Multi-mode Saliency Dynamics Model for Analyzing Gaze and Attention
(Authors Version)

Ryo Yonetani*
Kyoto University

Hiroaki Kawashima†
Kyoto University

Takashi Matsuyama‡
Kyoto University

Abstract

We present a method to analyze a relationship between eye movements and saliency dynamics in videos for estimating attentive states of users while they watch the videos. The multi-mode saliency-dynamics model (MMSDM) is introduced to segment spatio-temporal patterns of the saliency dynamics into multiple sequences of primitive modes underlying the saliency patterns. The MMSDM enables us to describe the relationship by the local saliency dynamics around gaze points, which is modeled by a set of distances between gaze points and salient regions characterized by the extracted modes. Experimental results show the effectiveness of the proposed model to classify the attentive states of users by learning the statistical difference of the local saliency dynamics on gaze-paths at each level of attentiveness.

CR Categories: H.1.2 [MODELS AND PRINCIPLES]: Information Systems—User/Machine Systems;

Keywords: switching linear dynamical system, saliency map, attentive state estimation

1 Introduction

"Eyes are a window into the mind" — eye movements are often regarded as crucial clues to understand user states [Just and Carpenter 1976; Calder et al. 2002]. Estimation techniques of the states such as interests [Brandherm et al. 2007; Nakano and Ishii 2010; Hirayama et al. 2010], attentions [Fletcher and Zelinsky 2009; Doshi and Trivedi 2010] or fatigues [Ji et al. 2006] allow machines to perform an intelligent interaction with humans. Our goal is to estimate users' attentive states (e.g., highly attentive to or distracted from the task) from eye movements while they watch general videos.

Estimation methods of user states often include an analysis of relationships between eye movements and contents or surrounding environments being looked at. It is because eye movements can be affected not only by the user states but also by the contents or environments as human information processing is classified into two types: a controlled and an automatic processing [Schneider and Shiffrin 1977]. Related work focuses on "what types of dynamics does the user look at" for the estimation, and they can be classified based on how to describe features of the eye movements using the properties of gaze-related objects in the contents or the surrounding environments. For instance, many studies on interactive systems begin the analysis by specifying objects being looked at, and then extract unique features such as the duration or the frequency of gazing targets [Qvarfordt and Zhai 2005; Brandherm et al. 2007], 3-gram sequences of the targets [Nakano and Ishii 2010], or reaction times to dynamic content updates [Hirayama et al. 2010]. Some studies on driving assistance systems investigate the correlation between eye gazes and the environments. They detect specific objects from the environments using optical flow [Doshi and Trivedi 2010] or obstacle, sign and pedestrian detection [Fletcher and Zelinsky 2009], and analyze the relationship between gaze directions and the object locations to estimate the driver's attention. Those studies employ heuristic specifications for the extraction of gaze-related objects and their dynamics from the contents or environments.

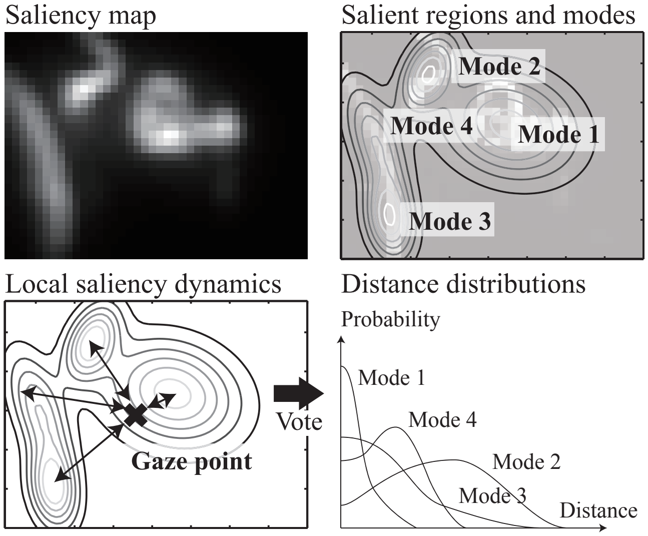

Figure 1: *The multi-mode saliency-dynamics model describes time-varying patterns of saliency dynamics as multiple sequences of primitive modes extracted from the dynamics. The local saliency dynamics around gaze points can be represented by using a set of distances between the gaze points and salient regions characterized by the modes. We statistically learn the variation of the local saliency dynamics on gaze scan-paths in order to estimate the attentive states toward videos being looked at.*

On the other hand, we aim to estimate user states toward general videos from eye movements. Eye movements on the videos such as movies [Goldstein et al. 2007], animations [Munn et al. 2008], or dynamic natural scenes [Dorr et al. 2010] are complicated due to the variety of the videos, and existing studies mainly focus on similarities of the eye movements among subjects. Consequently, it remains unclear how the eye movements can be correlated with various dynamics in the videos. General videos usually contain enormous variety of objects, and the objects have complex dynamics such as appearances and disappearances, motions, shape deformations, temporal texture variations, and temporal variations of visual saliency (i.e., the strength of targets' attractiveness to users' bottom-up attention). These dynamics are all expected to affect eye movements, and modeling of such complex dynamics can be a fundamental approach to estimate the user states toward the videos.

*e-mail:yonetani@vision.kuee.kyoto-u.ac.jp
†e-mail:kawashima@i.kyoto-u.ac.jp
‡e-mail:tm@i.kyoto-u.ac.jp

In this paper we propose a method that models the complex dynamics using simpler descriptions and examines "what types of dynamics does the user look at" from eye movements, in order to estimate the attentive states of users toward general videos. The main contribution is to introduce a novel model called the *multi-mode saliency-dynamics model (MMSDM)*, which describes dynamics of multiple visually-salient regions by simpler dynamics referred to as *modes* (Figure 1). The MMSDM is a model composed of multiple switching linear dynamical systems (SLDSs). Since each SLDS enables us to model the dynamics of a single salient region as the switching between modes, the MMSDM can capture the overall saliency dynamics in the videos consisting of multiple salient regions by the set of modes.

The MMSDM can describe the local saliency dynamics around gaze points by the spatial relationship between the gaze points and salient regions characterized by modes. We estimate the attentive state by statistically learning variations of the local saliency dynamics on gaze scan-paths, which is conditioned by several levels of attentiveness. Namely, the MMSDM provides a framework to classify user's attentive states based on "what types of local saliency dynamics does the user look at".

2 Overview of the proposed method

2.1 Problem setting and approach

Assume that general videos such as TV commercials are displayed on a screen, and a user watches the videos. The user's eye movements can be observed as a sequence of gaze points on the screen by using an eye tracker. The attentive state is measured as how strong users pay attention to the videos; that is, we assume the attentive state can be quantified into several levels. Attentive-state estimation is a problem consisting of classifying the levels of attentiveness based on the videos and the observed eye movements.

