



A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network

June 26, 2008

Derek Pang^(1,2), Akisato Kimura⁽¹⁾, Tatsuto Takeuchi⁽¹⁾,
Junji Yamato⁽¹⁾, Kunio Kashino⁽¹⁾

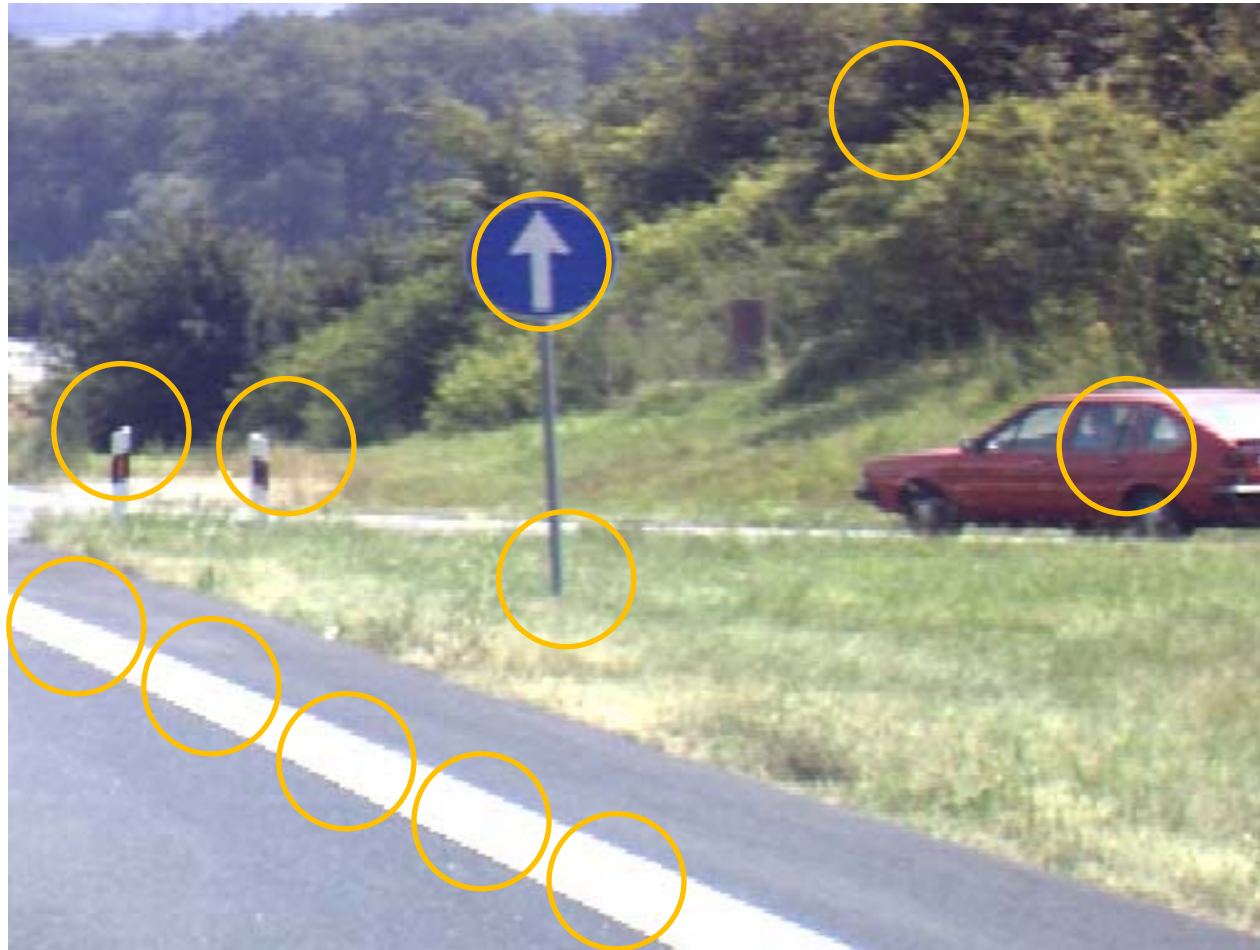
(1) NTT Communication Science Laboratories
Media Recognition Group, Media Information Laboratory



(2) Simon Fraser University
School of Engineering Science

SFU

Where would you focus?



Where would you focus?

