

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

Derek PANG^{†,††}, Akisato KIMURA^{††}, Tatsuto TAKEUCHI^{††},
Junji YAMATO^{††}, and Kunio KASHINO^{††}

^{††} NTT Communication Science Laboratories, NTT Corporation,
3-1 Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198 Japan.

[†] School of Engineering Science, Simon Fraser University,
8888 University Drive, Burnaby, British Columbia, V5A 1S6 Canada.

Abstract Recent studies in signal detection theory 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. To predict the likelihood of where humans typically focus on a video scene, we propose a new stochastic model of visual attention by introducing a dynamic Bayesian network. When describing the network of visual attention, the principle of the signal detection theory is introduced, namely, the position where a saliency response takes the maximum is the eye focusing position. Experimental results have demonstrated that our model performs significantly better in predicting human visual attention compared to the previous deterministic model.

Key words Visual attention, saliency, dynamic Bayesian network, state space model, hidden Markov model.

1. Introduction

Developing an accurate computational model of human visual attention has been a long-standing challenge. Such model may allow any system to select just relevant information from a complex and cluttered visual input in numerous artificial vision applications, such as robotics [1], [2], active vision [3], multimedia recognition [4] and retrieval [5].

The first biologically plausible model for explaining the human visual system was proposed by Koch and Ullman [6], and later implemented by Itti et al [7]. Many attempts [8] ~ [10] have been made to improve the Koch-Ullman model. The basic concept underlying all the above researches is the *feature integration theory* [11] which has been one of the most influential models 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. Although the feature integration theory well models the early human visual system, it suffers from a crucial problem in that the saliency responses are assumed to be deterministic. This conflicts our intuition that people may attend to different locations on the same visual input at the same time. A typical example can be seen in Fig. 1. Let us consider a search task with a single 45° target among a lot of distractors. We can

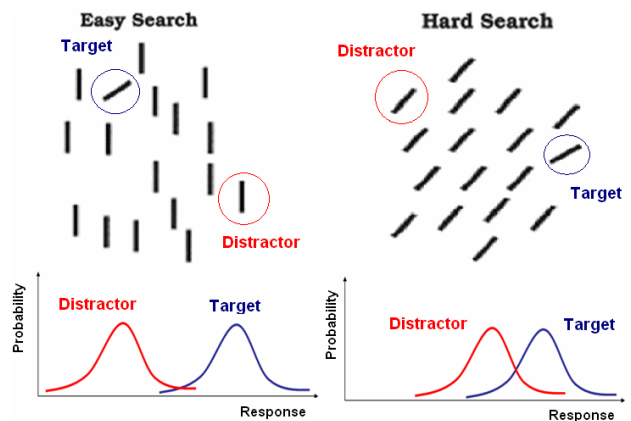


Fig. 1 Visual response based on the signal detection theory.

easily understand from the figure that the single target on the left is easy to find unlike that on the right. The previous models could not explain the above intuition since the models assume that every time we first select the location where the response of the detector tuned to the visual property of the target is greater than at any other locations.

Recently, another psychophysical theory to understand visual attention has been presented called the *signal detection theory* [12]. According to the signal detection theory, the elements in a visual display are internally represented as independent, noisy random variables. Again let us consider the search task shown in Fig. 1. The response of a detector

tuned to the target orientation is a kind of 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 right case, the target density is identical to the left one, 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 a longer time to detect the target.

Based on the paradigm of the signal detection theory, this report 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 state-space model, 3) an *eye movement pattern* that predicts the human viewing patterns using a hidden Markov model (HMM), and 4) an *eye position density map* that estimates the probable human-attended regions. 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 incorporates another property that eye movements may be affected also by the previous eye focusing positions.

Several previous researches that focused on stochastic modeling of visual attention has been studied. Itti and Baldi [9] investigated a Bayesian approach to detecting surprising events in video signals. Koike and Saiki [13] introduced a stochastic winner-take-all (WTA) mechanism into the Koch-Ullman model. Other researches that formulate a stochastic model based on a certain aspect of the attention system has been developed by Rimey and Brown [14]. Our main contributions against the previous researches include the introduction of a unified stochastic model that integrates the bottom-up information (i.e. saliency) with top-down information (i.e. eye movement pattern) by using a dynamic Bayesian network.

