p(\boldsymbol{x}(t)|p(S(t)))$ derived from a stochastic saliency map, and 3) re-sampling if necessary. We now describe each step in detail.

3.4.2 Propagation with eye movement patterns

The second and third terms of Equation (3) suggests that the current eye focusing position depends on the previous eye focusing position, and the degree of eye movements is driven by one's eye movement pattern $u(t)$.

---

2. The notation $p(\boldsymbol{x}(t)|p(S(t)))$ seems to be unusual, however, the PDF of eye focusing positions $\boldsymbol{x}(t)$ estimated from the stochastic saliency map $S(t)$ can be determined by the PDF of the stochastic saliency map, not the stochastic saliency map itself, as shown in Section 3.4.3.

$$p(z(t)|p(S(t)), z(t-1))$$
$$= p(\boldsymbol{x}(t), u(t)|p(S(t)), \boldsymbol{x}(t-1), u(t-1))$$
$$\propto \underline{p(\boldsymbol{x}(t)|p(S(t)))} \cdot \underline{p(u(t)|u(t-1)) \cdot p(\boldsymbol{x}(t)|\boldsymbol{x}(t-1), u(t))}$$
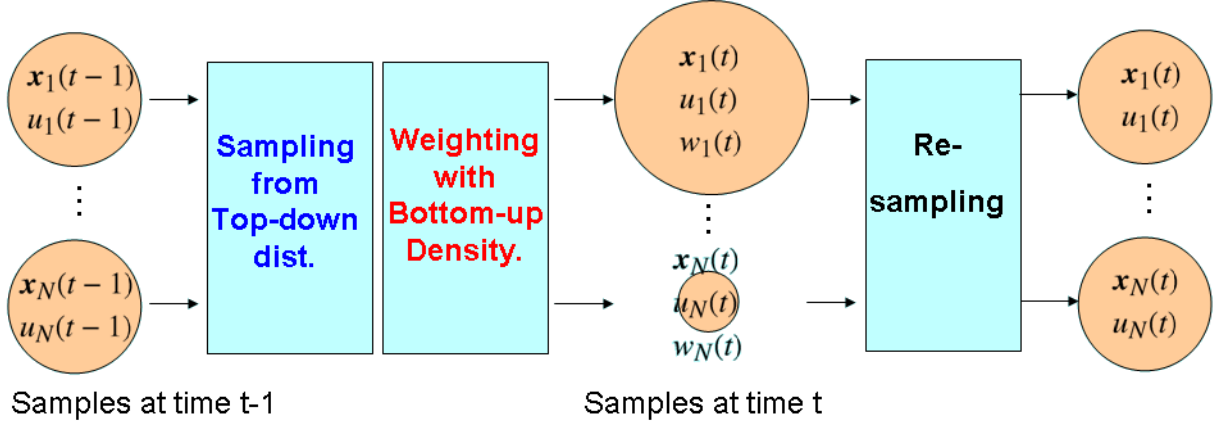
Bottom-up   **Top-down**

Fig. 6. Strategy for calculating eye focusing density maps

The second term $p(u(t)|u(t-1))$ of Equation (3) is characterized by the the transitional probability $\Phi = \{\phi_{(i,j)}\}$ of eye movement patterns defined by a 2×2 matrix given in advance.

$$p(u(t)|u(t-1)) = \phi_{(u(t),u(t-1))} \tag{7}$$

The third term $p(\boldsymbol{x}(t)|\boldsymbol{x}(t-1), u(t))$ of Equation (3) represents the transition PDF of eye focusing positions governed by the eye movement patterns at the current time (See Figure 7), defined as

$$p(\boldsymbol{x}(t)|\boldsymbol{x}(t-1), u(t)) = \mathcal{L}(\boldsymbol{x}(t-1), \gamma_{x,u(t)}, \sigma_{x,u(t)}), \tag{8}$$

where γ_{xi} and σ_{xi} $(i = 1, 2)$ are model parameters that represents the average and standard deviation of distances of eye movements, and $\mathcal{L}(\overline{\boldsymbol{x}}, \gamma, \sigma)$ is a shifted 2D Gaussian PDF with mean $\overline{\boldsymbol{x}}$, indent γ and variance σ^2 such that

$$\mathcal{L}(\overline{\boldsymbol{x}}, \gamma, \sigma) \propto \exp\left\{-\frac{(\|\boldsymbol{x} - \overline{\boldsymbol{x}}\| - \gamma)^2}{2\sigma^2}\right\}.$$

