

# A STOCHASTIC MODEL OF SELECTIVE VISUAL ATTENTION WITH A DYNAMIC BAYESIAN NETWORK

Derek Pang<sup>\*†</sup>, Akisato Kimura<sup>†</sup>, Tatsuto Takeuchi<sup>†</sup>, Junji Yamato<sup>†</sup> and Kunio Kashino<sup>†</sup>

<sup>†</sup> NTT Communication Science Laboratories, NTT Corporation, Japan

<sup>\*</sup> School of Engineering Science, Simon Fraser University, 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. Our model simulates and combines the visual saliency response and the cognitive state of a person to estimate the most probable attended regions. Experimental results have demonstrated that our model performs significantly better in predicting human visual attention compared to the previous deterministic model.

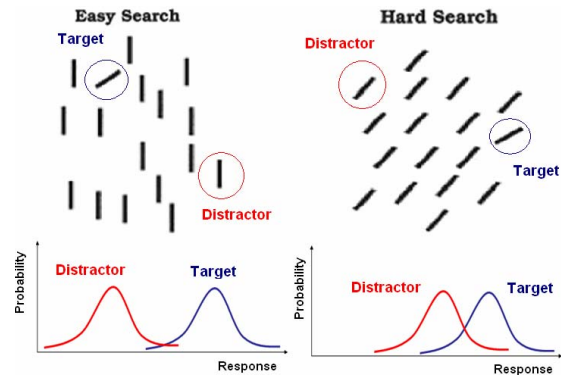
**Index Terms**— Visual attention model, saliency, dynamic Bayesian network, Kalman filter, hidden Markov model

## 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 critical information from a complex and clustered visual input in numerous artificial vision applications, such as robotics, surveillance and multimedia recognition.

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 was suggested by Koch and Ullman in 1985 [1], which follows the latter approach based on the feature integration theory (FIT) developed by Treisman and Gelade [2]. Many attempts [3, 4, 5, 6] have been made to improve the Koch-Ullman model. Although FIT well models the early human visual system, it only selects a *fixed* attended location where some visual properties are maximally different from its neighbors.

Recent studies [7] in the signal detection theory (SDT) offers a different approach to understanding visual processing. Let us consider a visual search task of finding a  $45^\circ$  target on Fig. 1. Based on FIT, we can immediately identify the  $45^\circ$

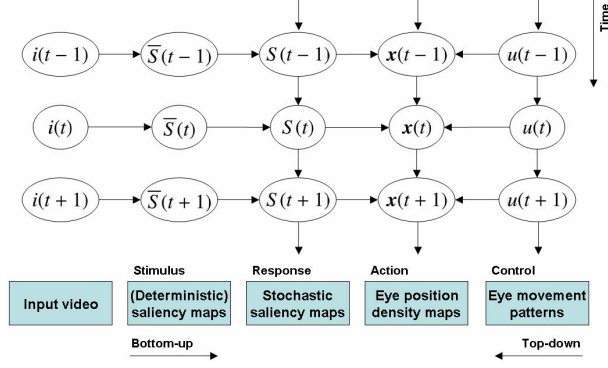


**Fig. 1.** Visual search response based on the signal detection theory (SDT) proposed by Eckstein et al.[7]

target for both easy and hard searches since we always select the location where visual properties are maximally different from its neighbors. Meanwhile, SDT idealizes the response to the visual target and distractors are represented as independent Gaussian random variables. For the hard search case, SDT shows the distribution of the distractor overlaps with the distribution of the target more significantly compared to the easy search case. Therefore, identifying the desired visual target for the hard search case becomes harder as the probability of responding to the distractors becomes higher.

Based on the paradigm of SDT, we propose a new stochastic model of visual attention. With this model can automatically predict the likelihood of where humans typically focus on a visual input. The proposed model comprises a dynamic Bayesian network with four layers: 1) A *saliency map* that shows the average saliency response. 2) A *stochastic saliency map* (SSM) that converts the saliency map into a natural human response through a Kalman filter. 3) An *Eye movement pattern* that predicts the human viewing patterns using a hidden Markov model (HMM). 4) A *eye position density map* that estimates the probable human-attended regions. Unlike previous researches [5, 8] that formulate a stochastic model based on a certain aspect of the attention system, we aim to introduce a unified stochastic model that integrates the bottom-up information (i.e. saliency) with the top-down information (i.e. eye movement pattern).