2. Stochastic visual attention model

2.1 Overview

Fig. 2 illustrates the graphical model 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.

$I = i(1 : T) = \{i(t)\}_{t=1}^T$ 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 . Also the

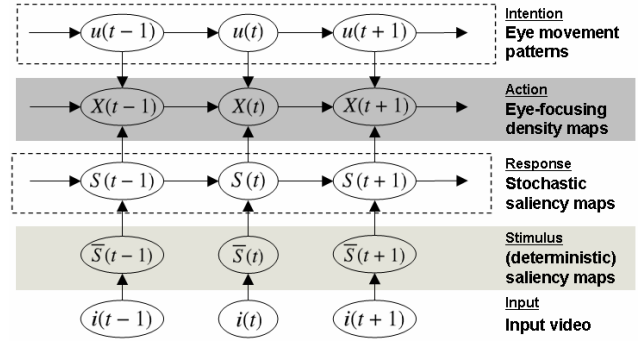


Fig. 2 Graphical representation of the proposed stochastic model of human visual attention.

symbol I denotes a sequence of frames $i(t)$ as well as a set of coordinates in the frame. For example, we denote a position \mathbf{y} in a frame as $\mathbf{y} \in I$.

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

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

$U(I) = u(1 : T; I) = \{u(t; I)\}_{t=1}^T$ denotes a sequence of *eye movement patterns* each of which represents a pattern of eye movements. Two typical patterns of eye movements are found when one is watching a video: 1) Passive state $u(t; I) = 0$ in which one tends to stay around one particular position to continuously capture important visual information, and 2) active state $u(t; I) = 1$ in which one actively moves around and searches various visual information on the scene. Eye movement patterns reflect purposes or intentions of human eye movements.

$X(I) = X(1 : T; I) = \{x(t; I)\}_{t=1}^T$ denotes a sequence of eye focusing positions each of which depends on the input video I . The eye focusing position is determined by integrating the bottom-up information (stochastic saliency maps) and the top-down information (eye movement patterns).

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

2.2 Saliency model implementation

We used Itti-Koch saliency model [7] 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). 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.

2.3 Stochastic saliency maps

We introduce the following state space model to estimate stochastic saliency maps from (deterministic) saliency maps.

$$\begin{aligned} s(t, \mathbf{y}) &= \bar{s}(t, \mathbf{y}) + \eta_{s1} \Leftrightarrow \bar{s}(t, \mathbf{y}) = s(t, \mathbf{y}) + \eta_{s1}, \quad (1) \\ s(t, \mathbf{y}) &= s(t-1, \mathbf{y}) + \eta_{s2}, \quad (2) \end{aligned}$$

where η_{si} ($i = 1, 2$) is a Gaussian random variable with mean 0 and variance σ_{si}^2 . Eq. (1) implies that a saliency map is observed through a Gaussian random process. Eq. (2) exploits the temporal characteristics of the human visual system.

The model represented by Eq. (1) (2) can be rewritten into the following stochastic relationships:

$$\begin{aligned} p(\bar{s}(t)|s(t)) &= \mathcal{G}(\bar{s}(t); s(t), \sigma_{s1}), \\ p(s(t)|s(t-1)) &= \mathcal{G}(s(t); s(t-1), \sigma_{s2}), \end{aligned}$$

where \mathbf{y} is omitted for simplicity and $\mathcal{G}(s; \bar{s}, \sigma)$ is the Gaussian density with argument s , mean \bar{s} and variance σ^2 .

$$\mathcal{G}(s; \bar{s}, \sigma) \stackrel{\text{def.}}{=} \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(s - \bar{s})^2}{2\sigma^2} \right\}.$$

To estimate the response of the stochastic saliency map $s(t)$, we apply Kalman filter (e.g. [15]). Suppose that the density $p(s(t-1)|\bar{s}(1:t-1))$ of the previous stochastic saliency map is given by the following Gaussian density:

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

Then, the density $p(s(t)|\bar{s}(1:t))$ of the current stochastic saliency map is obtained as follows:

$$\begin{aligned} p(s(t)|\bar{s}(1:t)) &= \mathcal{G}(s(t); \hat{s}(t|t), \sigma_s(t|t)), \\ \hat{s}(t|t) &= \frac{\sigma_s^2(t|t)}{\sigma_s^2(t|t-1) + \sigma_s^2(t|t)} \hat{s}(t|t-1) + \frac{\sigma_s^2(t|t)}{\sigma_s^2(t|t)} \bar{s}(t), \quad (3) \end{aligned}$$

$$\sigma_s^2(t|t) = \frac{\sigma_{s1}^2 \cdot \sigma_s^2(t|t-1)}{\sigma_{s1}^2 + \sigma_s^2(t|t-1)}, \quad (4)$$

$$\begin{aligned} \hat{s}(t|t-1) &= \hat{s}(t-1|t-1), \\ \sigma_s^2(t|t-1) &= \sigma_{s2}^2 + \sigma_s^2(t-1|t-1). \end{aligned}$$

2.4 Estimating eye motions

By incorporating the stochastic saliency map $S(t)$ and the *eye movement pattern* $u(t)$, we introduce the following relationships to estimate the *eye focusing position* $\mathbf{x}(t)$:

$$\mathbf{x}(t) = f_1(S(t)), \quad (5)$$

$$\mathbf{x}(t) = f_2(\mathbf{x}(t-1), u(t)), \quad (6)$$

where $f_i(\cdot)$ ($i = 1, 2$) is a stochastic function.

Eq. (5) represents that the eye focusing position is selected

from a stochastic process based on the signal detection theory. The signal detection theory implies that the position in which the stochastic saliency takes the maximum is determined as the eye focusing position. Here, let us define $p(S(t))$ as the density of the stochastic saliency map $S(t)$ at time t and $P(s(t, \tilde{\mathbf{y}}) \leq s)$ as the distribution function (i.e. the cumulative density) that corresponds to the density $p(s(t, \tilde{\mathbf{y}}))$ of the stochastic saliency $s(t, \tilde{\mathbf{y}})$

$$\begin{aligned} p(S(t)) &\stackrel{\text{def.}}{=} \{p(s(t, \mathbf{y}))\}_{\mathbf{y} \in I}, \\ p(s(t, \mathbf{y})) &\stackrel{\text{def.}}{=} p(s(t, \mathbf{y})|\bar{s}(1:t, \mathbf{y})) \quad \forall \mathbf{y} \in I, \\ P(s(t, \tilde{\mathbf{y}}) \leq s) &\stackrel{\text{def.}}{=} \int_{-\infty}^s p(s(t, \tilde{\mathbf{y}}) = s') ds'. \end{aligned}$$

Based on the above definitions, Eq. (5) can be rewritten by the following stochastic relationships:

$$\begin{aligned} p(\mathbf{x}(t)|p(S(t))) \\ = \int_{-\infty}^{\infty} p(s(t, \mathbf{x}(t)) = s) \prod_{\tilde{\mathbf{x}} \neq \mathbf{x}(t)} P(s(t, \tilde{\mathbf{x}}) \leq s) ds, \quad (7) \end{aligned}$$

where the first term indicates the density so that the stochastic saliency at the position $\mathbf{x}(t)$ equals s and the second term represents the probability so that the stochastic saliency values at all other positions are smaller than s . Integration computations in Eq. (7) is intractable. Therefore in the implementation, we take into account only at the positions where the average stochastic saliency takes local maxima.