Samples $\{\boldsymbol{z}_n(t)\}_{n=1}^N$ of eye focusing states are generated with a technique of MCMC sampling. Suppose that samples $\{\boldsymbol{z}_n(t-1)\}_{n=1}^N$ of eye focusing states at time $t-1$ have already been obtained. Then, samples $\{\boldsymbol{z}_n(t)\}_{n=1}^N$ at time t are drawn by using the second and third terms of Equation (3) with the Metropolis algorithm [27] such as
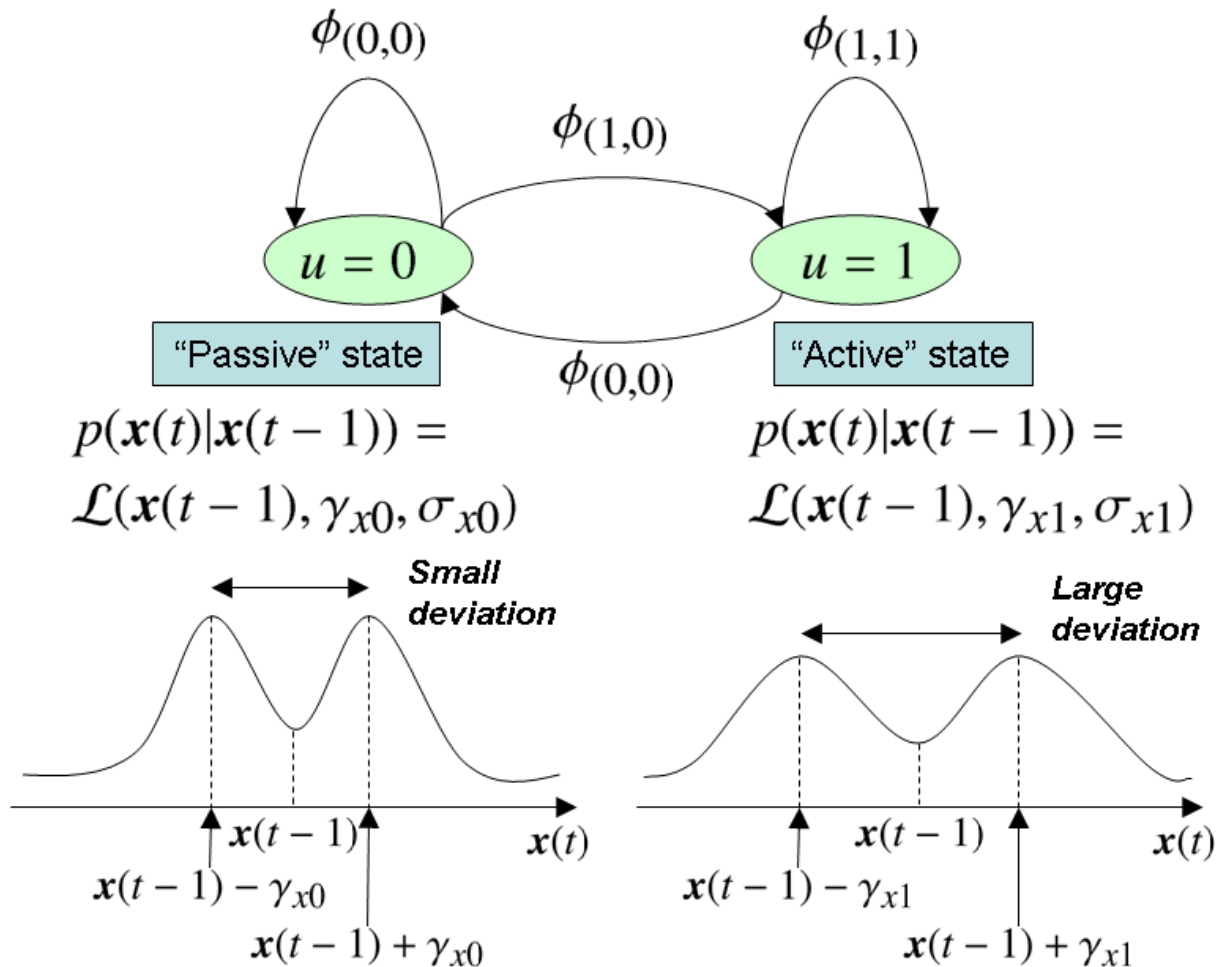
$$\boldsymbol{z}_n(t) = \{\boldsymbol{x}_n(t), \boldsymbol{u}_n(t)\},$$

$\phi_{(0,0)}$

$\phi_{(1,1)}$

$\phi_{(1,0)}$

$u = 0$

$u = 1$

"Passive" state

"Active" state

$\phi_{(0,0)}$

$$p(x(t)|x(t-1)) = \mathcal{L}(x(t-1), \gamma_{x0}, \sigma_{x0})$$

$$p(x(t)|x(t-1)) = \mathcal{L}(x(t-1), \gamma_{x1}, \sigma_{x1})$$

Small deviation

Large deviation

$x(t-1)$

$x(t)$

$x(t-1) - \gamma_{x0}$

$x(t-1) + \gamma_{x0}$

$x(t-1)$

$x(t)$

$x(t-1) - \gamma_{x1}$

$x(t-1) + \gamma_{x1}$

Fig. 7. Transition PDF of eye focusing positions governed by the eye movement patterns

$$u_n(t) \sim p(u(t)|u_n(t-1)),$$

$$x_n(t) \sim p(x(t)|x_n(t-1), u_n(t)),$$

where $x \sim p(x)$ indicates that a sample x is drawn from a PDF $p(x)$. This top-down part corresponds to the propagation step of a particle filter.

3.4.3 Updating with stochastic saliency maps

As the second step, sample weights $\{w_n(t)\}_{n=1}^N$ are updated based on the first term $p(x(t)|p(S(t)))$ of Equation (3). Formally, the weight $w_n(t)$ of the n-th sample $z_n(t)$ at time t can be calculated as

$$w_n(t) \propto w_n(t-1) \cdot p(x(t) = x_n(t)|p(S(t))).$$

As shown in Equation (6), samples $\{z_n(t)\}_{n=1}^N$ of eye focusing states and the associated weights $\{w_n(t)\}_{n=1}^N$ comprise an eye focusing density map at time t. This step corresponds to the update step of a particle

filter.

The first term of Equation (3) 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. In other words, this term computes the probability at each position that the stochastic saliency takes the maximum, which can be calculated as

$$
\begin{aligned}
&p(\boldsymbol{x}(t)|p(S(t))) \\
&= \int_{-\infty}^{\infty} p(s(t, \boldsymbol{x}(t)) = s) \prod_{\substack{\widetilde{\boldsymbol{x}} \in I, \\ \widetilde{\boldsymbol{x}} \neq \boldsymbol{x}(t)}} P(s(t, \widetilde{\boldsymbol{x}}) \leq s) ds,
\end{aligned}
\tag{9}
$$

where $P(s(t, \widetilde{\boldsymbol{y}}) \leq s)$ is the cumulative density function (CDF) that corresponds to the PDF $p(s(t, \widetilde{\boldsymbol{y}}))$ of the stochastic saliency $s(t, \widetilde{\boldsymbol{y}})$. The first part of Equation (9) stands for the probability such that a stochastic saliency value at position $\boldsymbol{x}(t)$ equals s, and the second part represents the probability such that stochastic saliency values at any other positions are smaller than s.