- This example illustrates that

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

Feature Integration Theory

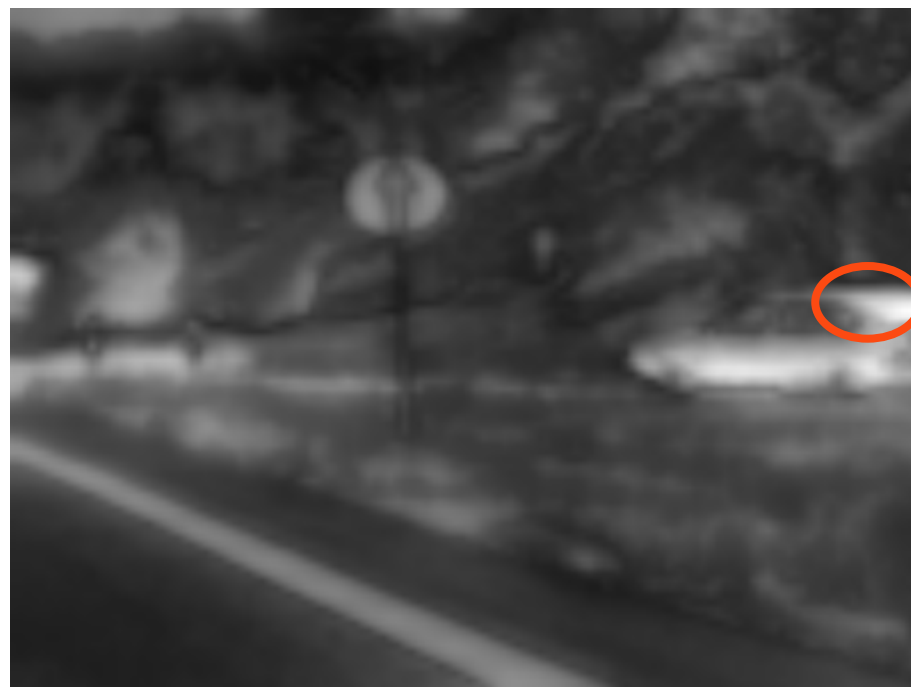
- The vast visual information are first broken down into several primitive visual features, or namely, *feature maps*.
- The *feature maps* are then processed and integrated to form a *saliency map*
- The *saliency map* measures the perceptual quality that makes certain regions of a visual input immediately catches our attention.

Deterministic Nature of Current Models

- Most current saliency models only selects a fixed attended location every time for the same visual input based on the feature-integration theory.



Input Image



Saliency map

Objective

- To develop an accurate and *non-deterministic* computational model of human visual attention
- To identify relevant visual information from a visual video without any prior experiences of the inputs
- Application: multimedia information retrieval, robotics, surveillance, driving assistance, video recognition, consumer video camera etc.

Our Proposed Model

A Stochastic Model of Selective Attention with a Dynamic Bayesian Network

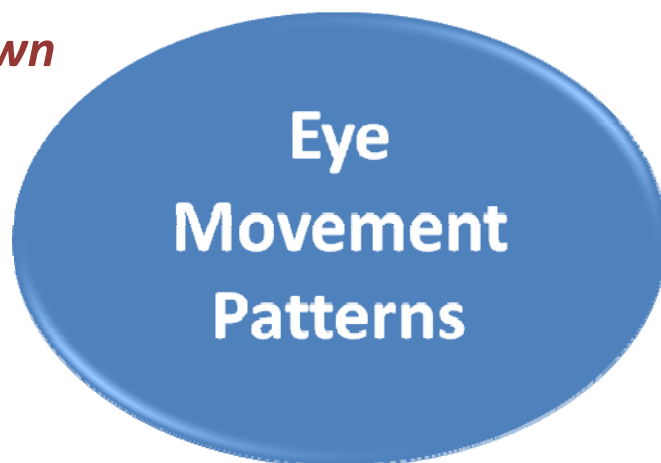
Presented by Derek Pang

(¹) NTT Communication Science Laboratories
Media Recognition Group, Media Information Laboratory