The second relationship (Eq. (6)) 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 $u(t)$. We can estimate eye focusing positions using an HMM, in which the hidden states are the eye movement patterns. 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 the eye focusing position being observed is governed by the following emission probability:

$$\begin{aligned} p(\mathbf{x}(t)|\mathbf{x}(t-1), u(t)) \\ = \prod_{i=0}^1 \{\mathcal{L}(\mathbf{x}(t); \mathbf{x}(t-1), \gamma_{xi}, \sigma_{xi})\}^{u(t)_i}, \quad (8) \end{aligned}$$

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 such that

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

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

Combining the above stochastic relationships, we can define the following relationship:

$$\begin{aligned} p(\mathbf{x}(t), u(t)|p(S(t)), \mathbf{x}(t-1), u(t-1)) \\ \stackrel{\text{def.}}{=} \frac{1}{Z} p(\mathbf{x}(t)|p(S(t))) \\ \cdot p(u(t)|u(t-1)) \cdot p(\mathbf{x}(t)|\mathbf{x}(t-1), u(t)), \quad (9) \end{aligned}$$

where Z is a normalizing constant. Since it is simply impractical to calculate Eq. (9) for each combination of variables, we utilize Monte-Carlo sampling. 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 Eq. (9), where N is the number of samples. The empirical distribution of samples $\tilde{X}(t)$ can then be represented as the eye position density map $X(t)$.

3. Parameter estimation

This section focuses on the problem of estimating maximum likelihood (ML) model parameters. We can utilize saliency maps calculated from the input video and eye focusing positions obtained by eye tracking devices (e.g. [16]) as observations. Simultaneous estimation of all ML parameters can be optimal but impractical due to the substantial calculation cost. Therefore, we separate parameter estimation into two stages. The first stage derives parameters for computing stochastic saliency maps, and the second stage for estimating eye focusing points.

3.1 Parameters for stochastic saliency maps

The first stage derives parameters $\theta_s = (\sigma_{s1}, \sigma_{s2})$ for computing stochastic saliency maps. Here, we introduce the EM algorithm (see e.g. [17]). In this case, the observations are the saliency maps $\bar{S} = \bar{S}(1:T)$ and the hidden variables are the stochastic saliency maps $S = S(1:T)$. The EM algorithm for estimating θ_s is as follows:

- The objective function:

We introduce the following objective functions to maximize the joint density $p(\bar{S}, S; \theta_s)$:

$$\begin{aligned} F_s(q(S), \theta_s) &\stackrel{\text{def.}}{=} \int_S q(S) \log p(\bar{S}, S; \theta_s) dS - \int_S q(S) \log q(S) dS \\ &= -p(\bar{S}; \theta_s) D(q(S) \| p(S|\bar{S}; \theta_s)), \end{aligned}$$

where $q(S)$ is a dummy density of S , $p(\cdot; \theta_s)$ is a density with parameter θ_s and $D(q||p)$ is the Kullback-Leibler divergence between densities q and p . We recursively update the dummy density $q(S)$ and parameters θ_s so as to maximize the objective function $F_s(q(S), \theta_s)$.

- $(k+1)$ -th E step:

The E step updates the dummy density $q(S)$ to maximize the objective function $F_s(q(S), \theta_{s,k})$ with the parameter $\theta_{s,k} \stackrel{\text{def.}}{=} (\sigma_{s1,k}, \sigma_{s2,k})$ updated at the previous step.

$$\begin{aligned} Q_{k+1}(S) &\leftarrow \arg \max_{q(S)} F_s(q(S), \theta_{s,k}) \\ &= p(S|\bar{S}; \theta_{s,k}), \end{aligned}$$

The density $p(S|\bar{S}; \theta_{s,k})$ can be calculated by Kalman smoother. Suppose that the density $p(s(t+1)|\bar{S}; \theta_{s,k})$ of the stochastic saliency $s(t+1)$ at time $t+1$ is given by the following Gaussian density:

$$\begin{aligned} p(s(t+1)|\bar{S}; \theta_{s,k}) &= \mathcal{G}(s(t+1); \hat{s}_k(t+1|T), \sigma_{s,k}(t+1|T)). \end{aligned}$$