Direct computation of Equation (9) is intractable. Instead, we introduce an alternative expression of Equation (9) that is applicable to stream processing with multi-core processors.

$$
\begin{aligned}
&p(\boldsymbol{x}(t)|p(S(t))) \\
&= \int_{-\infty}^{\infty} \frac{p(s(t, \boldsymbol{x}(t)) = s)}{P(s(t, \boldsymbol{x}(t)) \leq s)} \prod_{\widetilde{\boldsymbol{x}} \in I} P(s(t, \widetilde{\boldsymbol{x}}) \leq s) ds.
\end{aligned}
\tag{10}
$$

The latter part of Eq. (10) does not depend on the position $\boldsymbol{x}(t)$, which implies that it can be calculated in advance for every s. This calculation can be executed in $\mathcal{O}(\log |I|)$ time through a tree-based multiplication and parallelization at each level (cf. Fig. 8). Also, the former part of Eq. (10) can be calculated independently for each position $\boldsymbol{x}(t)$. Therefore, once the calculation of the latter part has finished, Eq. (10) can be calculated in $\mathcal{O}(\log |S|)$ time with a combination of tree-based addition and pixel-wise parallelization, where $|S|$ stands for the resolution of the integral in Eq. (10).

3.4.4 Re-sampling

Finally, re-sampling is performed to eliminate samples with low importance weights and multiply samples with high importance weights. This step enables us to avoid "degeneracy" problem, namely, to avoid the situation where all but one of the importance weights are close to zero. Although the effective number of samples [28] is frequently used as a criterion for re-sampling, we execute re-sampling at regular time intervals.

Remark 2: We have to note that the whole procedure which includes the propagation, updating and re-sampling steps for estimating eye focusing density maps is equivalent to a particle filter with MCMC sampling since the the PDFs used in the propagation and update steps are mutually independent with each other. *(Remark 2 ends.)*
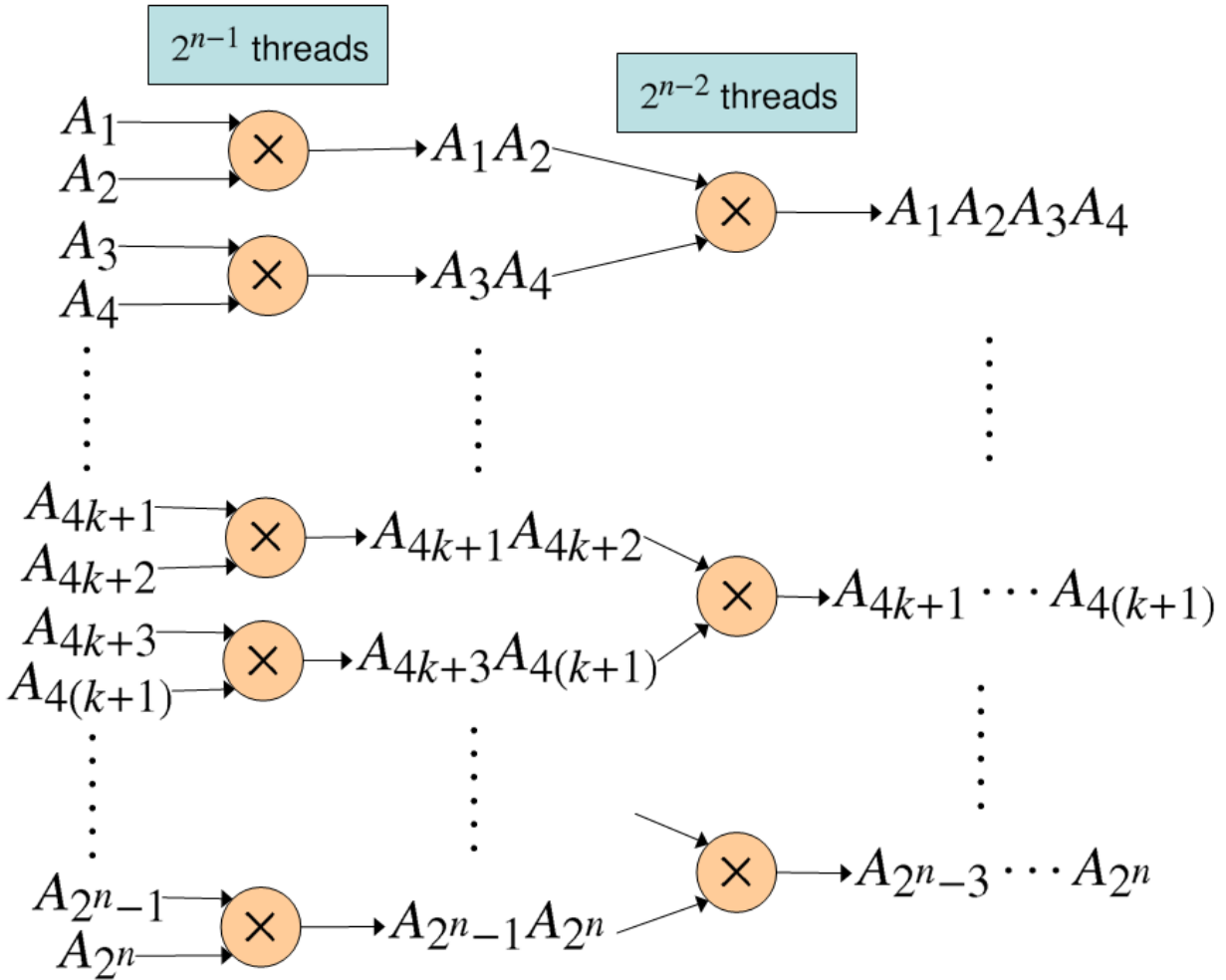
Fig. 8. Tree-based multiplication for computing the product $A_1 A_2 \cdots A_{2^n}$