Our Motivation

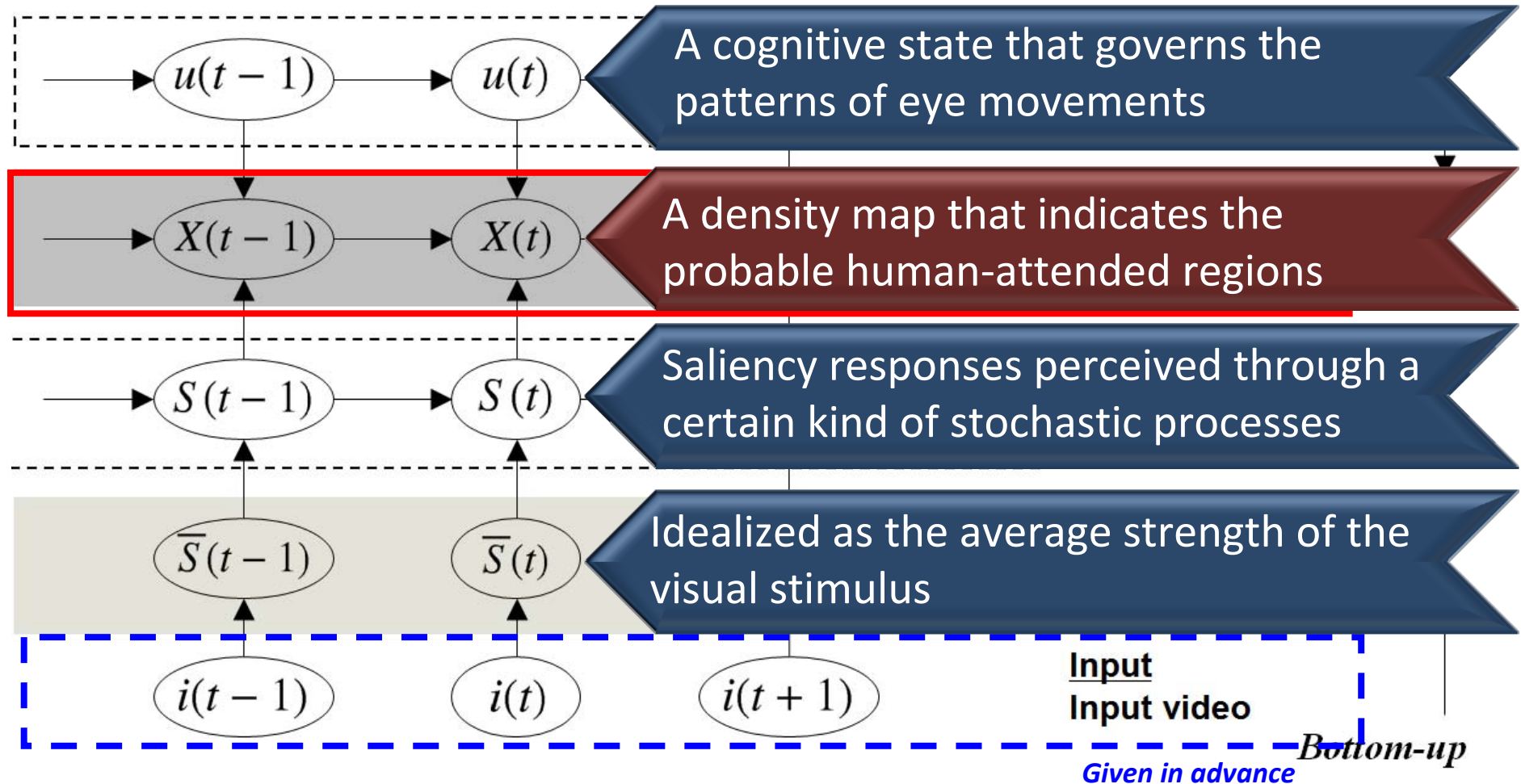
Top-down



Bottom-up



Stochastic Visual Attention Model

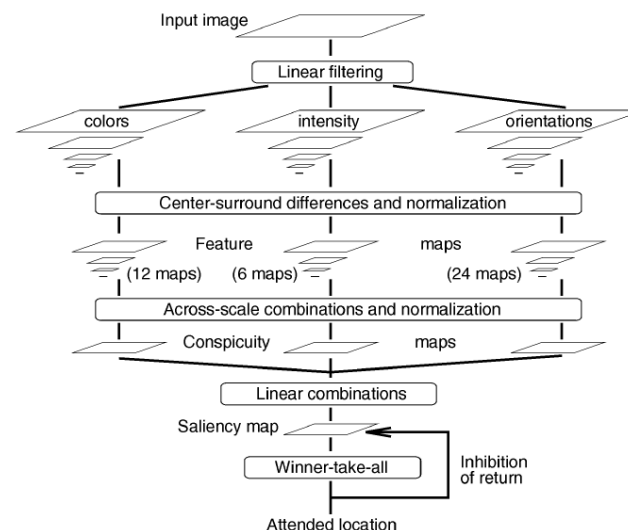


Extracting Deterministic Saliency Map

- Itti-Koch Saliency Model (Itti et al. 1998)
- Include a 'Retinal' Filter

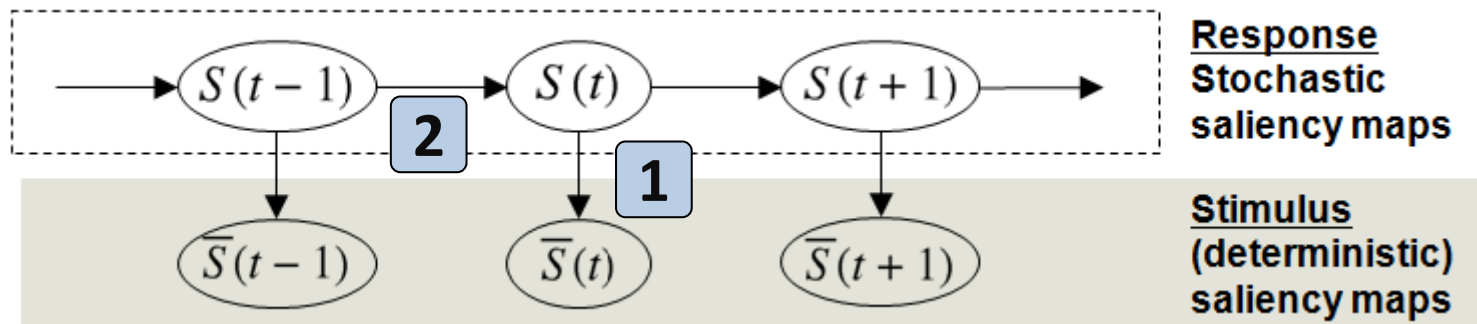
Ten Feature channels :

- 2 color opponents (Red/Green, Blue/Yellow)
- luminance
- temporal luminance flicker