The first author contributed to this work during his internship at NTT.



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

## 2. STOCHASTIC VISUAL ATTENTION MODEL

### 2.1. Saliency maps

Fig. 2 illustrates the structure of the proposed stochastic visual attention model. Consider an input video  $I = \{i(t)\}_{t=1}^T$ , where  $i(t)$  is the  $t$ -th frame and  $T$  is the total duration of the video  $I$ . Then, the *saliency map* at time  $t$ ,  $\bar{S}(t; I) = \{\bar{s}(t, \mathbf{y}; I)\}_{\mathbf{y}}$ , is obtained from video  $I$  using an existing saliency model [3], where  $\bar{s}(t, \mathbf{y}; I)$  indicates the saliency value at the position  $\mathbf{y}$ . Hereafter, we will occasionally omit the video  $I$ , the position  $\mathbf{y}$  or the time  $t$  for simplicity where explicit expression is unnecessary.

### 2.2. Stochastic saliency maps

The *stochastic saliency map* (SSM)  $S(t) = \{s(t, \mathbf{y})\}_{\mathbf{y}}$  are defined by assuming the following two relationships,

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

where  $\eta_{si}$  ( $i = 1, 2$ ) is a Gaussian random variable with mean 0 and variance  $\sigma_{si}^2$ . The first relationship implies that a saliency map is observed through an idealized Gaussian random process. The second relationship exploits the temporal characteristic of the human visual system. Both relationships contain a separate parameter  $\sigma_{si}$  to characterize the visual noises observed in both processes.

We can rewrite the above relationships into the following stochastic relationships:

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

where  $\mathcal{G}(s; \bar{s}, \sigma)$  is the Gaussian density with argument  $s$ , mean  $\bar{s}$ , and variance  $\sigma^2$ . To estimate the response of SSM, we can apply the Kalman filter to recursively compute the mean  $\hat{s}(t|t)$  and variance  $\sigma_s^2(t|t)$  of the stochastic saliency response  $p(s(t)|\{\bar{s}(i)\}_{i=1}^t)$  given saliency maps  $\{\bar{s}(i)\}_{i=1}^t$  for each pixel.

### 2.3. Estimating eye positions

By incorporating the stochastic saliency map  $S(t) = \{s(t, \mathbf{y})\}_{\mathbf{y}}$  and the *eye movement pattern*  $u(t)$ , we can estimate the *eye focusing position*  $\mathbf{x}(t)$  such that

$$\mathbf{x}(t) = f_1(S(t)), \quad (1)$$

$$\mathbf{x}(t) = f_2(\mathbf{x}(t-1), u(t)), \quad (2)$$

where  $f_i(\cdot)$  is a stochastic function. In Eq. (1), the eye position is selected from a stochastic process based on SDT. The probability that an eye position  $\mathbf{x} = \mathbf{x}(t)$  has the maximum saliency response is determined by

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

where

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

Eq. (2) suggests the current eye focusing position depends on the previous position. The degree of eye movements is driven by one's eye movement pattern  $u(t)$ . Two typical eye movement patterns [9] are found when one is watching a video: 1) Passive state, in which one tends to stay around one particular position to continuously capture important visual information, and 2) active state, in which one actively moves around and searches for various visual information on the scene. We can estimate eye positions using a HMM, in which the hidden states are the eye movement patterns. Let  $u(t) = (u(t)_0, u(t)_1)^T$  be a 2-dimensional binary vector such that  $(1, 0)^T$  denotes the passive state, and  $(0, 1)^T$  represents the active state, where the superscript  $T$  of a vector stands for the transposition. Then the transitional probability of going from the hidden state  $j$  to hidden state  $i$  is defined as

$$p(u(t)|u(t-1)) = \prod_{i=0}^1 \prod_{j=0}^1 \{\phi_{(i,j)}\}^{u(t)_i u(t-1)_j}$$

where  $\Phi = \{\phi_{(i,j)}\}_{(i,j)}$ . Given the hidden state  $u(t)$ , the probability of the eye focusing position being observed is governed by the emission probability