4 MODEL PARAMETER ESTIMATION

This section focuses on the problem of estimating maximum likelihood (ML) model parameters. Fig. 9 shows the block diagram of our model parameter estimation. We can automatically estimate almost all the model parameters in advance by utilizing saliency maps calculated from the input video and eye focusing positions obtained by some eye tracking devices as observations. Simultaneous estimation of all ML parameters can be optimal but impractical due to the substantial calculation cost. Therefore, we separate our parameter estimation into two independent stages. The first stage derives parameters for computing stochastic saliency maps, and the second stage for estimating eye focusing points.
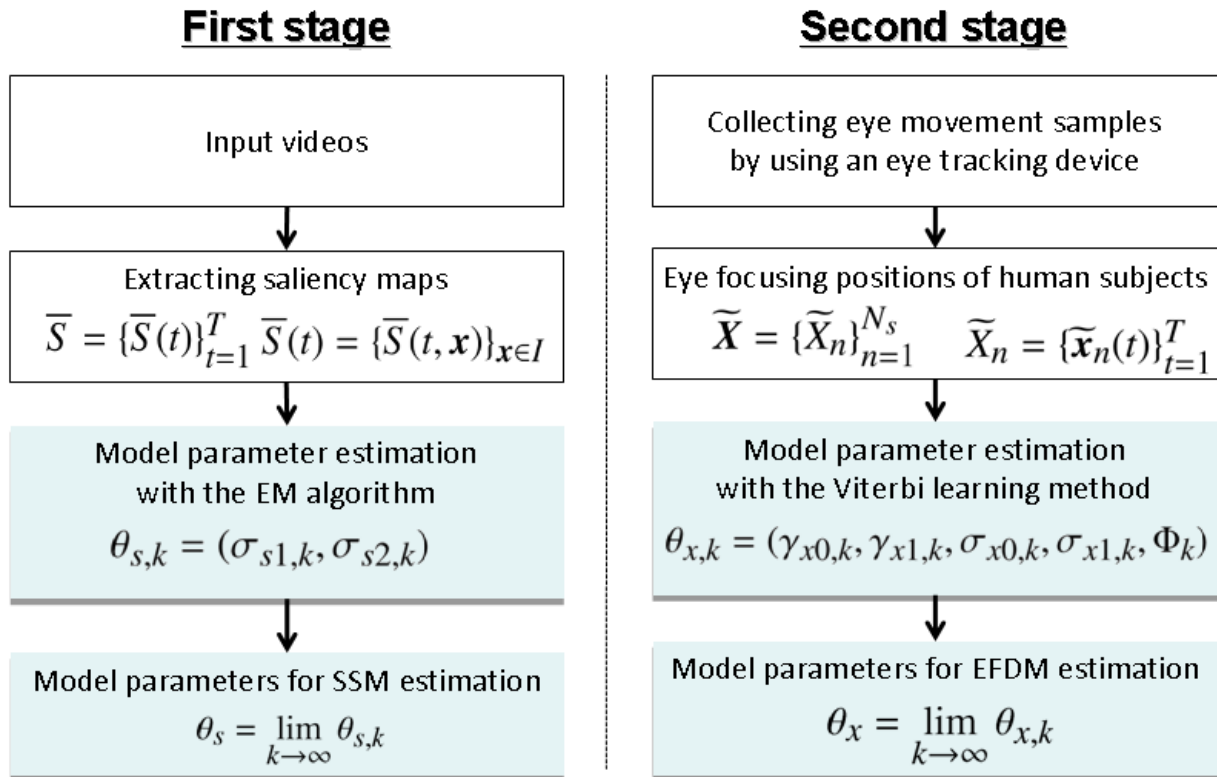
First stage

Input videos

↓

Extracting saliency maps
$$\overline{S} = \{\overline{S}(t)\}_{t=1}^{T} \quad \overline{S}(t) = \{\overline{S}(t, \boldsymbol{x})\}_{\boldsymbol{x} \in I}$$

↓

Model parameter estimation with the EM algorithm
$$\theta_{s,k} = (\sigma_{s1,k}, \sigma_{s2,k})$$

↓

Model parameters for SSM estimation
$$\theta_s = \lim_{k \to \infty} \theta_{s,k}$$

Second stage

Collecting eye movement samples by using an eye tracking device

↓

Eye focusing positions of human subjects
$$\widetilde{X} = \{\widetilde{X}_n\}_{n=1}^{N_s} \quad \widetilde{X}_n = \{\widetilde{\boldsymbol{x}}_n(t)\}_{t=1}^{T}$$

↓

Model parameter estimation with the Viterbi learning method
$$\theta_{x,k} = (\gamma_{x0,k}, \gamma_{x1,k}, \sigma_{x0,k}, \sigma_{x1,k}, \Phi_k)$$

↓

Model parameters for EFDM estimation
$$\theta_x = \lim_{k \to \infty} \theta_{x,k}$$

Fig. 9. Block diagram of our model parameter estimation

4.1 Parameters for stochastic saliency maps

The first stage derives parameters for computing stochastic saliency maps. Here, we introduce the EM algorithm. In this case, the observations are the saliency maps $\overline{S} = \overline{S}(1:T)$ and the hidden variables are the stochastic saliency maps $S = S(1:T)$. Remember that T is the duration of the video. The EM algorithm for estimating $\theta_s = (\sigma_{s1}, \sigma_{s2})$ is as follows:

$(k+1)$-th E step

The E step updates the PDF $p(S|\overline{S}; \theta_{s,k})$ of the stochastic saliency maps S given the saliency maps $\overline{S}$ with the previously estimated parameter $\theta_{s,k} = (\sigma_{s1}, \sigma_{s2})$ by using Kalman smoother. In detail, the objective is to recursively compute the mean $\widehat{s}(t|T)$ and standard deviation $\sigma_s(t|T)$ of the stochastic saliency $s(t)$ at time $t = 1, 2, \cdots, T$, where all the saliency maps $\overline{S}$ are used as observations. Note that the position $\boldsymbol{y}$ is omitted for simplicity.

Suppose that the PDF of the stochastic saliency at time $t+1$ is given by the following Gaussian PDF:

$$p(s(t+1)|\overline{S}; \theta_{s,k}) \quad = \quad \mathcal{N}(\widehat{s}_k(t+1|T), \sigma_{s,k}(t+1|T)).$$

Then, the PDF of the stochastic saliency at time t is obtained by the following recurrence relation:

$$p(s(t)|\overline{S}; \theta_{s,k}) \quad = \quad \mathcal{N}(\widehat{s}_k(t|T), \sigma_{s,k}(t|T)),$$

where

$$\widehat{s}_k(t|T) \quad = \quad \frac{\sigma_{sq,k}^2(t|t)}{\sigma_{s,k}^2(t|t)}\widehat{s}_k(t|t) + \frac{\sigma_{sq,k}^2(t|t)}{\sigma_{s2,k}^2}\widehat{s}_k(t+1|T),$$

$$\sigma_{s,k}^2(t|T) \quad = \quad \sigma_{sq,k}^2(t|t) + \left(\frac{\sigma_{sq,k}^2(t|t)}{\sigma_{s2,k}^2}\right)^2 \sigma_{s,k}^2(t+1|T),$$

$$\sigma_{sq,k}^2(t|t) \quad = \quad \frac{\sigma_{s2,k}^2 \sigma_{s,k}^2(t|t)}{\sigma_{s2,k}^2 + \sigma_{s,k}^2(t|t)},$$

and $\widehat{s}_k(t|t)$ and $\sigma_{s,k}^2(t|t)$ can be obtained by Eqs. (1)(2) with the parameter $\theta_{s,k}$.