- 4 orientations (0° , 45° , 90° , 135°)
- 2 oriented motion energies (horizontal and vertical)



Estimating Stochastic Saliency Map

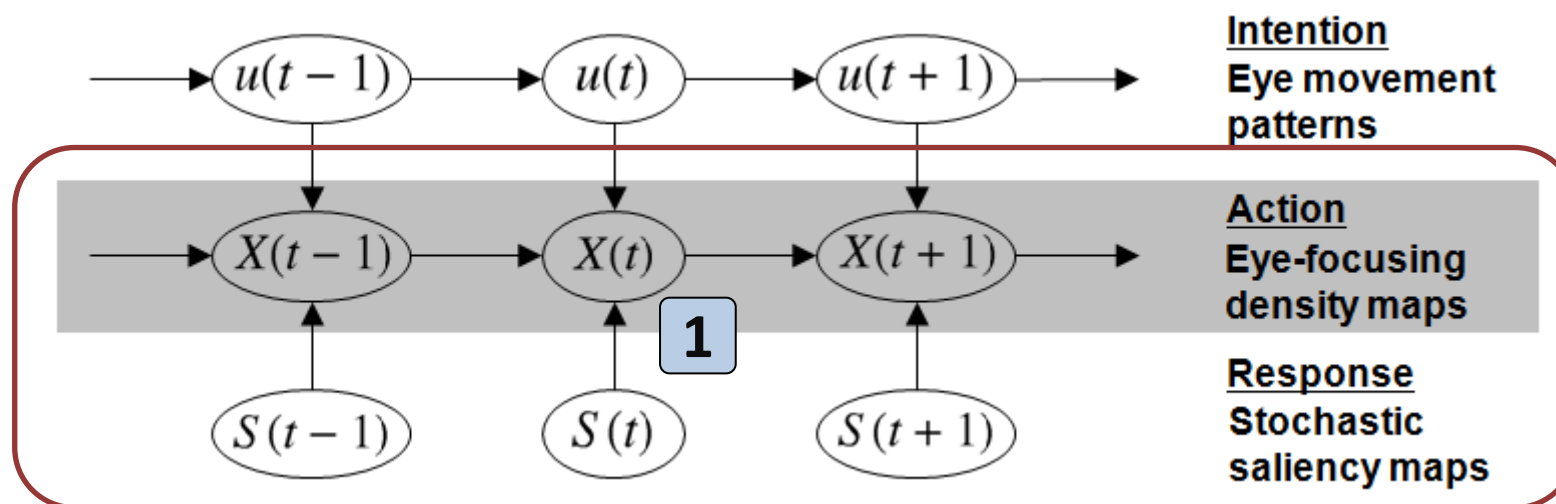
- A fundamental state-space model is introduced.



1. Saliency map is observed through a Gaussian random process
 2. Exploits the temporal smoothness
- The state of the stochastic saliency map can be predicted using Kalman Filter

Estimating Eye Focusing Density Maps (1)