$$\begin{aligned} p(\mathbf{x}(t)|\mathbf{x}(t-1), u(t)) &= \{\mathcal{L}(\mathbf{x}(t); \mathbf{x}(t-1), \gamma_{x0}, \sigma_{x0})\}^{u(t)_0} \\ &\quad \cdot \{\mathcal{L}(\mathbf{x}(t); \mathbf{x}(t-1), \gamma_{x1}, \sigma_{x1})\}^{u(t)_1}, \end{aligned}$$

where  $\mathcal{L}(\mathbf{x}; \bar{\mathbf{x}}, \gamma, \sigma)$  is a shifted Gaussian density with argument  $\mathbf{x}$ , average  $\bar{\mathbf{x}}$ , indent  $\gamma$ , and variance  $\sigma^2$  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 Eq. (1) and (2), we can define the following probabilistic relationship

$$\begin{aligned} p(\mathbf{x}(t), u(t)|p(S(t)), \mathbf{x}(t-1), u(t-1)) \\ = \frac{1}{Z} p(\mathbf{x}(t)|p(S(t))) \\ \cdot p(u(t)|u(t-1)) \cdot p(\mathbf{x}(t)|\mathbf{x}(t-1), u(t)), \quad (3) \end{aligned}$$

where  $Z$  is a normalizing constant. Since Eq. (3) is impractical to calculate, we utilize Monte-Carlo sampling to approximate the eye focusing density  $p(\mathbf{x}(t)|\{p(S(i))\}_{i=1}^t)$ . 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. (3), 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

We have to start with some a priori knowledge about the model parameters. Simultaneous estimation of all maximum likelihood (ML) parameters can be optimal but impractical due to the substantial calculation cost. Therefore we divide the parameter estimation into a two-stage estimation process.

#### 3.1. Parameters for stochastic saliency

Using the expectation-maximization (EM) algorithm, the first stage estimates the parameter set  $\theta_s = (\sigma_{s1}, \sigma_{s2})$  for computing SSM. In this case, the prior observations are the saliency maps  $\bar{S} = \{\bar{s}(i)\}_{i=1}^T$ , and the hidden variables are the SSMs  $S = \{s(i)\}_{i=1}^T$ . In the  $k$ -th E-step, the mean  $\hat{s}_k(t|T)$  and variance  $\sigma_{s,k}^2(t|T)$  of the Gaussian density  $p(s(t)|\bar{S}; \theta_{s,k-1})$  is updated recursively using Kalman smoother and the previously estimated parameter set  $\theta_{s,k-1}$ . In the  $k$ -th M-step, the parameter set  $\theta_{s,k}$  is updated using the mean and variance of the Gaussian density  $p(s(t)|\bar{S}; \theta_{s,k-1})$ .

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

#### 3.2. Parameters for eye positions estimation

The second stage derives the parameter set  $\theta_x = (\gamma_{x0}, \gamma_{x1}, \sigma_{x0}, \sigma_{x1}, \Phi)$  for computing the eye position density map. Instead of the EM algorithm, we utilize the Viterbi learning method which allows a greater flexibility in initializing our hidden states given the training data. The training data are

the eye focusing positions recorded from human subjects, and the hidden states are eye movement patterns. In the first step of the  $k$ -th learning iteration, a new hidden state sequence  $\{u_k(t)\}_{t=1}^T$  is computed by using the Viterbi algorithm and the previously estimated parameter set  $\theta_{x,k-1}$ . The second step then updates the ML parameter set  $\theta_{x,k}$  accordingly by examining the statistical distribution of the new sequence of eye movement patterns and the training data:

$$\begin{aligned} \gamma_{xi,k} &= \frac{\sum_{t=2}^T \|\mathbf{x}(t) - \mathbf{x}(t-1)\| u_k(t)_i}{\sum_{t=2}^T u_k(t)_i}, \\ \sigma_{xi,k}^2 &= \frac{\sum_{t=2}^T (\|\mathbf{x}(t) - \mathbf{x}(t-1)\| - \gamma_{xi,k-1})^2 u_k(t)_i}{\sum_{t=2}^T u_k(t)_i}, \\ \phi_{(i,j),k} &= \frac{\sum_{t=2}^T u_k(t)_i u_k(t-1)_j}{\sum_{t=2}^T u_k(t-1)_j}. \end{aligned}$$