$(k+1)$-th M step

The M step updates the parameter θ_s to maximize the expected log likelihood of the PDF $p(\overline{S}, S; \theta_s)$. We can derive a new parameter $\theta_{s,k+1}$ from the result of the E step by taking the derivatives of the log likelihood in terms of θ_s and setting to 0.

$$\sigma_{s1,k+1}^2 = \frac{1}{T}\sum_{t=1}^{T}\left\{(\overline{s}(t) - \widehat{s}_k(t|T))^2 + \sigma_{s,k}^2(t|T)\right\}$$

$$\sigma_{s2,k+1}^2 = \frac{1}{T-1}\sum_{t=2}^{T}\left[(\widehat{s}_k(t|T) - \widehat{s}_k(t-1|T))^2\right.$$

$$\left. +\sigma_{s,k}^2(t-1|T) + \frac{\sigma_{s2,k}^2 - \sigma_{s,k}^2(t-1|t-1)}{\sigma_{s2,k}^2 + \sigma_{s,k}^2(t-1|t-1)}\sigma_{s,k}^2(t|T)\right]$$

4.2 Parameters for eye focusing positions

The second stage derives parameters $\theta_x = (\gamma_{x0}, \sigma_{x0}, \gamma_{x1}, \sigma_{x1}, \Phi)$ for computing eye focusing positions. The observations are the sequence of eye focusing positions $\boldsymbol{X} = \boldsymbol{x}(1:T)$ obtained from some eye tracking devices, and the hidden states are the eye movement patterns $U = u(1:T)$. In this section, we introduce an alternative notation of eye movement patterns as $u(t) = (u(t)_0, u(t)_1)^T$, which is a 2-dimensional binary vector such that $(1,0)^T$ denotes the passive state, and $(0,1)^T$ represents the active state.

We take a Viterbi learning approach for its quick convergence. It recursively updates the eye movement patterns $U = u(1:T)$ and the ML parameter set θ_x to maximize the posterior $p(U|\boldsymbol{X}; \theta_x)$.

Initializing eye movement patterns

We have to start with determining an initial sequence $U_0 = u_0(1:T)$ of eye movement patterns. We introduce the following decision rule:

$$u_0(t) \quad = \quad \begin{cases} 0 & \text{if } \|\boldsymbol{x}(t) - \boldsymbol{x}(t-1)\| \leq \kappa_x \\ 1 & \text{if } \|\boldsymbol{x}(t) - \boldsymbol{x}(t-1)\| > \kappa_x \end{cases},$$

where κ_x is a given threshold.

<u>The $(k+1)$-th step for updating hidden variables</u>

This step updates the sequence U of eye movement patterns to maximize the posterior density $p(U|\boldsymbol{X};\theta_{x,k})$ given the parameter set $\theta_{x,k}$ obtained in the previous step.

$$U_{k+1} \quad \leftarrow \quad \arg\max_{U} p(U|\boldsymbol{X};\theta_{x,k}).$$

Viterbi algorithm (e.g. [29], [30]) can derive the ML sequence U_{k+1} of eye movement patterns.

<u>The $(k+1)$-th step for updating the parameter set</u>

This step updates the parameter set θ_x to maximize the posterior density $p(U_{k+1}|\boldsymbol{X};\theta_x)$.

$$\theta_{x,k+1} \quad = \quad \arg\max_{\theta_x} p(U_{k+1}|\boldsymbol{X};\theta_x).$$

Taking the derivative of the log likelihood in terms of θ_x, we obtain

$$\gamma_{xi,k+1}$$
$$= \quad \frac{\sum_{t=2}^{T} \|\boldsymbol{x}(t) - \boldsymbol{x}(t-1)\| u_{k+1}(t)_i}{\sum_{t=2}^{T} u_{k+1}(t)_i},$$
$$\sigma_{xi,k+1}^2$$
$$= \quad \frac{\sum_{t=2}^{T}(\|\boldsymbol{x}(t) - \boldsymbol{x}(t-1)\| - \gamma_{xi,k})^2 u_{k+1}(t)_i}{2\sum_{t=2}^{T} u_{k+1}(t)_i},$$
$$\phi_{(i,j),k+1}$$
$$= \quad \frac{\sum_{t=2}^{T} u_{k+1}(t)_i u_{k+1}(t-1)_j}{\sum_{t=2}^{T} u_{k+1}(t-1)_j}.$$

5 EVALUATION

5.1 Evaluation conditions

For the accuracy evaluation, we used CRCNS eye-1 database created by University of South California. This database includes 100 video clips (MPEG-1, 640×480 pixels, 30fps) and eye traces when showing these video clips to 8 human subjects (4-6 available eye traces for each video clip, 240fps). Other details for the database can be found in *https://crcns.org/files/data/eye-1/crcns-eye1-summary.pdf*. In this evaluation, we used 50 video clips (about 25 minutes in total) called "original experiment" and associated eye traces.

Model parameters were derived in advance with the learning algorithm presented in Section 4. In this time, we used 5-fold cross validation so that 40 video clips and associated eye traces were used as the training data for evaluating the remaining data (10 videos and associated eye traces).

All the algorithms were implemented with a standard C++ platform and NVIDIA CUDA, and the evaluation were carried out on a standard PC with a graphics processor unit (GPU). The detailed information for the platform used in this evaluation is listed in Table 1.