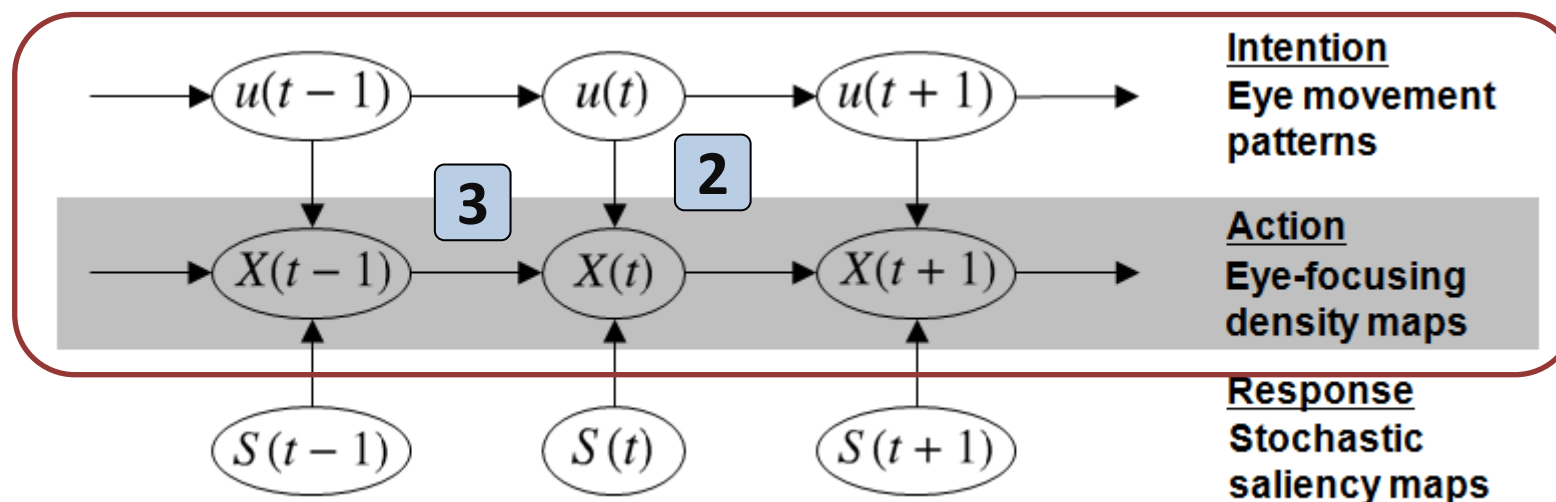
- A kind of hidden Markov model (HMM) is used.



1. The probability having the maximal saliency response, and being the eye focusing position

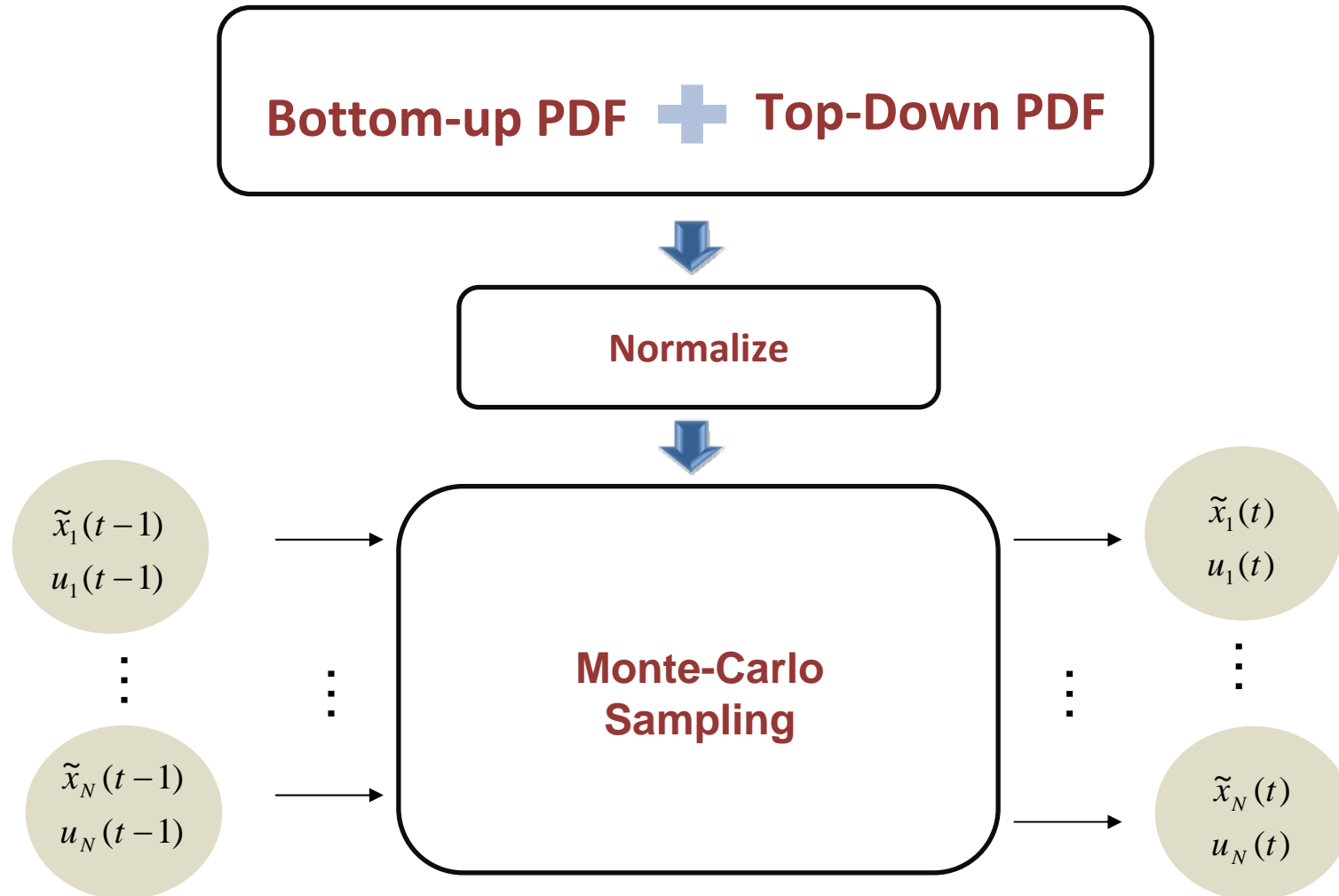
Estimating Eye Focusing Density Maps (2)