Then, the density $p(s(t)|\bar{S}; \theta_{s,k})$ of the stochastic saliency $s(t)$ at time t is obtained as follows:

$$\begin{aligned} p(s(t)|\bar{S}; \theta_{s,k}) &= \mathcal{G}(s(t); \hat{s}_k(t|T), \sigma_{s,k}(t|T)), \\ \hat{s}_k(t|T) &= \frac{\sigma_{sq,k}^2(t|t)}{\sigma_{s,k}^2(t|t)} \hat{s}_k(t|t) + \frac{\sigma_{sq,k}^2(t|t)}{\sigma_{s2,k}^2} \hat{s}_k(t+1|T), \\ \sigma_{s,k}^2(t|T) &= \sigma_{sq,k}^2(t|t) + \left(\frac{\sigma_{sq,k}^2(t|t)}{\sigma_{s2,k}^2} \right)^2 \sigma_{s,k}^2(t+1|T), \\ \sigma_{sq,k}^2(t|t) &= \frac{\sigma_{s2,k}^2 \sigma_{s,k}^2(t|t)}{\sigma_{s2,k}^2 + \sigma_{s,k}^2(t|t)}. \end{aligned}$$

$\hat{s}_k(t|t)$ and $\sigma_{s,k}^2(t|t)$ can be obtained by Kalman filter with the parameter $\theta_{s,k}$ (see Eqs. (3) (5)).

- $(k+1)$ -th M step:

The M step updates the parameter θ_s to maximize the objective function $F_s(Q_{k+1}(S), \theta_s)$.

$$\begin{aligned} \theta_{s,k+1} &\leftarrow \arg \max_{\theta_s} F_s(Q_{k+1}(S), \theta_s) \\ &= \arg \max_{\theta_s} \langle \log p(\bar{S}, S; \theta_s) \rangle_k, \end{aligned}$$

$$\langle f(S) \rangle_k \stackrel{\text{def.}}{=} \int_S f(S) p(S|\bar{S}, \theta_{s,k}) dS.$$

Taking derivatives of the log density in terms of θ_s , we obtain

$$\begin{aligned} \sigma_{s1,k+1}^2 &\leftarrow \frac{1}{T} \sum_{t=1}^T \langle \{\bar{s}(t) - s(t)\}^2 \rangle_k \\ &= \frac{1}{T} \sum_{t=1}^T \left\{ \langle \{\bar{s}(t) - \hat{s}_k(t|T)\}^2 + \sigma_{s,k}^2(t|T) \rangle, \right. \\ \sigma_{s2,k+1}^2 &\leftarrow \frac{1}{T-1} \sum_{t=2}^T \langle \{s(t) - s(t-1)\}^2 \rangle_k \\ &= \frac{1}{T-1} \sum_{t=2}^T \left[\langle \{\hat{s}_k(t|T) - \hat{s}_k(t-1|T)\}^2 \right. \\ &\quad \left. + \sigma_{s,k}^2(t-1|T) + \frac{\sigma_{s2,k}^2 - \sigma_{s,k}^2(t-1|t-1)}{\sigma_{s2,k}^2 + \sigma_{s,k}^2(t-1|t-1)} \sigma_{s,k}^2(t|T) \right] \end{aligned}$$

3.2 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 eye focusing positions \mathbf{X} obtained from the eye tracking data, and the hidden states are the eye movement patterns U . Instead of the EM algorithm, we introduce the Viterbi learning method. This approach recursively updates the eye movement patterns $U = u(1:T)$ and the ML parameter set θ_x to maximize the posterior $p(U|\mathbf{X}; \theta_x)$.

- Initializing eye movement patterns:

We have to start with determining an initial sequence $U_0 = u_0(1:T)$ of eye movement patterns. In this report, the following decision rule is introduced:

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

where κ_x is a given threshold. This means that eye movement patterns can be strongly correlated with the degree of eye movements.

- The $(k + 1)$ -th step for updating hidden variables:

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

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

Viterbi algorithm (e.g. [18], [19]) can derive the ML sequence U_{k+1} of eye movement patterns.

