Regular Paper

Mental Focus Analysis Using the Spatio-temporal Correlation between Visual Saliency and Eye Movements

Ryo Yonetani[1,a)] Hiroaki Kawashima[1] Takatsugu Hirayama[1,†1] Takashi Matsuyama[1]

Received: April 11, 2011, Accepted: September 12, 2011

Abstract: The spatio-temporal correlation analysis between visual saliency and eye movements is presented for the estimation of the mental focus toward videos. We extract spatio-temporal dynamics patterns of saliency areas from the videos, which we refer to as saliency-dynamics patterns, and evaluate eye movements based on their correlation with the saliency-dynamics patterns in view. Experimental results using TV commercials demonstrate the effectiveness of the proposed method for the mental-focus estimation.

Keywords: mental focus, spatio-temporal analysis, saliency map, eye movement

1. Introduction

Understanding mental states of users, who face interactive display systems to browse various kinds of information, allows building a smooth interaction between users and systems. Our goal is to estimate the *mental focus* of users from their eye movements, which indicates whether they pay attention to a specific task. Especially in this paper, we focus on the estimation of the mental focus while users watch general videos such as TV commercials; that is, we consider the mental focus as the strength of users' attention toward videos, and besides, we assume the strength can be quantified into several levels (e.g., high and low).

Eyes are a window into the mind; eye movements are often regarded as one of the crucial clues to estimate user states such as the attention [1], [2], [3]. Analysis of the eye movements includes eye blinks [1], PERCLOS (PERcentage of eyelid CLOSure) [4], or their combination using Bayesian networks [5]. The basic concept behind these studies is that eye movements can be affected by the user states as well as contents being looked at. And there the analysis of relationships between contents and eye movements plays a key role. Existing work can be characterized by what kind of relationships the method uses. For example, many studies on interactive systems basically begin the analysis by specifying objects being looked at, and then extract the features such as the gaze duration [6], 3-gram sequence of gaze targets [7], or the reaction time to dynamic content updates [8]. In order to obtain the detailed psycho-cognitive processes, the consideration of content semantics is furthermore required [9]. Some works on driving assistance systems investigate the correlation between eye gazes and surrounding environments. They look for salient objects from surrounding environments using optical flow [2] or obstacle, sign and pedestrian detection [3], and analyze the relationship between gaze directions and the object positions to estimate the drivers attention.

Those related works basically specify gaze-related objects or their semantic relationships based on some heuristics about contents or surrounding environments in advance. However, because general videos contain enormous kinds of objects, it is difficult to specify all the objects and their relationships using the heuristics applied in the related works. Besides, eye movements have a large variety of dynamics caused by contents and users' states. Therefore, the analysis of eye movements detailed enough to realize the mental-state estimation essentially requires (1) the spatial structure that describes gaze targets and the other surrounding objects in view, and (2) the temporal structure of dynamics between both eye movements and the targets.

In this paper, we propose a novel method for the mental-focus analysis that utilizes the spatio-temporal relationship of dynamics in both objects and eye movements. The main contribution is to introduce an analysis of eye-movement dynamics by categorizing them from the aspect of objects' dynamics in view. This enables us to switch an appropriate feature set of eye movements according to objects' dynamics, and is expected to enhance the accuracy of the estimation of mental focus. In order to analyze dynamics in displayed contents without using specific semantic heuristics, we employ the saliency map [10], which is known as a model of a visual attention system. With consideration of the influence on eye movements, we classify the dynamics of extracted saliency areas into several patterns called *saliency-dynamics patterns*, which specify corresponding eye movements and their features. Thus, once a saliency-dynamics pattern is identified from contents data in a certain time window, we can evaluate the eye movements to estimate the level of the mental focus.