- A kind of hidden Markov model (HMM) is used.

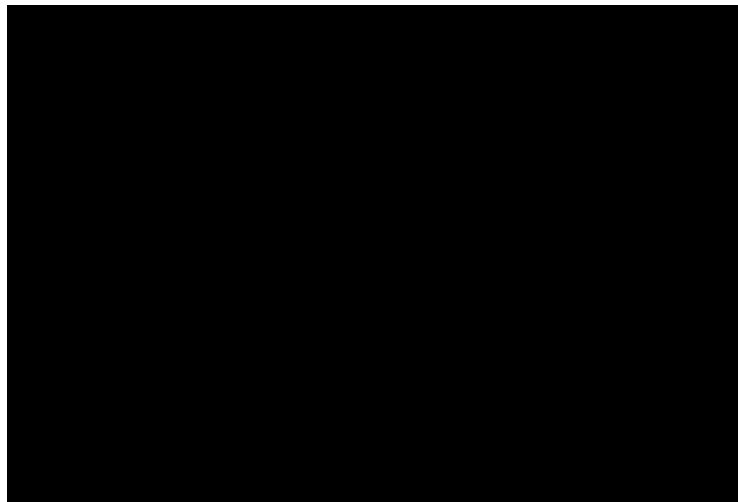


- The degree of eye movements is driven by eye movement patterns
- The current eye focusing position depends on the previous position

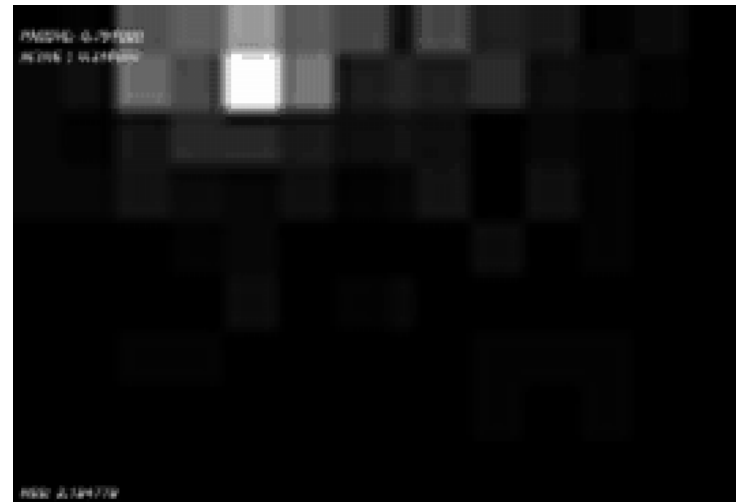
Generating Eye-Focusing Density Map



Demo



Input Video



Eye Positions Density Maps



Evaluation

A Stochastic Model of Selective Attention with a Dynamic Bayesian Network

Presented by Derek Pang

Media Recognition Group, Media Information Laboratory
NTT Communication Science Laboratories