The basic concept behind the proposed method is that different gaze scan-paths can be observed depending on the level of attentiveness. Kahneman proposed the attention theory that likens attention to a limited resource which is allocated to specific tasks [Kahneman 1973]. Following this theory, the level of attentiveness can be regarded as the amount of attention resource that allocates to video-viewing tasks in this study. And thus, we assume that users watch videos more actively using the attention resource when they are highly attentive to the videos.

A challenge arises here, since the eye movements can be affected not only by the level of attentiveness but also saliency dynamics that attract human gaze in a video as mentioned in Section 1. We therefore assume that the observed eye movements can be conditioned by both of them. Let S and E be features of saliency dynamics and those of eye movements respectively, and let A be the level of attentiveness. When E is statistically learned under the condition of both S and A, the unknown level of attentiveness $\hat{A}$ is estimated from a newly-observed pair of $\hat{S}$ and $\hat{E}$ as follows:

$$\begin{aligned}\hat{A} &= \arg\max_{A} P(A \mid \hat{S}, \hat{E}) \\ &= \arg\max_{A} P(\hat{E} \mid \hat{S}, A) P(A),\end{aligned} \tag{1}$$

where we assumed that the level of attentiveness A is independent of the saliency S for simplicity.

This study introduces the descriptions of features S and E. As mentioned in Section 1, the important clue to the estimation is "what types of dynamics does the user look at", and thus eye-movement features can be described by the properties of objects in contents. From this point of view, S can be described as "the overall saliency dynamics in videos", and E as "the types of local saliency dynamics being looked at". General videos have complex dynamics including appearances and disappearances of salient regions at various locations, motions, shape deformations, temporal texture variations or temporal saliency variations of the regions. Likewise, local saliency dynamics around gaze points can be also complex because they are extracted from the overall dynamics. Such complex dynamics require a large number of parameters of S and E, and make it difficult to learn their probabilistic dependencies.

The MMSDM is introduced to overcome this difficulty. Since the MMSDM describes the overall saliency dynamics as the sequences of modes (see Section 3 and 4 for details), it provides the descriptions of S and E in much more compact forms. Specifically, with regard to the features of saliency dynamics S, a set of modes simultaneously existing in a frame, which we refer to as a scene, is introduced instead of the complex saliency dynamics themselves. And with regard to eye-movement features E, a set of distances between gaze points and salient regions characterized by modes of the scene is employed in order to describe the local saliency dynamics around the gaze points (see Section 5 for details). Consequently, the proposed method can handle the probabilistic dependencies between complex dynamics by using simpler descriptions.

2.2 The proposed method

The overview of our proposed method is as follows. As shown in Figure 2(A), we first extract salient regions to be looked at, and their dynamics patterns from videos. We employ the saliency map [Itti et al. 1998], which is known as a model of the bottom-up visual attention. The saliency map enables us to detect the salient regions automatically from the general videos without using specific heuristics. In order to model the dynamics of salient regions parametrically, we employ the Gaussian mixture model (GMM) and parameterize the temporal variations in a sequence of saliency maps. By tracking the spatio-temporally continuous Gaussian components, we obtain multiple saliency patterns that represent dynamics such as motions, approximate shape deformations and temporal saliency variations of the regions.

We then learn modes, primitive dynamics that appear in the saliency patterns, based on the MMSDM (Figure 2(B)). The MMSDM segments all the saliency patterns into sequences of modes.

As a result of the segmentation, the dynamics in a single frame can be characterized by a scene: a set of modes appeared simultaneously in the frame. The feature of the overall saliency dynamics, S in Eq. (1), is described by the scene. Besides, the local saliency dynamics around gaze points, E, is described by using the distances between the gaze points and regions characterized by modes of the scene. In the learning phase, these distances are learned as distributions under the condition of the scenes and the levels of attentiveness (Figure 2(C)). In the estimation phase, once we newly observe a pair of videos and eye movements, we first identify the scene and then estimate the attentiveness based on Eq. (1).

3 Extraction of saliency patterns

This section introduces the extraction and the parametric representations of saliency-dynamics patterns. We utilize the saliency map here, a bottom-up computational model of visual attention. The saliency map typically includes the extraction of low-level visual features such as intensities, colors, orientations, or motions from sequential images at multiple scales, the normalization and the integration of the features into a 2D map with a saliency value at each

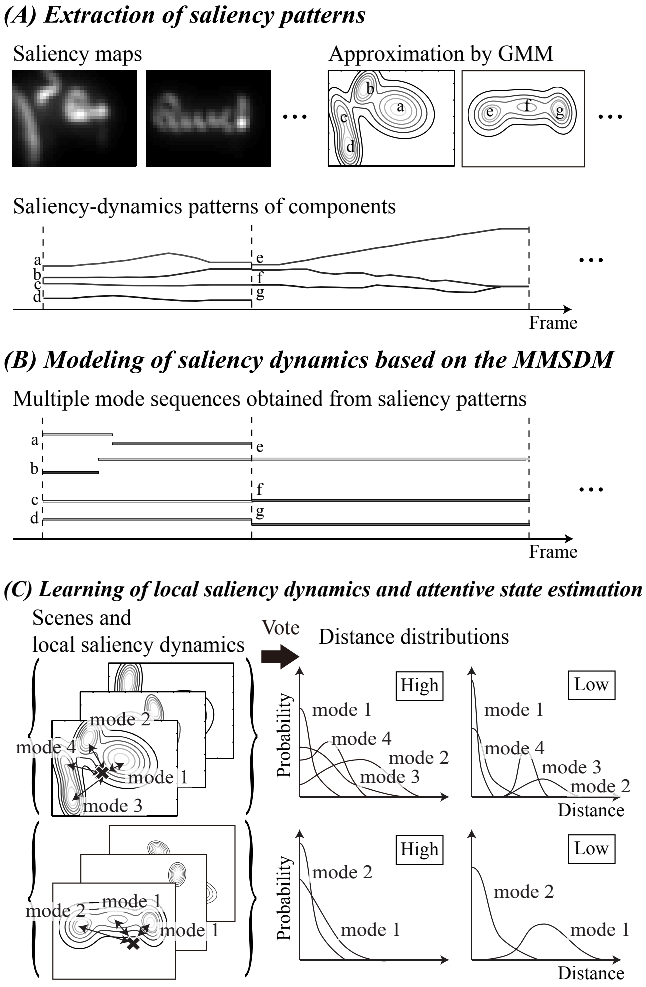

Figure 2: *Overview of the proposed method.*

pixel. We obtain a sequence of saliency maps from a video [1] , i.e., a saliency map $\hat{i}_t$ is computed from an input frame i_t at frame t.