The overview of our proposed analysis is as follows. As shown in **Fig. 1** (a), we first extract the spatio-temporal saliency volumes $\mathcal{S}$ referred to as *saliency flows*. $\mathcal{S}$ are regarded as the candidates

[1] Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[†1] Presently with Nagoya University
[a)] yonetani@vision.kuee.kyoto-u.ac.jp

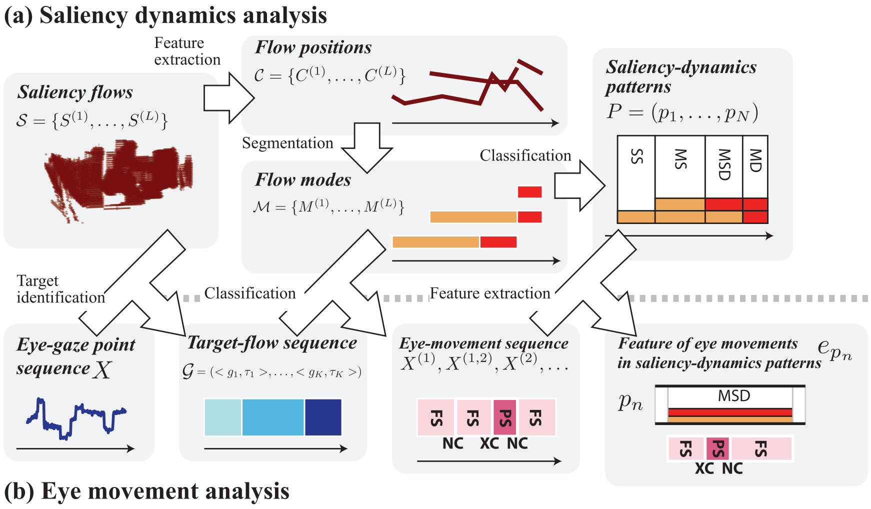
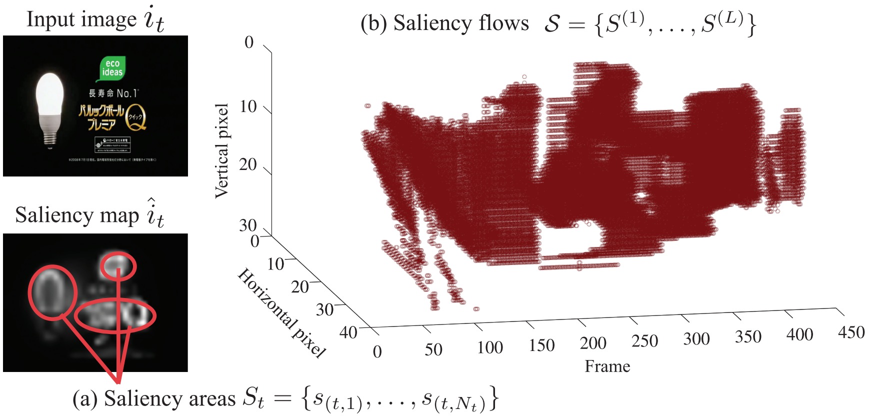

(a) Saliency dynamics analysis
Feature extraction
Flow positions
$\mathcal{C} = \{C^{(1)}, \ldots, C^{(L)}\}$
Saliency flows
$\mathcal{S} = \{S^{(1)}, \ldots, S^{(L)}\}$
Segmentation
Flow modes
$\mathcal{M} = \{M^{(1)}, \ldots, M^{(L)}\}$
Classification
Saliency-dynamics patterns
$P = (p_1, \ldots, p_N)$
SS
MS
MSD
MD
Target identification
Classification
Feature extraction
Eye-gaze point sequence X
Target-flow sequence
$\mathcal{G} = (< g_1, \tau_1 >, \ldots, < g_K, \tau_K >)$
Eye-movement sequence
$X^{(1)}, X^{(1,2)}, X^{(2)}, \ldots$
FS
FS
PS
FS
NC XC NC
Feature of eye movements in saliency-dynamics patterns e_{p_n}
p_n
MSD
FS
PS
FS
XC NC
(b) Eye movement analysis

Fig. 1 Overview.