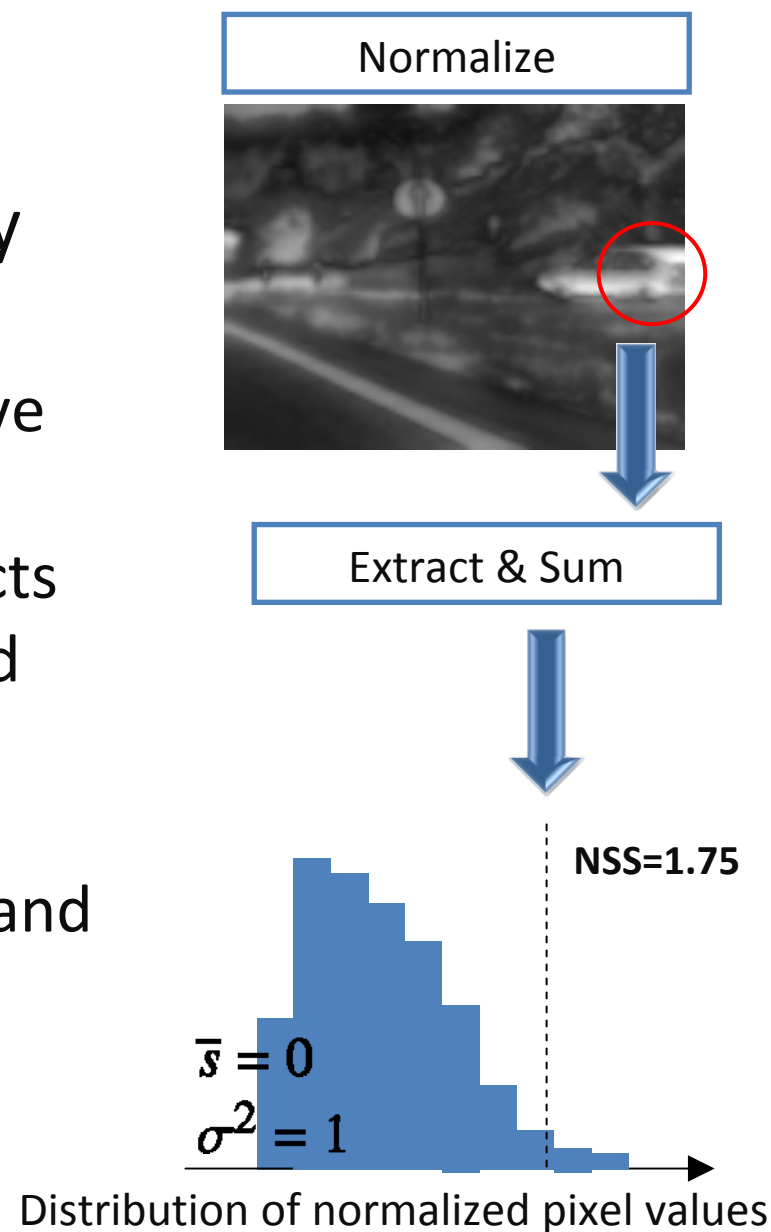


Experiment Setup

- Collected eye movement samples from six human subjects using an eye tracking device based on corneal reflection
- Evaluation data: 13 Video clips
 - 3 video clips from “Movie Task” video demonstration distributed from VisCog Production
 - Each of the 10 other video clips contain a sequence of five to six different natural scenes
- Video clip length : 30 to 90 seconds
- No specific instruction is given to the viewers (passive viewing)

Evaluation Metric

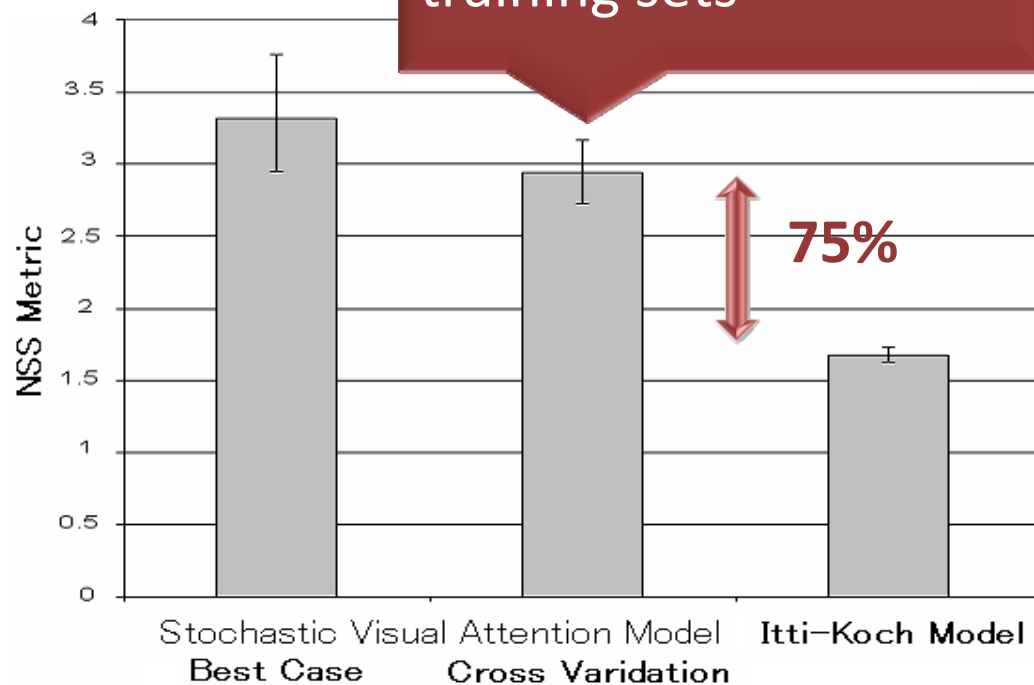
- Normalized scanpath saliency (NSS)
 - Each map is normalized to have mean=0 and dev=1.
 - Eye positions of human subjects are overlaid on the normalized map.
 - Normalized pixel values are extracted from each fixation, and summed up to give the NSS.
 - NSS can be compared with the distribution of random eye fixations.