Because videos often contain multiple objects in a frame, the obtained map $\hat{i}_t$ is expected to include several salient regions. These regions individually include the dynamics such as motions, shape deformations, temporal texture variations, and temporal saliency variations. Moreover, the number of regions is time varying; in other words, regions can appear and disappear at any frame throughout a single video.

In the following subsections, we first model a saliency pattern produced by salient regions in Subsection 3.1, and then introduce the technique to track the number of regions in Subsection 3.2.

3.1 Modeling saliency patterns using the GMM

Let us first assume that the numbers of regions in all the frames are given. The video can be then segmented into an interval sequence, $(I^{(1)}, \dots, I^{(K)})$, where the interval $I^{(k)} = [b_k, e_k]$ consists of frames with the same number of regions, $C^{(k)}$. We describe the saliency dynamics in each interval by a set of saliency patterns produced by salient regions.

To parameterize the pattern of the regions, we approximate the saliency maps by the GMM. That is, salient regions are modeled by Gaussian components and their saliency patterns can be obtained by tracking spatio-temporally continuous components. This modeling sacrifices representation of a detailed contour and texture of the regions. Instead, the GMM allows us to describe locations, approximate shapes, and the strength of saliencies of the regions by means, covariances and weights of the components, respectively.

The concrete procedure begins with normalization of the saliency map $\hat{i}_t$ as a probability distribution on the image (i.e., 2-d plane), and approximation of the distribution by massive samples. We then estimate parameters of the GMM based on the expectation-maximization (EM) algorithm. Let us denote the probability distribution of the GMM, which is fitted to $\hat{i}_t$, as Ψ_t. The mean, covariance, and weight vector of c-th component in Ψ_t are respectively denoted as $\boldsymbol{\mu}_t^{(c)}$, $\Sigma_t^{(c)}$, and $\phi_t^{(c)}$. The overall properties of region c at frame t is denoted as $\boldsymbol{\theta}_t^{(c)} \in \mathbb{R}^6$ $(c \in \{1, \dots, C^{(k)}\}, t \in I^{(k)})$. It can be modeled as $\boldsymbol{\theta}_t^{(c)} = ((\boldsymbol{\mu}_t^{(c)})^{\mathrm{T}}, (\boldsymbol{\sigma}_t^{(c)})^{\mathrm{T}}, \phi_t^{(c)})^{\mathrm{T}}$, where $\boldsymbol{\sigma}_t^{(c)} \in \mathbb{R}^3$ consists of the variances and the covariance (i.e., elements of $\Sigma_t^{(c)}$). By tracking the spatio-temporally continuous components in the interval $I^{(k)}$, the saliency pattern of the region c is obtained as the vector sequence $\boldsymbol{\Theta}^{(k,c)} = (\boldsymbol{\theta}_{b_k}^{(c)}, \dots, \boldsymbol{\theta}_{e_k}^{(c)})$.

While the convergence of the iterative procedure of the EM algorithm is guaranteed, the results strongly depend on given initial values. We assume that the salient regions change “smoothly” except for shot changes. And there, for the initial values of the iteration, we employ the optimized parameters in the previous frame.

3.2 Tracking the number of components

This subsection introduces the tracking technique of the time-varying number of components so as to segment a video into intervals. In order to determine the number when a video displays a complex scene and salient regions have ambiguous contours, we prepare the candidates of GMMs with various numbers of components for each frame, and optimize the time-varying number of components. Specifically, we first set the minimum and maximum number of components as $C_{\min}$ and $C_{\max}$, respectively. The GMM distributions and their parameters consisting of each of $\{C_{\min}, \dots, C_{\max}\}$ components are obtained for every frame. The number of components is tracked based on the greedy algorithm.

Let us denote the probability distribution of γ-component GMM on the image at frame t as $\Psi_t^{(\gamma)} = \sum_{c=1}^{\gamma} \phi_t^{(c,\gamma)} \psi_t^{(c,\gamma)}$ $(\gamma \in \{C_{\min}, \dots, C_{\max}\})$, where $\psi_t^{(c,\gamma)} = \mathcal{N}(\boldsymbol{\mu}_t^{(c,\gamma)}, \Sigma_t^{(c,\gamma)})$ describes c-th Gaussian distribution and $\phi_t^{(c,\gamma)}$ denotes the weight of the distribution. The estimated number of components at frame t is denoted as $\hat{C}_t$, and it is derived from the following equation:

$$\hat{C}_t = \mathop{\arg\max}_{\gamma \in \{C_{\min}, \dots, C_{\max}\}} \{V_{\mathrm{D}}(\gamma) - \alpha V_{\mathrm{P}}(\gamma) + \beta \omega_{(t,t-1)} V_{\mathrm{S}}(\gamma, \hat{C}_{t-1})\}. \tag{2}$$

V_{D} is a data term that measures the similarity between the saliency map $\hat{i}_t$ and the fitted γ-component GMM distribution $\Psi_t^{(\gamma)}$ by the Bhattacharyya coefficient denoted as $V_{\mathrm{D}}(\gamma) = BC(\hat{i}_t, \Psi_t^{(\gamma)})$, where we here normalize $\hat{i}_t$ so that it can be treated as a probability distribution. In order to keep components apart and eliminate the redundancy of the models, V_{P} is employed as the parameter term. V_{P} evaluates a similarity among components; this term calculates

[1] The implementation of saliency extraction is in MATLAB using the Saliency Toolbox [Walther and Koch 2006]. The features computed here include intensities, colors, orientations, and inter-frame motions.