- The $(k + 1)$ -th step for updating the parameter set:

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

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

Taking derivatives of the log density in terms of θ_x , we derive the following equations, where the function $\delta(a, b) = 1 - |a - b|$ shows whether a equals b :

$$\begin{aligned} \gamma_{xi,k+1} &\leftarrow \frac{\sum_{t=2}^T \|\mathbf{x}(t) - \mathbf{x}(t-1)\| \cdot \delta(i, u_{k+1}(t))}{\sum_{t=2}^T \delta(u_{k+1}(t), i)}, \\ \sigma_{xi,k+1}^2 &\leftarrow \frac{\sum_{t=2}^T (\|\mathbf{x}(t) - \mathbf{x}(t-1)\| - \gamma_{xi,k+1})^2 \cdot \delta(u_{k+1}(t), i)}{2 \sum_{t=2}^T \delta(u_{k+1}(t), i)}, \\ \phi_{(i,j),k+1} &\leftarrow \frac{\sum_{t=2}^T \delta(u_{k+1}(t), i) \cdot \delta(u_{k+1}(t-1), j)}{\sum_{t=2}^T \delta(u_{k+1}(t-1), j)}. \end{aligned}$$

4. Evaluation

4.1 Collecting eye tracking data

For the purpose of parameter training and model evaluation, we collected samples of eye focusing positions from six human subjects under a protocol approved by the Institutional Review Board of NTT Communication Science Laboratories. 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. The total length of each video varies from 30 to 90 seconds, and the size of video clips was 640 x 480 pixels. Each subject's right eye position was recorded at 30 Hz with an eye tracking device [16] 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 the actual human eye focusing positions, we used the normalized scan-path saliency (NSS) [20]. 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 test subjects, $\bar{p}(\mathbf{x})$ and $\sigma(p(\mathbf{x}))$ are the mean and the variance of the model's output density map respectively. An NSS value of unity indicates the subjects' eye positions fall on a region whose predicted

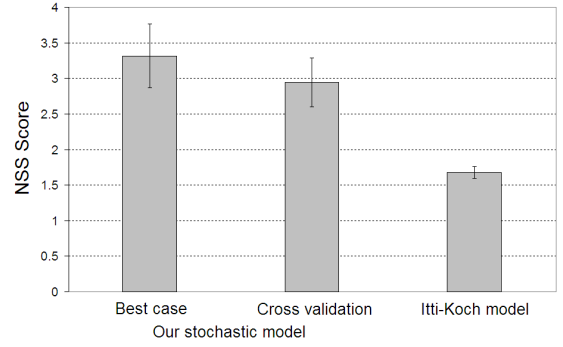


Fig. 3 Evaluation result of the proposed stochastic visual attention model and Itti-Koch model. The Y error bar indicates the standard error.

density is one standard deviation above average. Meanwhile, an NSS value of zero or lower means that the model performs no better than picking a random position on the map.

4.3 Results

We evaluated the performance of our basic model by comparing it against Itti-Koch model [7]. Two different data training scenarios were conducted to show the model dependency on the training data set: best case scenario and 3-fold cross validation scenario. In the best case scenario, the parameter of each video is trained by its own set of eye focusing data. In the cross-validation scenario, the video clips are divided into three data sets. Only one data set was retained for evaluation each time with the remaining sets being the training data.

Fig. 3 shows the average NSS scores of all video clips for the Itti-Koch model and our basic model trained with the two different scenarios for parameter estimation. The average NSS result indicates that our model trained from either scenario in each viewing situation performs substantially better than Itti-Koch model by more than 75%. Not just in the average performance but also in the result of every video clip from every experiment type, the proposed stochastic model have outperformed the Itti-Koch model. The result in the cross validation scenario verified that our purposed model still performed significantly well independent of which training set is utilized.

Fig. 4 shows samples of eye focusing density maps estimated from our basic model, and Fig. 5 shows an example of NSS score distributions of the Itti-Koch model and our proposed model for a particular input video. Each figure indicate that our proposed model well estimated a tendency of eye movements compared with the Itti-Koch model.

5. 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,

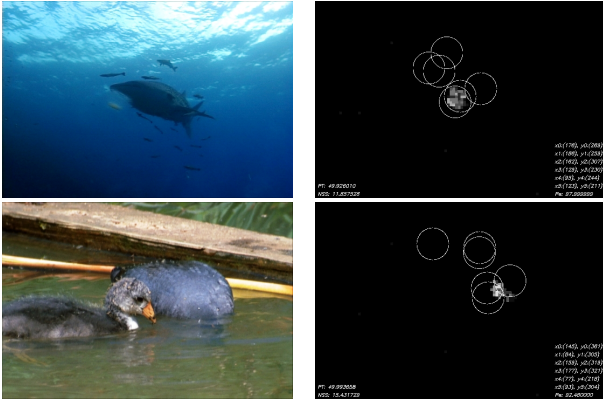


Fig. 4 Samples of results. (Left) Input videos. (Right) Eye focusing density maps, where each white circle corresponds to the eye focusing position of a subject.