TABLE 1

Platform used in the evaluation

OS	Windows Vista Ultimate
Development platform	Microsoft Visual C++ 2008 OpenCV 1.1pre & NVIDIA CUDA 2.2
Optimization	disabled
CPU	Intel Core2 Quad Q6600 (2.40GHz)
RAM	4.0GB
GPU	NVIDIA GeForce GTX275 ×2 SLI

5.2 Evaluation metric

As a metric to quantify how well a model predicts the actual human eye focusing positions, we used the normalized scan-path saliency (NSS) used in the previous work [24]. Let $R_j(t)$ be a set of all pixels in a circular region centered on the eye focusing position of test subject j with a radius of 30 pixels. Then, the NSS value at time t is defined as

$$NSS(t) = \frac{1}{N_s} \sum_{j=1}^{N_s} \frac{1}{\sigma(p(\boldsymbol{x}))} \left\{ \max_{\boldsymbol{x}(t) \in R_j(t)} p(\boldsymbol{x}(t)) - \overline{p}(\boldsymbol{x}) \right\},$$

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

5.3 Results

We compared our proposed method with 3 existing computational models: 1) a simple control measuring local pixel variance (denoted "variance") [16], 2) a saliency map (denoted "CIOFM") [10], and 3) Bayesian surprise (denoted "surprise") [16]. All the outputs emitted from the above existing models are included in CRCNS eye-1 database, and therefore we directly utilized them for the evaluation.

Fig. 10 shows the model accuracy measured by the average NSS score with standard errors for all the video clips, and Fig. 11 details the average NSS score for each video clip. The order of video clips is sorted beforehand to keep the visibility. The result shown in Fig. 10 indicates that the our new method achieved significantly better scores than all 3 existing methods, which implies that our proposed method can estimate human visual attention with high accuracy. Also, the result shown in Fig. 11 indicates that our proposed method marked almost the same as or much better than all the existing methods for most of the video clips.

Fig. 12 shows snapshots of outputs from Itti model (the second and fifth rows) and our proposed method (the third and sixth rows). It illustrates that outputs from Itti model included several large
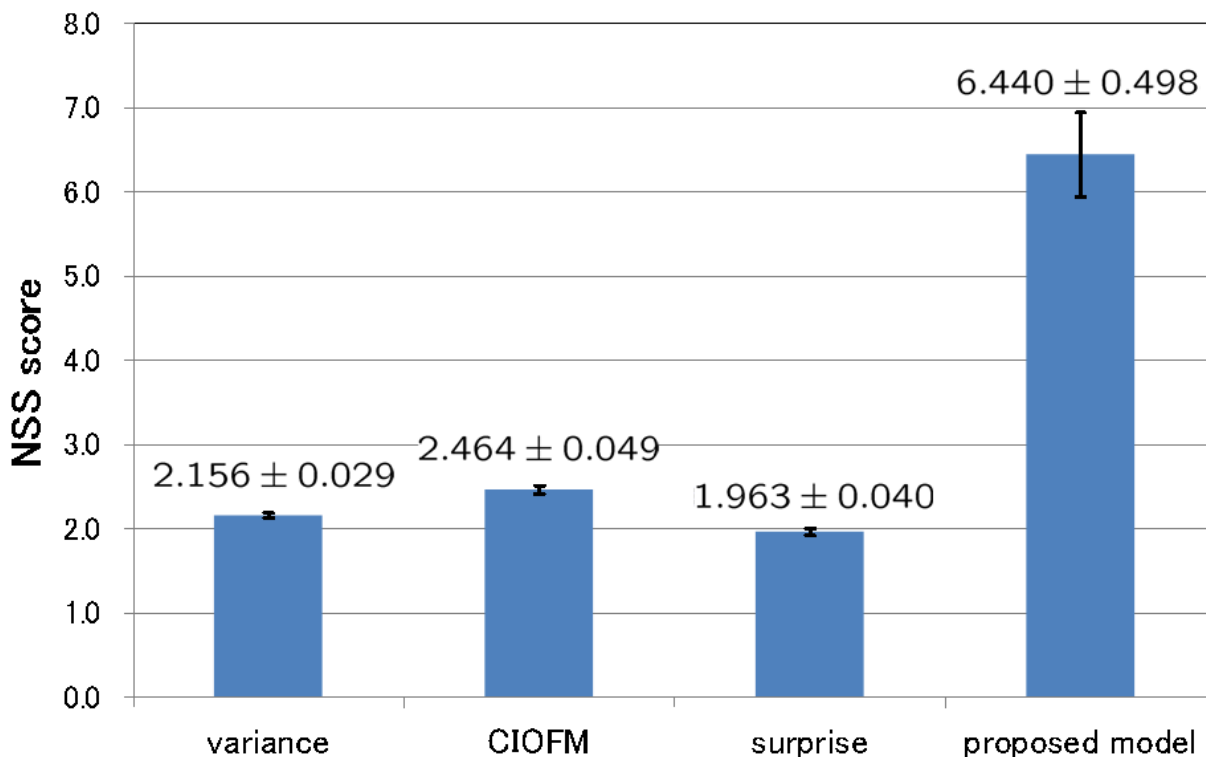
Fig. 10. Average NSS score for all the video clips

salient regions. On the other hand, outputs from our proposed method included only a few small eye focusing areas. This implies that our new method picked up probable eye focusing areas accurately.

Fig. 13 shows the total execution time of 1) calculating Itti's saliency map without CUDA, 2) the proposed method without CUDA, and 3) the proposed method with CUDA. The result indicates that the proposed method has achieved near real-time estimation (40-50 msec/frame), and almost the same processing time as the one for Itti's model.