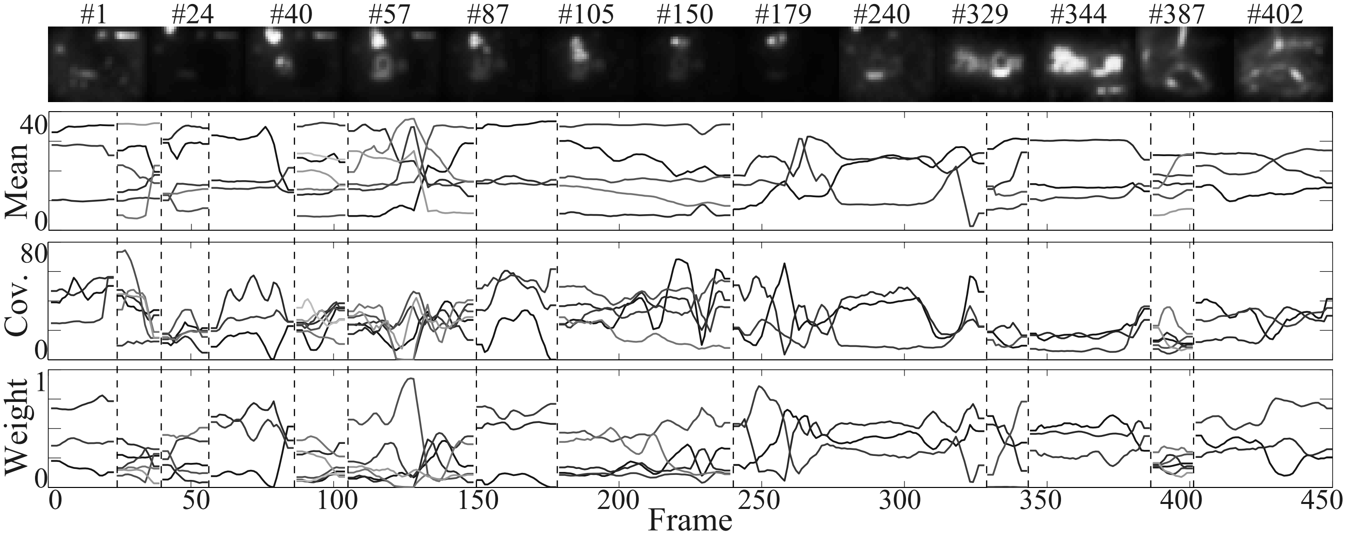

Figure 3: *Example of time-varying saliency patterns. The 1st row of the figure depicts some of the saliency maps and the rest of the rows depict the saliency patterns of salient regions: means (2nd row), covariances (3rd row) and weights (4th row). With regard to means and covariances, the figure depicts the 1st principal component of the signals for visualization. The vertical dashed lines describe the frames in which the number of components changes.*

the minimum distance among components by the maximum Bhattacharyya coefficient as follows:

$$V_{\mathrm{P}}(\gamma) = \max_{i,j\in\{1,\ldots,\gamma\},i\neq j} \left(BC(\psi_t^{(i,\gamma)}, \psi_t^{(j,\gamma)}) \right). \tag{3}$$

$\alpha(> 0)$ is a scale parameter defined as a ratio between the standard deviation of V_{D} sequence and that of V_{P} sequence.

V_{S} is a smoothness term described as $V_{\mathrm{S}}(\gamma, \hat{C}_{t-1}) = \delta(\gamma, \hat{C}_{t-1})$ where $\delta(i, j)$ denotes the Kronecker delta. This term is regularized by $\omega_{(t-1,t)}$ that measures the similarity between the successive saliency maps by $\omega_{(t-1,t)} = BC(\hat{\imath}_t, \hat{\imath}_{t-1})$. The parameter $\beta(> 0)$ describes the strength of the regularization. We gradually increase β from zero as long as we obtain short intervals consisting of the same number of components. In this paper we set β to make all the intervals longer than 0.5 sec.

The interval $I^{(k)}$ is obtained with a set of saliency patterns $\left\{\Theta^{(k,c)} \mid k \in \{1,\ldots,K\},\ c \in \{1,\ldots,C^{(k)}\}\right\}$ as a result of the tracking. Figure 3 depicts an example of the patterns obtained from a TV commercial. The vertical dashed lines in the figure describe the frames in which the number of components changes.

4 Modeling saliency dynamics based on the MMSDM

4.1 The multi-mode saliency-dynamics model

The modeling of the overall saliency dynamics in videos is required for the estimation of attentive states as mentioned in Subsection 2.1. In general videos, multiple salient regions can appear and they can have a dynamics individually. And thus, the variety of the overall saliency dynamics can diverge because they are described as the combination of the dynamics of each region. Since we assume the eye movements conditioned by the saliency dynamics as Eq. (1), a simpler description and a smaller variety of the overall dynamics are essential for learning features of the eye movements. Although the GMM describes the salient regions with relatively small number of parameters, the Gaussian components obtained frame-wisely still have a large variety because they includes parameters of locations as well as shapes and saliencies. On the other hand, several typical patterns can be found in the saliency patterns as shown in Figure 3. That is, the classification of patterns in a small interval is expected to lead to a drastic decrease of the variety of the overall dynamics.

The multi-mode saliency-dynamics model (MMSDM) is a model to describe the dynamics of multiple salient regions by simpler descriptions, and it is composed of multiple switching linear dynamical systems (SLDSs). The model of SLDS, which has been widely studied in the field of control theory[2], has now become an efficient tool to represent complex dynamics in human motion [Bregler 1997; Pavlovic et al. 2000; North et al. 2000; Li et al. 2002]. The SLDS models the complex dynamics as the switching between simpler dynamics, where each of the time evolutions of simpler dynamics is formulated by a linear dynamical system. As we introduced in Section 1, we refer to each of these simpler (i.e., linear) dynamics as a mode. Applying the SLDS to a single salient region, we can model its complex dynamics (i.e., motions, approximate shape deformations, and temporal saliency variations) as the switching between multiple modes. In the case that multiple regions exist, multiple SLDSs can be used, where each of the regions is modeled by a different SLDS. Therefore, once we successfully model the overall saliency dynamics in a video, the MMSDM segments saliency patterns comprised in the video into multiple mode sequences.

Figure 4 depicts an example of multiple mode sequences obtained from the saliency patterns shown in Figure 3. Note that the number of mode sequences corresponds to that of components, and the number is estimated using the greedy algorithm in Subsection 3.2. As a consequence, the mode sequences can newly appear or disappear as we see in Figure 4. In the remaining of this section, we first explain the MMSDM more specifically, and we then introduce the

[2] Switching linear dynamical system is often referred to as the “switched linear system” in the field of control theory.