(a) Saliency areas $S_t = \{s_{(t,1)}, \ldots, s_{(t,N_t)}\}$

Fig. 2 Saliency extraction. A saliency flow $S^{(l)}$ is composed by temporally-continuous saliency areas with the same id l: $\{s_{(t,n)} \mid l_{(t,n)} = l\}$.

of objects to be looked at. We also represent the flow positions by their centroid sequences $\mathcal{C}$. Then, analyzing the change of *modes*, such as static and dynamic, occurred in each sequence in $\mathcal{C}$, we segment the flows into mode sequences $\mathcal{M}$. And from $\mathcal{M}$, we finally achieve a sequence of saliency-dynamics patterns P (see Section 2 for details). For eye-movement analysis (see Fig. 1 (b)), eye-gaze data is first segmented into a gaze-target sequence $\mathcal{G}$ using $\mathcal{S}$, and classified into several types of eye movements using their correlation with $\mathcal{M}$ (see Section 3 for details). Features that imply mental focus are then extracted from eye movements $X^{(1)}, X^{(1,2)}, \ldots$ differently according to their types. We bundle the features from eye movements within $p_n \in P$ into $\boldsymbol{e}_{p_n}$, and employ different discriminative models of mental focus for each saliency-dynamics pattern. In this research we discriminate two levels of mental focus, high and low, by switching the models based on the observed patterns (see Section 4 for details).

2. Video Saliency Extraction and Analysis

2.1 Visual Saliency in a Video

Videos have visual saliency areas which attract human eye gazes. We employ the saliency map [10] to obtain the saliency areas in a video, and utilize the areas as objects to be looked at. The saliency map is a bottom-up computational model of visual attention, which typically includes the extraction of multiple low-level visual features such as the intensity, the color, the orientation from an image at multiple scales, normalization and integration of features into a 2D map with a saliency value at each pixel. Studies on visual-attention systems such as the saliency map basically aim to evaluate their model by predicting human gaze behaviors [11], [12], with no consideration of mental states. On the other hand, the proposed method aims to analyze the relationship between gaze behaviors and the obtained saliency map, in order to estimate the mental focus.

We extract saliency areas from a video [*1]; a saliency map $\hat{i}_t$ is computed from an input frame i_t at the time t and each pixel $\boldsymbol{c} \in \mathbf{N}^2$ is given a saliency value $\hat{i}_t(\boldsymbol{c})$. $\hat{i}_t$ is thresholded at π_s and the remaining pixels are segmented into a set of saliency areas $S_t = \{s_{(t,1)}, \ldots, s_{(t,N_t)}\}$ by 8-connectivity labeling (see **Fig. 2** (a)).

2.2 Saliency Flow Construction

As mentioned in Section 1, the spatio-temporal saliency dynamics is utilized for eye-movement analysis. So we extend the visual saliency into spatio-temporal volumes referred to as saliency flows (see Fig. 2 (b)). The saliency flows are defined as simply-connected 3D volumes composed by temporally-continuous saliency areas, and they contain the time-varying pat-

[*1] The implementation of saliency extraction is in MATLAB using the Saliency Toolbox [13].

tern of their shape, position, and saliency value. Videos are often expected to contain several flows, so we assign an ID to each constructed flow. The following procedure gives an ID $l_{(t,n)}$ to a saliency area $s_{(t,n)}$ so that we bundle a set of saliency areas into a single flow.

For the area $s_{(t,n)}$, we look for a set of area $\hat{S}_{t-1}$ from S_{t-1}, the elements of which are spatio-temporally continuous to $s_{(t,n)}$. Since saliency flows are defined as simply-connected, branches in the flows should be avoided. Such branches are formed where multiple areas are connected to $s_{(t,n)}$, so we assigned $l_{(t,n)}$ based on the number of elements in $\hat{S}_{t-1}$ (denoted as $Card(\hat{S}_{t-1})$; the cardinality of $\hat{S}_{t-1}$):

(a) If $Card(\hat{S}_{t-1}) = 0$; no area can be found in $\hat{S}_{t-1}$, $s_{(t,n)}$ consists of the saliency flow which emerges at t. Then a new ID number is given to $l_{(t,n)}$.

(b) If $Card(\hat{S}_{t-1}) = 1$; a single area can be found in $\hat{S}_{t-1}$, $s_{(t,n)}$ is the following area to $s_{(t-1,m)} \in \hat{S}_{t-1}$.
This case $l_{(t,n)}$ receives the ID $l_{(t-1,m)}$.

