

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

Derek Pang <sup>(1)</sup>, Akisato Kimura <sup>(2)</sup>, Tatsuto Takeuchi <sup>(2)</sup>, Junji Yamato <sup>(2)</sup>, Kunio Kashino <sup>(2)</sup>  
 (1) School of Engineering Science, Simon Fraser University  
 (2) NTT Communication Science Laboratories, NTT Corporation

## Background

- Developing an accurate computational model of human visual system has been a long-standing challenge.
- Such a model Enable any system to select just relevant information from a complex and cluttered visual input.
- Applications: Robotics, surveillance, image/video recognition and retrieval



Input frame

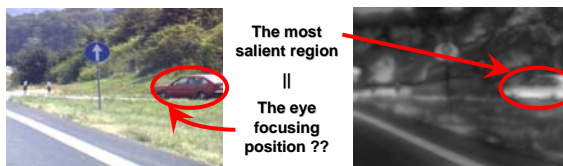
Saliency map

[Itti et al. 1998]

## Problems

**Different people may attend to different regions of a given visual input at the same time !**

- Most previous computational models only selects a fixed attended location every time for the same input.
- need to introduce a stochastic model of human visual attention

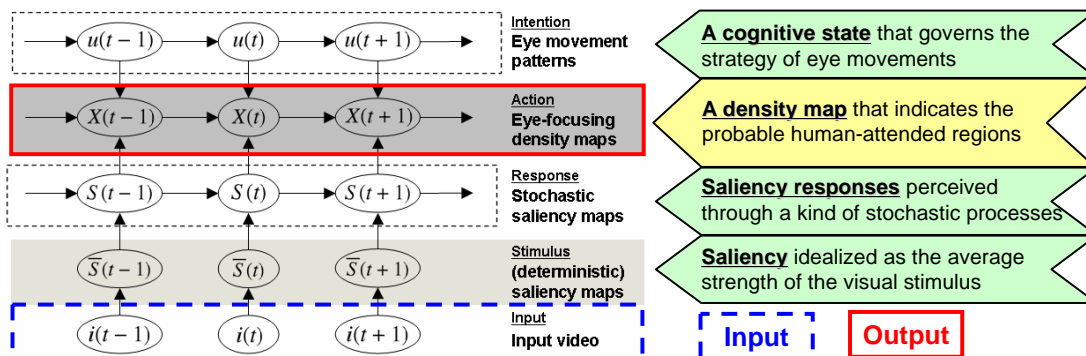


## Contributions

**The first stochastic model of human visual attention with a dynamic Bayesian network**

- Simulate and combine the visual saliency response and the cognitive state
- Automatically predict the likelihood of where humans focus on only from an input video

## Proposed stochastic model



### Saliency maps

Itti-Koch model [Itti et al. 1998]

### Stochastic saliency maps

- Fundamental Gaussian state-space model with a saliency map as observation
- Can be estimated by Kalman filter

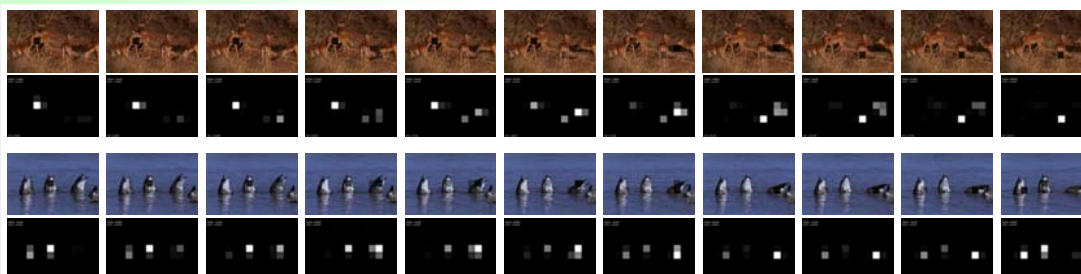
### Eye movement patterns

Independently transit from one to another

### Eye focusing density maps

- The probability having the maximal response, and being the eye focusing position
- The degree of eye movements is driven by the current eye movement pattern

## Snapshots



[Top] Input Video (masked with eye focusing density maps)

[Bottom] Eye focusing density maps (more white, more probable)

Demonstration videos are available here.