## 4. EVALUATION

### 4.1. Collecting eye tracking data

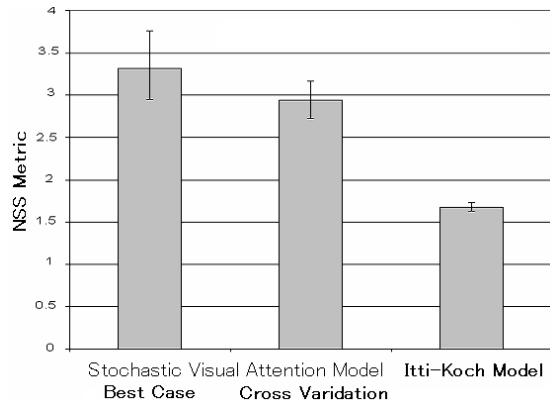
For the purpose of parameter training and model evaluation, we collected samples of eye focusing positions from six human subjects. Each subject viewed 13 different video clips. The first three video clips are taken from the "Movie Task" video demonstration distributed by VisCog Productions, Inc., and each of the remaining 10 clips comprises a sequence of natural scenes. The total length of each video varies from 30 to 90 seconds. Each subject only views each clip once with no specific instructions given. Each video clip was presented on a 18" computer monitor (1280 x 1024 pixels, 60 Hz), and played with a standard video file player placed on the top left of the monitor window. Subjects rested on a chin-rest and were seated at a viewing distance of 60cm. Each subject's right eye position was recorded at 30 Hz with an eye tracking device [10] based on corneal reflection.

### 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) [9]. 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 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.



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

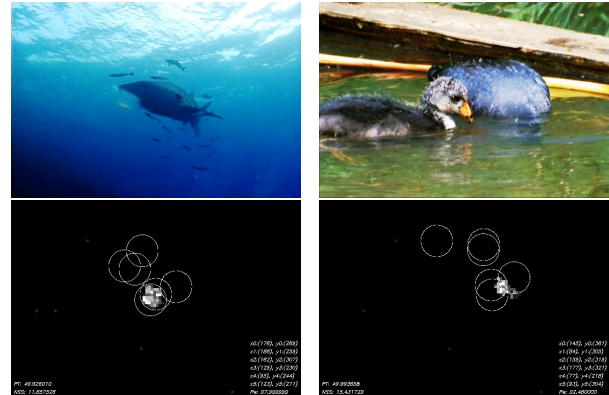
### 4.3. Results

We evaluated the performance of our model by comparing it against Itti-Koch model [3]. 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 model trained with the two different scenarios, and Fig. 4 shows samples of eye focusing density maps estimated from our model. The results of all video clips from all experiment types have outperformed the Itti-Koch model. The average NSS result indicates that our model trained from either scenario substantially better than Itti-Koch model by more than 75%. The result in the cross validation scenario verified that our model still performed significantly well independent of which training set is utilized.

### 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 1) the probability of having the maximum saliency response at a given region evaluated based on signal detection theory, 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 a previous deterministic model. Future work includes a unified approach to estimate all ML parameters, introduction of spatial relationships of stochastic saliency maps, and a better integration of the bottom- and the top-down information.



**Fig. 4.** Samples of results. The top row shows snapshots taken from the input videos, the bottom shows the corresponding eye focusing density maps, and each white circle in the density map is the eye focusing position of a subject.

### 6. ACKNOWLEDGEMENTS

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.

### 7. REFERENCES

- [1] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.
- [2] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.
- [3] L. Itti et al., "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. PAMI*, vol. 20, no. 11, pp. 1254–1259, November 1998.
- [4] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. PAMI*, vol. 22, no. 9, pp. 970–982, 2000.
- [5] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. CVPR2005*, June 2005, pp. 631–637.
- [6] C. Leung et al., "A computational model of saliency depletion/recovery phenomena for the salient region extraction of videos," in *Proc. ICME2007*, July 2007, pp. 300–303.
- [7] M. P. Eckstein et al., "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.
- [8] T. Koike and J. Saiki, "Stochastic saliency-based search model for search asymmetry with uncertain targets," *Neurocomputing*, vol. 69, pp. 2112–2126, October 2006.
- [9] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *Proc. CVPR2007*, June 2007, pp. 1–8.
- [10] T. Ohno et al., "FreeGaze: A gaze tracking system for everyday gaze interaction," in *Proc. Symposium on ETRA*, 2002, pp. 125–132.