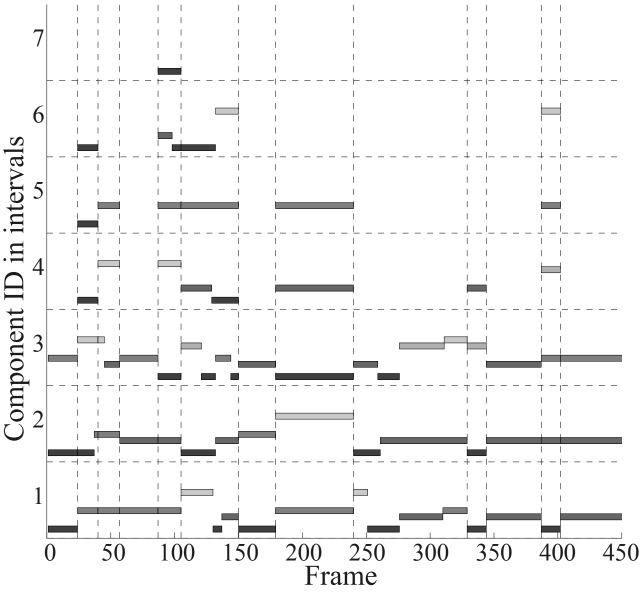

Figure 4: *The mode sequences obtained from saliency patterns shown in Figure 3. The vertical axis shows component ID $\{1, \ldots, C^{(k)}\}$ in each interval $I^{(k)}$, and the vertical position of rectangles describes the ID of modes. The vertical dashed lines describe the frames in which the number of components changes.*

learning and segmentation method of MMSDM briefly.

4.2 Modeling saliency dynamics based on the MMSDM

Let us denote a set of linear dynamical systems as $\mathcal{D} = \{D_1, \ldots, D_N\}$. Each system D_i models a transition of state vectors, from z_{t-1} to z_t, as follows:

$$z_t = F^{(i)} z_{t-1} + \boldsymbol{g}^{(i)} + \boldsymbol{w}_t^{(i)}, \qquad (4)$$

where $F^{(i)}$ is a system matrix and $\boldsymbol{g}^{(i)}$ is a bias vector. $\boldsymbol{w}_t^{(i)}$ is the process noise modeled by a Gaussian distribution $\mathcal{N}(\boldsymbol{w}_t^{(i)} \mid \boldsymbol{0}, Q^{(i)})$. That is, each dynamical system D_i has $F^{(i)}$, $\boldsymbol{g}^{(i)}$ and $Q^{(i)}$ individually. We assume that the systems are fully observable, i.e., the states correspond to signal variables obtained from data.

Our focus is to model a saliency pattern $\boldsymbol{\Theta}^{(k,c)} = (\boldsymbol{\theta}_{b_k}^{(c)}, \ldots, \boldsymbol{\theta}_{e_k}^{(c)})$ as a sequence of modes. The pattern includes motions, approximate shape deformations, and temporal saliency variations. Generally, the motion saliency is obtained by the relative changes such as temporal differences of successive frames [Wildes 1998]. In an analogous way, the other dynamics are expected to include an important clue in their relative changes. We so describe a relative pattern of saliency dynamics by modes; z_t is defined as $z_t = \boldsymbol{\theta}_t^{(c)} - \bar{\boldsymbol{\Theta}}^{(k,c)}$ $(t \in [b_k, e_k])$ where $\bar{\boldsymbol{\Theta}}^{(k,c)}$ denotes a temporal average of $\boldsymbol{\Theta}^{(k,c)}$ in the interval $[b_k, e_k]$. The use of the relative change also reduces the size of the parameter space.

On the other hand, the mode transitions are modeled by a finite state automaton. Let $\{m_1, ..., m_N\}$ be a set of modes that corresponds to $\mathcal{D}$, where the mode m_i corresponds to the linear dynamical system D_i. Since a mode, m_i, has a certain duration, let us introduce the notation $< m_i, \tau >$. That is, the segment $< m_i, \tau >$ corresponds to a saliency pattern modeled by D_i with length τ.

As a result, each saliency pattern is segmented into a mode sequence. Let $\mathcal{M}^{(k,c)} = (M_1^{(k,c)}, \ldots, M_{N_{(k,c)}}^{(k,c)})$ be a mode sequence of the saliency pattern $\boldsymbol{\Theta}^{(k,c)}$. $\mathcal{M}^{(k,c)}$ is generated by the automaton and it assumes the first-order Markov property for the generated intervals and that the adjacent intervals have no temporal gaps or overlaps. Then, the mode transition process can be modeled by the conditional probability, $P(M_n =< m_j, \tau >| \ M_{n-1} =< m_i, \tau_p >)$, where it denotes that the segment $< m_j, \tau >$ occurs after the segment $< m_i, \tau_p >$. Since the duration has a large variability in our case, we chose not to model the distribution of segment durations but to model only the transition probability between modes, i.e., we use $P(< m_j, \tau >|< m_i, \tau_p >) = P(m_j \mid m_i)$.

4.3 Learning and segmentation of saliency dynamics

In the learning phase, all the saliency patterns $\left\{\boldsymbol{\Theta}^{(k,c)} \mid k \in \{1, \ldots, K\},\ c \in \{1, \ldots, C^{(k)}\}\right\}$ obtained from a large number of videos are used for the training data to estimate the parameters of the SLDSs. Here, we assume that all the patterns that appear in videos share the same set of dynamical systems $\mathcal{D}$ in order to employ a large amount of training data. This reduces the size of the parameter space significantly. However, the number of modes, N, is also unknown; that is, the parameters to be estimated in the learning phase is: $\{F^{(i)}, g^{(i)}, Q^{(i)} \mid i = 1, \ldots, N\}, P(m_i \mid m_j)\ (i, j = 1, \ldots, N)$, and N itself. If N is given, then the MMSDM can be learned by solving the segmentation of the saliency patterns and the parameter estimation of linear dynamical systems simultaneously. Therefore, a method of non-hierarchical clustering (e.g., the EM algorithm) can be applied. However, our problem involves an unknown number of modes, and moreover, the non-hierarchical clustering algorithms often strongly dependent on the initial parameters.

Despite these difficulties, some algorithms have been developed for this problem recently. In particular, we employ the method proposed in [Kawashima and Matsuyama 2005]. This method has proposed the use of the model-based hierarchical clustering of linear dynamical systems that estimate their number of modes and the parameters. Hence, the estimation process is divided into two steps: a clustering step and a refinement step that refines all the parameters based on the EM algorithm including the transition probability $P(m_i \mid m_j)$ $(i, j = 1, \ldots, N)$. In parallel with the parameter estimation, this method segments all the training data into mode sequences (see [Kawashima and Matsuyama 2005] for the details of the learning algorithm).

5 Attentive state estimation

5.1 Scenes and local saliency dynamics

Our proposed method aims at the estimation of the attentive state (i.e., the level of attentiveness) by analyzing a relationship between eye movements and saliency dynamics in videos. It follows Eq. (1), and requires both the overall saliency dynamics in videos and local saliency dynamics around gaze points for S and E, respectively. For modeling the saliency dynamics, we introduce the MMSDM and segment the dynamics patterns into multiple mode sequences as shown in Section 4. This section presents the description of S and E using the mode sequences.