To be available online:

<http://www.brl.ntt.co.jp/people/akisato/saliency2-j.html>

### Saliency maps (SM)

Idealize the average strength of visual stimulus

Stimulus (Deterministic) saliency map

Input video

Itti-Koch-Niebur model [Itti et al. 1998]

- # Calculate SM from each frame independently
- # Extract and integrate inter-scale spatial contrasts of fundamental features

Fundamental features:  
Intensity, color opponents (red/green, blue/yellow), orientations (0,  $\pi/4$ ,  $\pi/2$ ,  $3\pi/4$ ), motion energies (horizontal, vertical)

### Stochastic saliency maps (SSM)

Saliency response through a Gaussian random process

Response Stochastic saliency maps

Stimulus (Deterministic) saliency map

Fundamental Gaussian state-space model

- (1) Saliency is perceived through a Gaussian process  

$$p(\mathbf{s}(t, \mathbf{x}); \bar{\mathbf{s}}(t, \mathbf{x})) = \mathcal{G}(\mathbf{s}(t, \mathbf{x}); \bar{\mathbf{s}}(t, \mathbf{x}), \sigma_{t1})$$
- (2) Exploit the temporal smoothness  

$$p(\mathbf{s}(t, \mathbf{x}); \mathbf{s}(t-1, \mathbf{x})) = \mathcal{G}(\mathbf{s}(t, \mathbf{x}); \mathbf{s}(t-1, \mathbf{x}), \sigma_{t2})$$

The state of the stochastic saliency map can be predicted using Kalman Filter

### Eye-focusing density maps (EFDM)

#### 1. [Bottom-up] Estimating EFDM from SSM

(1) The position in which stochastic saliency takes the maximum is determined as the eye focusing position.

→ The probability having the maximal response, and being the eye focusing position

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

#### 2. [Top-down] Estimating EFDM from EMP

(2) EMP independently transits from one to another (u=0: passive state / u=1: active state)

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

(3) The degree of eye movements is driven by EMP (u=0: smaller mean and var of eye movement dist) (u=1: larger mean and var of eye movement dist)

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

#### 3. Integrating bottom-up and top-down information

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

Bottom-up                      Top-down

Monte-Carlo sampling of eye positions and EMP to approximate EFDM

Time t-1                      Time t

### Model parameter estimation

Estimating maximum likelihood (ML) model parameters by utilizing

- # saliency maps calculated from the input videos
- # eye focusing positions obtained by eye tracking devices

Simultaneous estimation is impractical → Separate it into two stages

[First stage] Estimating model parameters for SSM with saliency maps

[Second stage] Estimating model parameters for EFDM with eye focusing positions of humans

#### First stage

Input videos

Extracting saliency maps  
 $\bar{\mathbf{S}} = \{\bar{\mathbf{S}}(t)\}_{t=1}^T, \bar{\mathbf{S}}(t) = \{\bar{\mathbf{S}}(t, \mathbf{x})\}_{\mathbf{x} \in \mathcal{I}}$

Model parameter estimation with the EM algorithm  
 $\theta_{x,t} = (\sigma_{x1,t}, \sigma_{x2,t})$

Model parameters for SSM estimation  
 $\theta_x = \lim_{k \rightarrow \infty} \theta_{x,t}^k$

#### Second stage

Collecting eye movement samples by using an eye tracking device

Eye focusing positions of human subjects  
 $\bar{\mathbf{X}} = \{\bar{\mathbf{X}}_n\}_{n=1}^{N_s}, \bar{\mathbf{X}}_n = \{\bar{\mathbf{x}}_n(t)\}_{t=1}^T$

Model parameter estimation with the Viterbi learning method  
 $\theta_{x,t} = (\gamma_{x1,t}, \gamma_{x2,t}, \sigma_{x1,t}, \sigma_{x2,t}, \Phi_x)$

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

### Experiments

#### Evaluating the accuracy by comparing human eye motions

[Input] 13 video clips including natural scenes (640x480, 15fps, 30-90sec)  
 [Setup] 1280x1024 LCD display, 6 subjects, no specific instructions  
 [Device] An eye tracking device based on corneal reflection (30fps)  
 [Metric] Normalized scanpath saliency (NSS)  
 (Intuitively, the degree of "non-randomness" in eye fixation)

Output

Distribution of normalized pixel values

Normalize

$\bar{\mathbf{x}} = \mathbf{0}$   
 $\sigma^2 \mathbf{x} = \mathbf{I}$

NSS= 1.75

#### Comparing average NSS scores

Model	Average NSS Score
Best case	~3.3
Cross validation	~2.9
Our stochastic model	~3.3
Itti-Koch model	~1.7

Proposed                      Previous