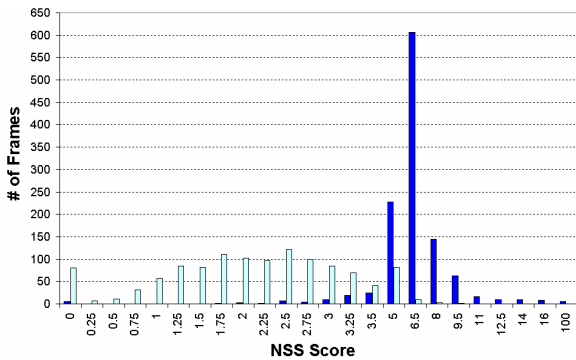


Fig. 5 NSS score distribution, where dark bars (resp. light bars) correspond to the our proposed model (resp. the Itti-Koch model)

and 2) the probability of matching the eye movement projection based on the predicted cognitive state. Experiments have revealed that our model offers a better eye-gazing prediction against previous deterministic model. To enhance our current model, future work may include introduction of spatial relationships of stochastic saliency maps, a better integration of the bottom-up and the top-down information, real-time computing with general-purpose GPU (GP-GPU) computation, and integration of the proposed method into some applications.

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. Hiromi Nakaiwa, Dr. Shoji Makino and Dr. Kenji Nakazawa of NTT Communication Science Laboratories for their help to the internship.

Bibliography

[1] N. Ouerhani and H. Hügli, "Robot self-localization using visual attention," Proc. International Symposium on Computational Intelligence in Robotics and Automation (CIRA),

pp.309–314, June 2005.

[2] Y. Nagai and K. Rohlfling, "Can motionese tell infants and robots: What to imitate?," Proc. International Symposium on Imitation in Animals and Artifacts (AISB), pp.299–306, April 2007.

[3] Y. Takeuchi, N. Ohnishi, and N. Sugie, "Active vision fixating a saliency in a scene: Information theoretical approach," IEICE Technical Report, Vol.96, pp.31–38, February 1997.

[4] A. Salah, E. Alpaydin, and L. Akarun, "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition," IEEE Trans. Pattern Anal. Mach. Intell., Vol.24, No.3, pp.420–425, March 2002.

[5] S. Li and M. Lee, "An efficient spatiotemporal attention model and its application to shot matching," IEEE Trans. Circuits Syst. Video Technol., Vol.17, No.10, pp.1383–1387, October 2007.

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

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

[8] 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.

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

[10] C. Leung, A. Kimura, T. Takeuchi, and K. Kashino, "A computational model of saliency depletion/recovery phenomena for the salient region extraction of videos," Proc. International Conference on Multimedia and Expo (ICME), pp.300–303, July 2007.

[11] A. Treisman and G. Gelade, "A feature-integration theory of attention," Cognitive Psychology, Vol.12, pp.97–136, 1980.

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

[13] T. Koike and J. Saiki, "Stochastic saliency-based search model for search asymmetry with uncertain targets," Neurocomputing, Vol.69, pp.2112–2126, October 2006.

[14] R. Rimey and C. Brown, "Controlling eye movements with hidden Markov models," International Journal of Computer Vision, Vol.7, pp.47–65, 1991.

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

[16] T. Ohno, N. Mukawa, and A. Yoshikawa, "FreeGaze: A gaze tracking system for everyday gaze interaction," Proc. Symposium on ETRA, pp.125–132, 2002.

[17] Z. Ghahramani, "An introduction to hidden Markov models and Bayesian networks," International Journal of Pattern Recognition and Artificial Intelligence, Vol.15, No.1, pp.9–42, 2001.

[18] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Trans. Inf. Theory, Vol.13, No.2, pp.260–269, April 1967.

[19] 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.

[20] R.J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," Proc. Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, June 2007.