6 CONCLUSION

We have presented the first stochastic model of human visual attention based on a dynamic Bayesian framework. Unlike many existing methods, we predict the likelihood of human-attended regions on a video based on two criteria: 1) The probability of having the maximum saliency response at a given region evaluated based on the signal detection theory, and 2) the probability of matching the eye movement projection based on the predicted state. Experiments have revealed that our model offers a better eye-gazing prediction against previous deterministic models. To enhance our current model, future work may include determination of initial parameters close to the global optima when estimating model parameters, unified approach to estimate all the model parameters, a better density model of eye movements, a
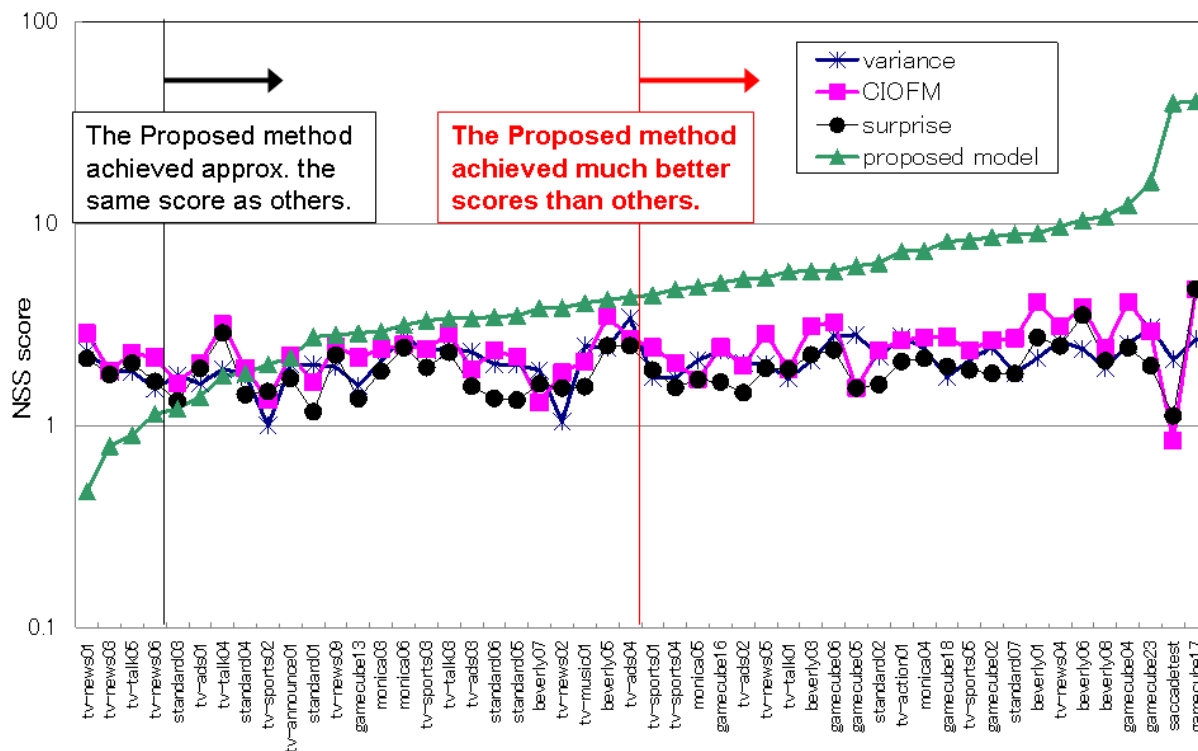
Fig. 11.   Average NSS score for each video clip, where the vertical axis uses a log scale and video clips on the horizontal axis are sorted in the ascending order of NSS scores.

better integration of the bottom-up and the top-down information, a better saliency model for extracting (deterministic) saliency maps, and integration of the proposed method into some applications such as driving assistance, active vision and video retrieval.

ACKNOWLEDGMENT

Fig. 12. Snapshots of model outputs

REFERENCES

[1] M. Hikita, S. Fuke, M. Ogino, T. Minato, and M. Asada, "Visual attention by saliency leads cross-modal body representation," in *Development and Learning, 2008. ICDL 2008. 7th IEEE International Conference on*, Aug. 2008, pp. 157–162.

[2] X.-P. Hu, L. Dempere-Marco, and G.-Z. Yang, "Hot spot detection based on feature space representation of visual search," *Medical Imaging, IEEE Transactions on*, vol. 22, no. 9, pp. 1152–1162, Sept. 2003.

[3] P. Santana, M. Guedes, L. Correia, and J. Barata, "Saliency-based obstacle detection and ground-plane estimation for off-road vehicles," in *ICVS*, 2009, pp. 275–284.

[4] G. Boccignone, "Nonparametric bayesian attentive video analysis," in *Proc. International Conference on Pattern Recognition (ICPR)*, Dec. 2008, pp. 1–4.

[5] G. Boccignone, A. Chianese, V. Moscato, and A. Picariello, "Foveated shot detection for video segmentation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 3, pp. 365–377, March 2005.

[6] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *Proc. International Conference on Multimedia and Expo (ICME)*, June 2009.

[7] O. Chum, J. Philbin, M. Isard, and A. Zisserman, "Scalable near identical image and shot detection," in *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*.   New York, NY, USA: ACM, 2007, pp. 549–556.

[8] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[9] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.
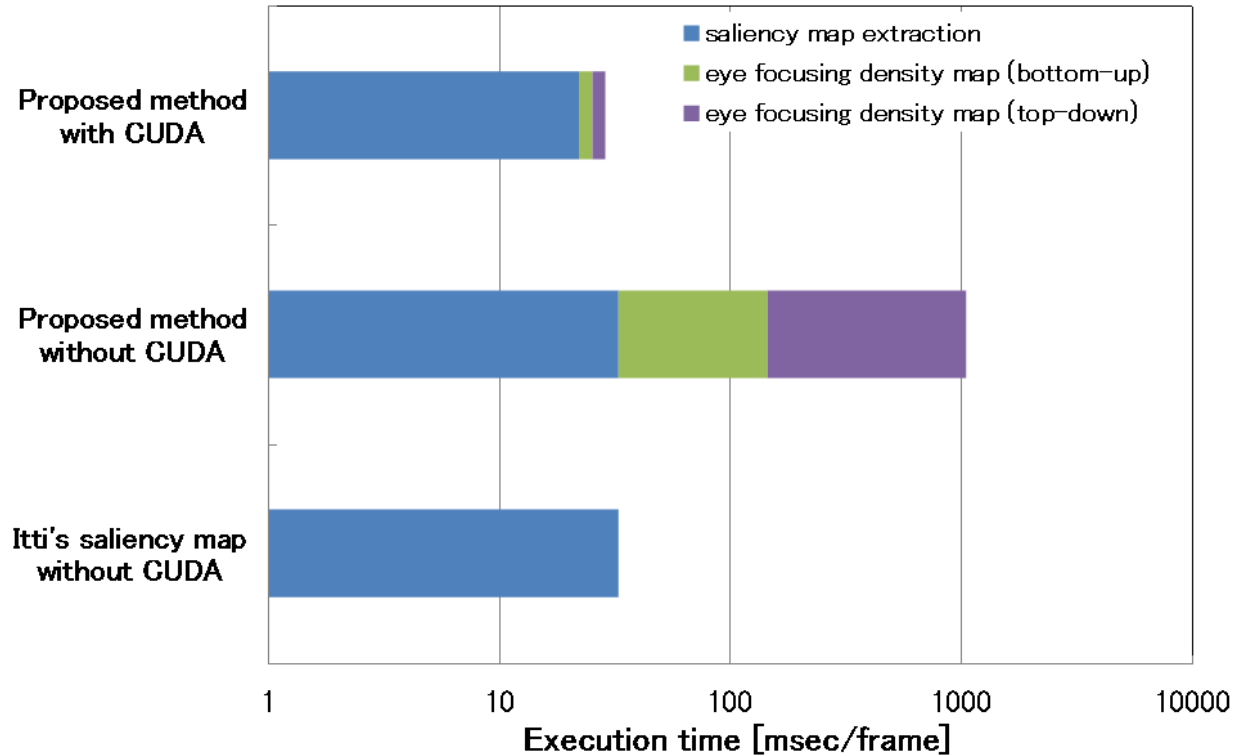
Fig. 13. Total execution time [msec/frame] (log-scaled)