(c) If $Card(\hat{S}_{t-1}) \geq 2$; more than one area can be found in $\hat{S}_{t-1}$, $s_{(t,n)}$ is the area with the collision of multiple flows. For this case $l_{(t,n)}$ receives the ID $l_{(t-1,\hat{m})}$ of area $s_{(t-1,\hat{m})} \in \hat{S}_{t-1}$ which locates the nearest position to $s_{(t,n)}$.

Once $l_{(t,n)}$ is assigned, we look for areas which have the same ID number in $S_t = \{s_{(t,1)}, \dots, s_{(t,N_t)}\}$. If any area exists, $s_{(t,n)}$ is one of the branches which emerges at t. In this case $l_{(t,n)}$ is relabeled to a new ID number.

Let us assume that the maximum ID L is given by the above procedure, and the ID $l_{(t,n)} \in \{1, \dots, L\}$ is defined for the area $s_{(t,n)}$. Besides, let us define here the function that returns the ID from a single pixel $\boldsymbol{c}$ at the time t as

$$SID_t(\boldsymbol{c}) = l \quad (l \in \{1, \dots, L, \zeta\}), \tag{1}$$

where ζ denotes the state in which no flow exists at $\boldsymbol{c}$. A saliency flow labeled as l is composed by a set of saliency areas $S^{(l)} \triangleq \{s_{(t,n)} \mid l_{(t,n)} = l\}$. This flow exists in the temporal interval $[b_l, e_l]$, and the area covered with $S^{(l)}$ at the time t is represented as $S_t^{(l)}$.

2.3 Saliency Dynamics Analysis

A set of saliency flows $\mathcal{S} = \{S^{(1)} \dots, S^{(L)}\}$ represents the saliency dynamics. Each element of $\mathcal{S}$, a saliency flow $S^{(l)}$, contains the time varying pattern of its shape, position and saliency value. Above all, a motion is well known as one of the important features to enhance the saliency [14]. The motion is therefore expected to affect an attractiveness of saliency flows, and it is also expected to affect eye-movement dynamics; the saliency flows with or without motion cause different types of the eye movements (see Section 3 for details).

So we extract motion features of saliency flows, and classify motion patterns of saliency dynamics using their motion features for eye-movement analysis. The flow position $\boldsymbol{c}_t^{(l)}$ using the centroid of $S_t^{(l)}$ is defined as follows:

$$\boldsymbol{c}_t^{(l)} = \frac{1}{Card(S_t^{(l)})} \sum_{\boldsymbol{c} \in S_t^{(l)}} \boldsymbol{c}. \tag{2}$$

A position sequene $C^{(l)} \triangleq (\boldsymbol{c}_t^{(l)} \mid t \in [b_l, e_l])$ is obtained from $S^{(l)}$.

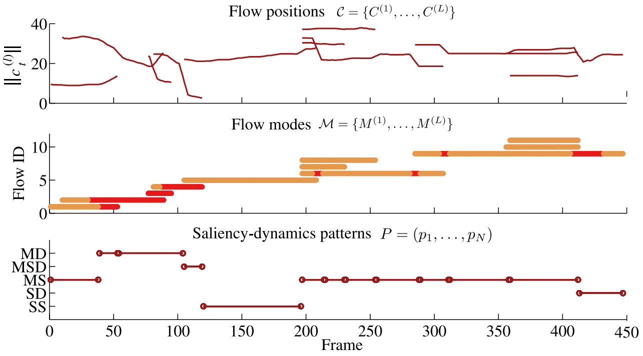


Fig. 3 Saliency-flow segmentation and classification. In the top of the figure, red lines show the Euclidean norm of flow positions. In the middle, red lines show dynamic modes and orange lines show static. In the bottom, SS shows Single Static, SD shows Single Dynamic, MS shows Multi Static, MSD shows Multi Static/Dynamic, and MD shows Multi Dynamic. π_c was defined as 2°/s and ω_d as 2° using the size of human central visual field. w_s is empirically defined as 0.1 sec and ω_p as 0.5 sec.

Table 1 Saliency-dynamics patterns and their specifications.

Saliency-dynamics pattern	Specifications
Single Static (SS)	Sole flow exists with static mode.
Single Dynamic (SD)	Sole flow exists with dynamic mode.
Multi Static (MS)	Multiple flows exist, and they are all static.
Multi Static/Dynamic (MSD)	Multiple flows exist. Some of them are static and the others are dynamic.
Multi Dynamic (MD)	Multiple flows exist, and they are all dynamic.