First of all, Eq. (1) models probabilistic dependencies in simultaneously-obtained S, E and A. That is, the estimation process is conducted within a certain time window. Here, as shown in Figure 4, modes of saliency dynamics can sometimes change rapidly. With consideration for such rapid changes, we assume frame-wise dependencies, i.e., $P(E_t \mid S_t, A_t)$.

How can S_t and E_t be described using the obtained modes? S_t describes the overall saliency dynamics, and thus we employ a set of modes appeared simultaneously in the frame. Let us first denote a mode of c-th component at frame t in interval $I^{(k)}$ as $s_t^{(c)} \in \{m_1, \ldots, m_N\}$. We refer to S_t as a scene, and denote it as follows:

$$S_t = \left\{ s_t^{(c)} \mid c \in \{1, \ldots, C^{(k)}\} \right\}. \tag{5}$$

The possible number of scenes with C components and N modes can be calculated by a homogenous product,
${}_N\mathrm{H}_C = {}_{N+C-1}\mathrm{C}_C = \binom{N+C-1}{C}$.

Recall that our framework examines "what types of dynamics does the user look at". This specifically requires not only the types but also the spatial relationship of saliency dynamics toward the observed eye movements. The modes describe relative patterns of the saliency dynamics and they do not include the information of the absolute position of the regions. Thus, regarding the description of local saliency dynamics E_t, we utilize the distances between an gaze point and salient regions with the modes. Let us describe E_t around an gaze point $\boldsymbol{x}_t \in \mathbb{R}^2$ by a set of distances as follows:

$$E_t = \left\{ \xi(s_t^{(c)}) \mid c \in \{1, \ldots, C^{(k)}\} \right\}, \tag{6}$$

where $\xi(s_t^{(c)})$ denotes the distance between $\boldsymbol{x}_t$ and the Gaussian component with the mode $s_t^{(c)}$. Specifically, we here define $\xi(s_t^{(c)})$ as follows:

$$\xi(s_t^{(c)}) = \sqrt{(\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(c)})^{\mathrm{T}}(\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(c)})}, \tag{7}$$

where $\boldsymbol{\mu}_t^{(c)}$ is a mean of the component.

5.2 Learning of local saliency dynamics and attentive state estimation

This subsection introduces a method for learning local saliency dynamics E_t. As shown in Eq. (6), E_t consists of a set of distances $\{\xi(s_t^{(c)})\}$. We learn the distances conditioned by the scene S_t (see Eq. (5)) and the level of attentiveness $A_t \in \{A_1, \ldots, A_{N_A}\}$.

With consideration of a large variety of scenes, we assume the independency of distances among modes, and learn the distances as the naive Bayes model. That is, we learn each distance $\xi(s_t^{(c)})$ as a probability distribution for the corresponding mode $s_t^{(c)} \in \{m_1, \ldots, m_N\}$ individually.

Let us denote the probability distribution of the distance $\xi(m)$ to the component with the mode m, where a set of modes equals to S, as $P_m(\xi(m) \mid S, A)$. $P(E_t \mid S_t, A_t)$ is then defined as follows:

$$P(E_t \mid S_t, A_t) = \prod_{s_t^{(c)} \in S_t} P_{s_t^{(c)}}(\xi(s_t^{(c)}) \mid S_t, A_t). \tag{8}$$

We collect various scenes with the same set of modes, and apply the kernel density estimation to the obtained data to estimate the probability distributions of distances.

Given that a new gaze point is observed under a scene $\hat{S}_t$, we calculate the local saliency dynamics $\hat{E}_t$ and estimate the level of attentiveness $\hat{A}_t$ based on the maximum likelihood estimation:

$$\hat{A}_t = \arg\max_A P(\hat{E}_t \mid \hat{S}_t, A). \tag{9}$$

In modifying Eq. (1) to Eq. (9), we assume $P(A)$ as a constant and equal in all the levels of attentiveness while watching each video.

6 Experiments

6.1 Experimental setup

We conducted some experiments and estimated the level of attentiveness. In this experiments, we aimed to discriminate two levels: high and low, as a relatively-simplified evaluation (i.e., $A_t \in \{A_{\mathrm{high}}, A_{\mathrm{low}}\}$). 10 subjects took part in the experiments, and 12 TV commercial videos (15 sec) were employed. The commercial videos are originally designed to attract the visual attention, and therefore are expected to include some obvious salient regions.

Environment and conditions

A subject sat in front of a screen[3], and an eye tracker [4] was installed below the screen. The eye-tracking accuracy was, on average, around 0.7°. The distance between the subject and the screen was around 1000 mm, and in this settings eye movements could be observed during experiments.

As mentioned in Subsection 2.1, the level of attentiveness specifies an amount of attention resource allocated to the video-viewing tasks. We therefore adopt the following two conditions in order to control the attentiveness in the experiments:

Condition 1 (high level of attentiveness) A subject watches a video and answers a simple interview after that.

Condition 2 (low level of attentiveness) A subject watches a video with doing a mental calculation.

For each condition, subjects were asked to orient their gaze to a screen as far as possible.

In the experiments, the videos are split into two groups, VA (six out of all the videos) and VB (the other six videos). Subjects watched all the videos twice by following the procedures below.

- The half of the subjects carried out the tasks as: **(1st trial)** VA —Condition 1, VB –Condition 2, and **(2nd trial)** VB—Condition 1, VA—Condition 2.
- The other half carried out the tasks as **(1st trial)** VB —Condition 1, VA –Condition 2, and **(2nd trial)** VA—Condition 1, VB—Condition 2.

Between the trials, subjects took a small break. The order of videos in video groups was randomized in each of the trials.

Preprocessing and parameters setting

Gaze data was acquired by the eye tracker at 30 Hz. Since we focus on eye movements on a screen, we do not regard the eye blinks or eyelid closures in this paper. As preprocessing, we applied a median filter with 0.5 sec window to the data to suppress spontaneous noises and to interpolate short defects by eye blinks. We also exclude the remaining defects in the data caused by eyelid closures from analysis, which constituted 23.6% of the total data.

The minimum and maximum number of components were set to $C_{\min} = 2$ and $C_{\max} = 8$, respectively, during the extraction of saliency patterns from the saliency maps. In addition, the strength of regularization for tracking, β in Eq. (2), was determined not to

[3] MITSUBISHI Diamondcrysta RDT262WH, 25.5 inch, W550 mm/H344 mm.

[4] Tobii X60 Eye Tracker. An approximate allowed range of head motion is 400×220×300mm.