[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[11] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, 2000.

[12] E. Gu, J. Wang, and N. Badler, "Generating sequence of eye fixations using decision-theoretic attention model," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2005, pp. 92–99.

[13] S. Frintrop, *Vocus: a Visual Attention System for Object Detection And Goal-directed Search (Lecture Notes in Computer Science)*. Springer-Verlag New York Inc (C), 3 2006. [Online]. Available: http://amazon.co.jp/o/ASIN/3540327592/

[14] S. Jeong, S. Ban, and M. Lee, "Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment," *Neural Networks*, vol. 21, pp. 1420–1430, October 2008.

[15] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: Computational principles, biological plausibility, and implications for neurophysiology and psychophysics," *Neural Computation*, vol. 21, no. 1, pp. 239–271, January 2009.

[16] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2005, pp. 631–637.

[17] 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)*, July 2007, pp. 300–303.

[18] S. Ban, I. Lee, and M. Lee, "Dynamic visual selective attention model," *Neurocomputing*, vol. 71, pp. 853–856, March 2007.

[19] 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*, vol. 62, pp. 425–451, 2000.

[20] P. Verghese, "Visual search and attention: A signal detection theory approach," *Neuron*, vol. 31, pp. 525–535, August 2001.

[21] B. J. Scholl, Ed., *Objects and Attention (Cognition Special Issue)*. The MIT Press, 8 2002. [Online]. Available: http://amazon.co.jp/o/ASIN/0262692805/

[22] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 693–708, 2009.

[23] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 171–177, 2009.

[24] 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)*, June 2007, pp. 1–8.

[25] K. Myers and B. Taplay, "Adaptive sequential estimation with unknown noise statistics," *IEEE Trans. Autom. Control*, vol. 21, no. 4, pp. 520–523, August 1976.

[26] J. Leathrum, "On sequential estimation of state noise variances," *IEEE Trans. Autom. Control*, vol. 26, no. 3, pp. 745–746, June 1981.

[27] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, vol. 21, pp. 1087–1092, 1953.

[28] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: Particle filters for tracking applications*. Boston: Artech House Publishers, 2004.

[29] A. Viterbi, "Error bounds for comvolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, April 1967.

[30] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

Akisato Kimura (M'00-SM'07) received B.E., M.E. and D.E. degrees in Communications and Integrated Systems from Tokyo Institute of Technology, Japan in 1998, 2000 and 2007, respectively. Since 2000, he has been with NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, where he is currently a Research Scientist in Media Information Laboratory. He has been working on content-based multimedia copy detection, image/video retrieval, automatic image/video annotation, computational models of human visual attention and visual scene understanding. His research interests include pattern recognition, computer vision, image/video processing, human visual perception, cognitive science, statistical signal processing, machine learning and information theory.

Derek Pang (S'04) received a B.A.Sc. degree from Simon Fraser University, Canada in 2009. Since September 2009, he has been with Department of Electrical Engineering, Stanford University, where he is a Ph.D student in the Image, Video and Multimedia Systems (IVMS) group. From June to December 2007, he stayed at Media Information Laboratory, NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation as a research intern. He received the Student Travel Award in IEEE International Conference on Multimedia and Expo (ICME2009). His research interests are in the areas of computer vision, machine learning and multimedia processing and coding.

Tatsuto Takeuchi received B.A. and M.A. degrees from Kyoto University, Japan and the Ph.D. degree in experimental psychology from the University of Tokyo, Japan in 1994. He performed postdoctoral research at the University of California at Berkeley. In 1995, he joined NTT Basic Research Laboratories (now NTT Communication Science Laboratories), Nippon Telegraph and Telephone Corporation. He is currently working on basic and applied research concerning visual motion and color perception. He is a member of the Association for Research in Vision and Ophthalmology, the Japanese Psychological Association and the Vision Society of Japan.
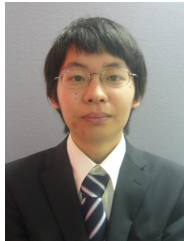
Kouji Miyazato has been with Department of Information and Communication Systems Engineering, Okinawa National College of Technology, Japan, where he is a bachelor student. In September 2008, he stayed at Media Information Laboratory, NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation as a research intern. During his internship, he was working on research of stream processing with GPUs. He received the Outstanding Performance Award in the 6th Japanese Olympiad in Informatics (JOI 2006/2007).

Kunio Kashino (S'89-M'95-SM'05) received a Ph.D. degree in electrical engineering from the University of Tokyo in 1995. In 1995, he joined NTT Corporation, where he is currently a Senior Research Scientist, Supervisor. He has been working on multimedia information retrieval. His research interests include acoustic signal processing and Bayesian information integration.
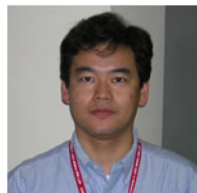
Junji Yamato (M'90-SM'05) received B.E. and M.E. degrees from the University of Tokyo, Japan, in 1988 and 1990, respectively. He received an M.S. degree from the MIT in 1998, and a Ph.D. degree from the University of Tokyo in 2001. He is currently the Executive Manager of Media Information Laboratory, NTT Communication Science Laboratories. His research interests include computer vision, gesture recognition and human-robot interaction.