The following procedures transform $C^{(l)}$ into a sequence of *modes* (motion states) $M^{(l)}$, and classify saliency dynamics as several patterns based on the modes (see **Fig. 3**).

Flow segmentation

Two modes: m_{d} and m_{s}, where m_{d} represents that a flow has a motion, and m_{s} shows it is static, are introduced for saliency-flow segmentation. We use the notation $< m_n^{(l)}, \tau_n^{(l)} >$ to represent the interval that has a mode $m_n^{(l)} \in \{m_{\mathrm{d}}, m_{\mathrm{s}}\}$ and duration $\tau_n^{(l)} (\sum_n \tau_n^{(l)} = e_l - b_l + 1)$. We first set threshold to the speed (the first order differential value) of $\|C^{(l)}\|$ at π_c, and segment it into an initial mode sequence $(< m_1^{(l)}, \tau_1^{(l)} >, \dots, < m_{N_l}^{(l)}, \tau_{N_l}^{(l)} >)$. Static modes with smaller intervals than w_s are then suppressed by merging them with subsequent dynamic modes. Dynamic modes with a smaller motion than an amplitude ω_d are also merged with subsequent static modes because such small modes cause no eye motions. Finally we renew the subindices to obtain the mode sequence $M^{(l)} = (m_1^{(l)}, \dots, m_{N_l}^{(l)})$.

Pattern classification in saliency dynamics

Saliency-dynamics patterns describe characteristics of the spatio-temporal structure of saliency dynamics in videos. We consider video scenes with and without dynamic saliency flows separately because the dynamic flows originally tend to attract a more attention than the static ones. Besides, we take the number of flows into account because it affects gaze distributions. Thus, from the existence and modes of the saliency flows, the patterns consisting of Single Static, Single Dynamic, Multi Static, Multi-Static/Dynamic, and Multi Dynamic, are formed by a set of mode sequences $\mathcal{M} = \{M^{(1)}, \dots, M^{(L)}\}$ (see **Table 1** for their specifications).

$\mathcal{M}$ is first segmented based on changes in the number of flows and of their modes. As well as the preceding flow segmentation, we examine the duration of each segment and merge small segments with subsequent ones with the threshold ω_p. A sequence of saliency-dynamics patterns is finally acquired: $P = (p_1, \ldots, p_N)$ $(p_n \in \{\mathrm{SS}, \mathrm{SD}, \mathrm{MS}, \mathrm{MSD}, \mathrm{MD}\})$, from an input video.

3. Eye Movements Analysis Using Saliency Dynamics

3.1 Mental Focus and Eye Movements

Mental focus is here defined as a state that specifies whether humans pay attention to video viewing tasks, and we assume the state can be quantified into several levels. Kahneman proposed the attention theory that likens attention to a limited resource which is allocated to tasks [15]. Following this theory, the level of the mental focus can be regarded as the amount of attention resource that allocates to the tasks. Besides, human information processing can be classified into two types [16]: controlled processing is driven by human intentions, and automatic processing is on the other hand a passive outcome of visual stimuli. We define the eye movements in a controlled mode as *endogenous* eye movements, and in an automatic mode as *exogenous* eye movements.

Based on these two theories, we assume the following mental focus and eye movements relationship model. From mental factors such as intentions or physiological factors such as fatigues, the level of mental focus is determined and attention resource is allocated. This attention resource causes endogenous eye movements, and some visual stimuli in saliency dynamics cause exogenous eye movements as well. Eye gaze data is observed by mixing these eye movements.

Eye movements have different features according to their type, and therefore have to be evaluated in a different way. The first step toward the estimation is to classify the eye-movement type using saliency dynamics. After that we extract features from the eye movements differently according to their type.

3.2 Eye Movements in Video Viewing

We summarize the kinds of eye movements being observable on a screen, which take place during video viewing, so that we classify them (see **Fig. 4**). When we humans watch videos, iterative scanning and selection of objects are expected to be observed. Here, we employ the following primitive eye movements from a biomedical research [17]: saccade, fixation and smooth pursuit, and try to describe characteristics of the scanning and the selection using combinations of the primitive movements.