Experiment Result

- Best-case scenario
 - The model parameter is trained by its own data
- 3-fold cross validation scenario
 - Only one of 3 data sets is retained for evaluation with remaining sets being the training data

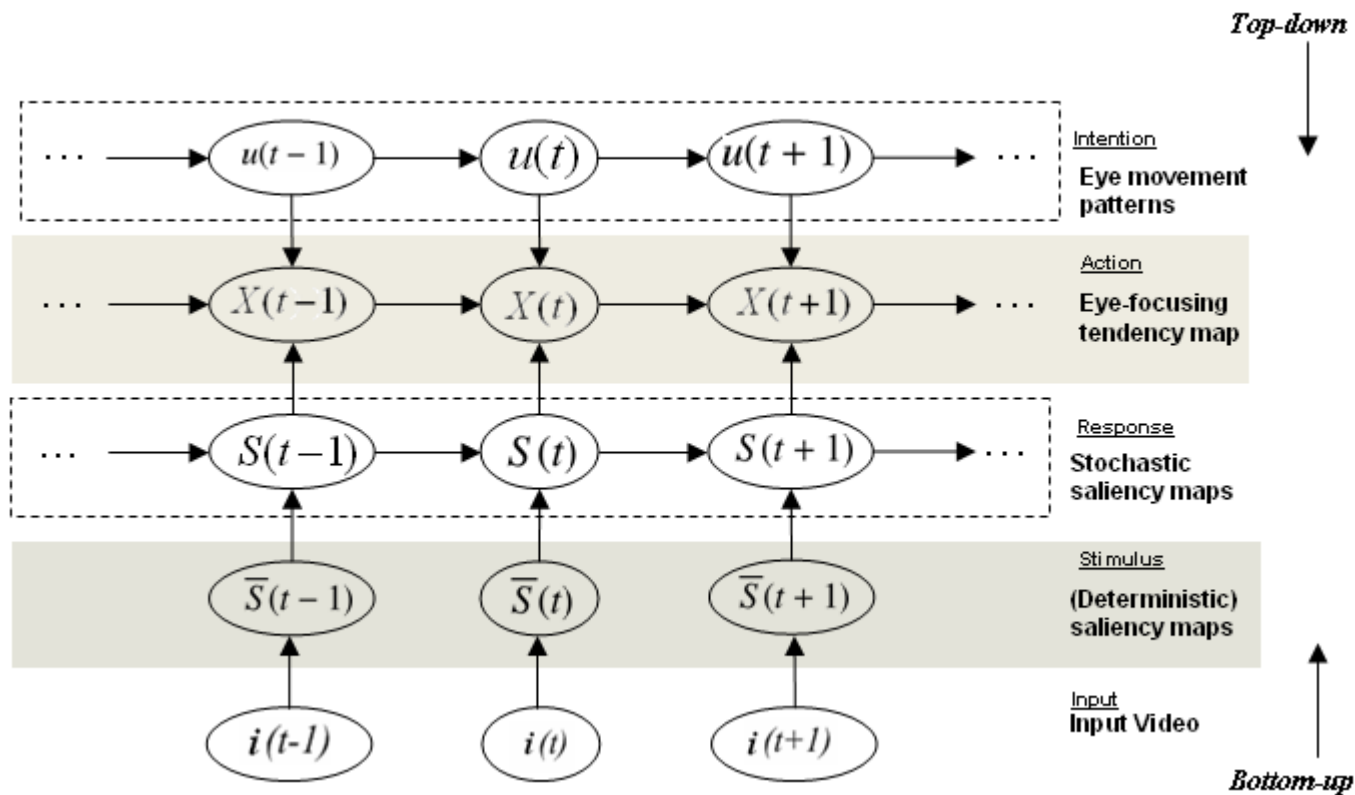
Our model performs significantly better independent of the training sets



Average result for each training scenario

Conclusion

- First unified stochastic model that integrates top-down and bottom-up information
- Predict the likelihood of human-attended regions without any prior experience.
- Experiment has revealed promising results against previous deterministic models.
- Future work:
 - Spatial relationship?
 - Better integration of information?
 - Computational time improvement?



Thank you. Questions/Comments