Table 1: *Estimation results.* $\mathrm{M}_{\mathrm{dur}}$*: the baseline method with the gaze-duration feature,* $\mathrm{M}_{\mathrm{com}}$*: the baseline method with the components clustering, and* $\mathrm{M}_{\mathrm{pro}}$*: the proposed method.*

Method	$\mathrm{M}_{\mathrm{dur}}$	$\mathrm{M}_{\mathrm{com}}$	$\mathrm{M}_{\mathrm{pro}}$
Accuracy (%)	59.3	68.9	**80.6**

generate intervals shorter than 0.5 sec. In the learning method described in Subsection 4.3, we employed all of the 12 videos for the training data, and the number of modes was determined as 7.

Evaluation

120 data for each trial (totally 240 data), which consist of two levels of attentiveness per each video, were obtained. We applied leave-one-out cross validation (LOOCV) to obtain the estimation accuracy. In order to exclude an order effect on watching of the same video under the opposite condition, LOOCV is applied to each of the trials. Namely, each of the validations we remove one of the 120 data to learn local saliency dynamics for all the scenes using the rest of the data, and test for all the frames in the removed data to be classified correctly. Thus, the accuracy is obtained as a ratio of correctly-classified frames to all the tested frames.

Two different estimation methods, $\mathrm{M}_{\mathrm{dur}}$ and $\mathrm{M}_{\mathrm{com}}$, were used to serve as baselines. $\mathrm{M}_{\mathrm{dur}}$ utilizes gaze durations toward any salient regions as a feature. Specifically, since the salient regions are modeled as a Gaussian component, we utilize the minimum distance between gaze points and component means, $\min_{s_t^{(c)}} \xi(s_t^{(c)})$ in Eq. (7), to judge whether subjects gaze at the regions. The gaze-duration feature is obtained as the ratio of intervals where the distances are within a threshold; the threshold was set to 4° empirically with the consideration of the screen size, in a 0.5 sec sliding window. The level of attentiveness is estimated by applying the Fisher's discriminant analysis to the features.

The other baseline $\mathrm{M}_{\mathrm{com}}$ verifies the efficiency of scene description by MMSDM. $\mathrm{M}_{\mathrm{com}}$ includes a frame-wise clustering of parameter changes of Gaussian components ($\boldsymbol{\theta}_t^{(c)} - \boldsymbol{\theta}_{t-1}^{(c)} \in \mathbb{R}^6$ in Subsection 3.1) by applying GMM. The number of clusters is set to 7, the same number as that of modes in the proposed method. $\mathrm{M}_{\mathrm{com}}$ follows the estimation scheme shown in Section 5. That is, the scenes and the local saliency dynamics in $\mathrm{M}_{\mathrm{com}}$ are described by using a set of clusters instead of a set of modes.

6.2 Results and discussions

Table 1 shows the accuracy of the estimations, the ratio of correctly-classified frames to all the tested frames. The results demonstrate that the proposed method $\mathrm{M}_{\mathrm{pro}}$ can work well even when the baselines have no clear discrimination of the two attentiveness levels.

The comparison of $\mathrm{M}_{\mathrm{dur}}$ with $\mathrm{M}_{\mathrm{pro}}$ and $\mathrm{M}_{\mathrm{com}}$ shows the effectiveness in handling saliency dynamics when analyzing eye movements. Regarding $\mathrm{M}_{\mathrm{pro}}$ and $\mathrm{M}_{\mathrm{com}}$, the common scheme underlying these methods is to learn eye-movement features (i.e., the local saliency dynamics) conditioned by scenes, and the difference is how to model the scenes. The proposed method models scenes by a set of modes. The number of clusters in $\mathrm{M}_{\mathrm{com}}$ is set to the same as that of the modes for verification, and the results indicate that the proposed method can capture the scenes more efficiently.

The scenes in the proposed method have specific modes that perform effectively/ineffectively for the estimation. Figure 5 depicts examples of distance distributions as well as saliency maps and their scenes. Distance distributions from scene (a) show that all the modes in its scene can contribute the estimation to some extent, however, the distributions from (b) and (c) demonstrate that the mode 1 and 3 have little difference between the states (i.e., high and low) and their contribution in the scene is limited. In addition, the scene with modes $\{1, 3, 3\}$ has no modes that contribute the estimation as shown by distributions from scene (d). In such cases, we can "prune" the useless modes from the scene and can describe the scene more simply. For instance, the scene $\{1, 3, 4, 5\}$ can be replaced by the scene $\{4, 5\}$ with regard to scene (b). This mode pruning will reduce the number of scenes to be considered, and provide more training data for some scenes.

Modeling of saliency maps by GMM can describe saliency dynamics such as motions, shape deformations and temporal saliency variations parametrically. However, this modeling has a side-effect that the Gaussian components are inescapably fitted to the locations wherever the strength of saliency is higher than surroundings. As a result, the intuitive numbers of salient regions do not always agree with the number of components (the scene from saliency map (c) for instance). Generally, scenes with a large number of components have a large variety in mode combinations and may cause an overfitting. More accurate estimation of the number of components is required for the proposed method as well as the mode pruning mentioned above. Besides, it still remains unclear that which parameters in saliency dynamics contribute the estimation. The specification of the contribution of each parameter for analyzing gaze and attentiveness should be addressed in future work.

7 Conclusion

We propose a method to estimate the attentive states of users while they watch general videos such as TV commercials. The multi-mode saliency-dynamics model (MMSDM) is proposed to describe complex saliency dynamics in videos. Since the MMSDM segments the saliency dynamics into multiple mode sequences, it enables us to describe local saliency dynamics around gaze points efficiently with a set of modes. The experimental results demonstrate that the difference in statistically-learned local saliency dynamics can be a crucial clue to estimate the level of attentiveness.

Acknowledgement This work is in part supported by MEXT Global COE program "Informatics Education and Research Center for a Knowledge-Circulating Society".

References

BRANDHERM, B., PRENDINGER, H., AND ISHIZUKA, M. 2007. Interest estimation based on dynamic bayesian networks for visual attentive presentation agents. In *Proc. of ICMI*, 346–349.

BREGLER, C. 1997. Learning and recognizing human dynamics in video sequences. In *Proc. of CVPR*, 568–574.

CALDER, A., LAWRENCE, A., KEANE, J., SCOTT, S., OWEN, A., CHRISTOFFELS, I., AND YOUNG, A. 2002. Reading the mind from eye gaze. *Neuropsychologia 40*, 8, 1129–1138.

DORR, M., MARTINETZ, T., GEGENFURTNER, K., AND BARTH, E. 2010. Variability of eye movements when viewing dynamic natural scenes. *Journal of vision 10*, 10, 1–17.

DOSHI, A., AND TRIVEDI, M. M. 2010. Attention estimation by simultaneous observation of viewer and view. In *Proc. of CVPR Workshop*, 21–27.

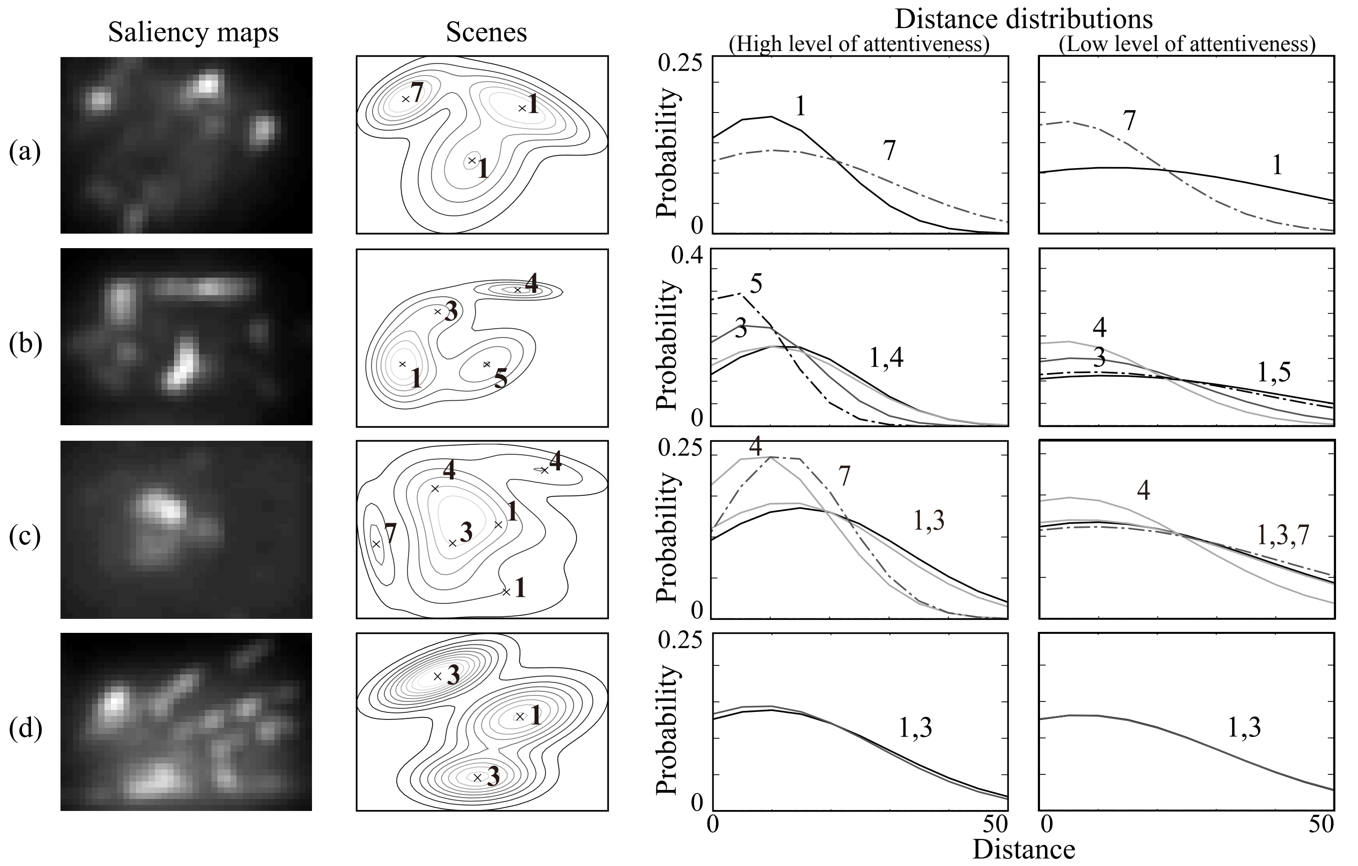

Figure 5: *Saliency maps, obtained scenes, and distance distributions for both high and low level of attentiveness. The numbers in the scenes and distance distributions represent the IDs of modes.*

Fletcher, L., and Zelinsky, A. 2009. Driver inattention detection based on eye gaze-road event correlation. *International Journal of Robotics Research 28*, 6, 774–801.

Goldstein, R. B., Woods, R. L., and Peli, E. 2007. Where people look when watching movies: do all viewers look at the same place? *Computers in Biology and Medicine 37*, 7, 957–64.

Hirayama, T., Dodane, J.-B., Kawashima, H., and Matsuyama, T. 2010. Estimates of user interest using timing structures between proactive content-display updates and eye movements. *IEICE Trans. on Information and Systems E-93D*, 6, 1470–1478.

Itti, L., Koch, C., and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI 20*, 11, 1254–1259.

Ji, Q., Lan, P., and Looney, C. 2006. A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans 36*, 5, 862–875.

Just, M., and Carpenter, P. 1976. Eye fixations and cognitive processes. *Cognitive Psychology 480*, 441–480.

Kahneman, D. 1973. *Attention and effort*. Prentice Hall.

Kawashima, H., and Matsuyama, T. 2005. Multiphase learning for an interval-based hybrid dynamical system. *IEICE Trans. on Fundamentals E88-A*, 11, 3022–3035.

Li, Y., Wang, T., and Shum, H.-Y. 2002. Motion texture: a two-level statistical model for character motion synthesis. *ACM Trans. on Graphics 21*, 3, 465–472.

Munn, S. M., Stefano, L., and Pelz, J. B. 2008. Fixation-identification in dynamic scenes. In *Proc. of APGV*, 33–42.

Nakano, Y. I., and Ishii, R. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proc. of IUI*, 139–148.

North, B., Blake, A., Isard, M., and Rittscher, J. 2000. Learning and classification of complex dynamics. *IEEE Trans. on PAMI 22*, 9, 1016–1034.

Pavlovic, V., Rehg, J. M., and MacCormick, J. 2000. Learning switching linear models of human motion. In *Proc. of NIPS*, 981–987.

Qvarfordt, P., and Zhai, S. 2005. Conversing with the user based on eye-gaze patterns. In *Proc. of CHI*, 221–230.

Schneider, W., and Shiffrin, R. M. 1977. Controlled and automatic human information processing: I. detection, search, and attention. *Psychological Review 84*, 1–66.

Walther, D., and Koch, C. 2006. Modeling attention to salient proto-objects. *Neural Networks : the Official Journal of the International Neural Network Society 19*, 9, 1395–1407.

Wildes, R. P. 1998. A measure of motion salience for surveillance applications. In *Proc. of ICIP*, 183–187.