The scanning movements present steady behaviors, and therefore are characterized by flow modes. Based on the modes, they are classified into two types:

Fixation scan (FS) is a scanning eye movement of static saliency flows. FSs can be described as the combination of fixations: maintenance of gaze on a single location, and saccades: rapid and jerky gaze shifts.

Pursuit scan (PS) is a scanning of moving (dynamic) saliency flows. PSs contain smooth pursuits of the flow motions, and saccades.

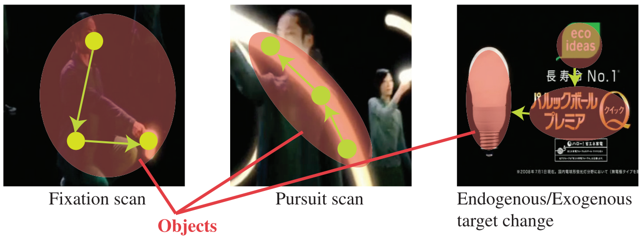


Fig. 4 Eye movements in video viewing.

On the other hand, the selection movements present transitional behaviors, and are caused voluntarily as well as by visual stimuli in saliency flows. Therefore, they can be classified based on the type of human information processing. We introduce the term *events* that represent visual stimuli in the flows. Events are defined here as emergences or mode transitions from static to dynamic. They cause exogenous gaze shifts, and according to the their association with target selections, the following eye movements are introduced:

Endogenous target change (NC) is a gaze shift between saliency flows that occurs asynchronously with events. Using NCs, humans voluntarily determine which flows to scan next, and manage to shift their gaze.

Exogenous target change (XC) is also a gaze shift between saliency flows but it occurs in synchronization with events in a destination flow. The feature of XCs is that humans do not consciously shift their gaze.

3.3 Eye Movements Classification

Let us assume that eye-gaze data $\boldsymbol{x}_t \in \mathbf{R}^2$, a 2D point on a screen is obtained using an eye tracker. The four eye movements mentioned above are specified from the observed sequence of eye-gaze points $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots)$ (see **Fig. 5**).

Segmentation

For the classification, we first produce intervals $< g_k, \tau_k >$ from X, each of which has a single target flow with ID g_k and duration τ_k. By following Eq. (1), the reference from a point $\boldsymbol{x}_t$ to a flow ID is described as follows:

$$g(t) = SID_t(\boldsymbol{x}_t) \quad \in \{1, \ldots, L, \zeta, \eta\}. \tag{3}$$

Notice that the numbers $1, \ldots, L$ denote the flow IDs, ζ shows that no flow locates at the point, and η shows that the human blinks or looks outside of a screen. Using target-flow changes, we obtain a target-flow sequence with duration $\mathcal{G} = (< g_1, \tau_1 >, \ldots, < g_K, \tau_K >)$, which consists of $g_k \notin \{\zeta, \eta\}$ (see Fig. 5 (a)). We then divide $< g_k, \tau_k >$ based on mode transitions in the flow $S^{(g_k)}$; $< g_k, \tau_k >$ is divided into $(< m_{k_1}, \tau'_{k_1} >, \ldots, < m_{k_M}, \tau'_{k_M} >)$ by a set of mode-transition time in $S^{(g_k)}$. After the division, we renew subindices to get a sequence of target-flow mode with duration: $\mathcal{G}' = (< m_1, \tau'_1 >, \ldots, < m_{K'}, \tau'_{K'} >)$ (see Fig. 5 (b)).

Classification

We classify the eye movements using the obtained mode sequence $\mathcal{G}'$. Let us assume that $< m_k, \tau'_k >$ exists in the interval $[b_k, e_k]$. We describe a sequence of eye-gaze points in $< m_k, \tau'_k >$ as $X^{(k)} \triangleq \{\boldsymbol{x}_t | t \in [b_k, e_k]\}$ and of eye-gaze points between $< m_{k-1}, \tau'_{k-1} >$ and $< m_k, \tau'_k >$ as $X^{(k-1,k)} \triangleq \{\boldsymbol{x}_t | t \in [e_{k-1}, b_k]\}$, which will individually correspond to specific eye movements.

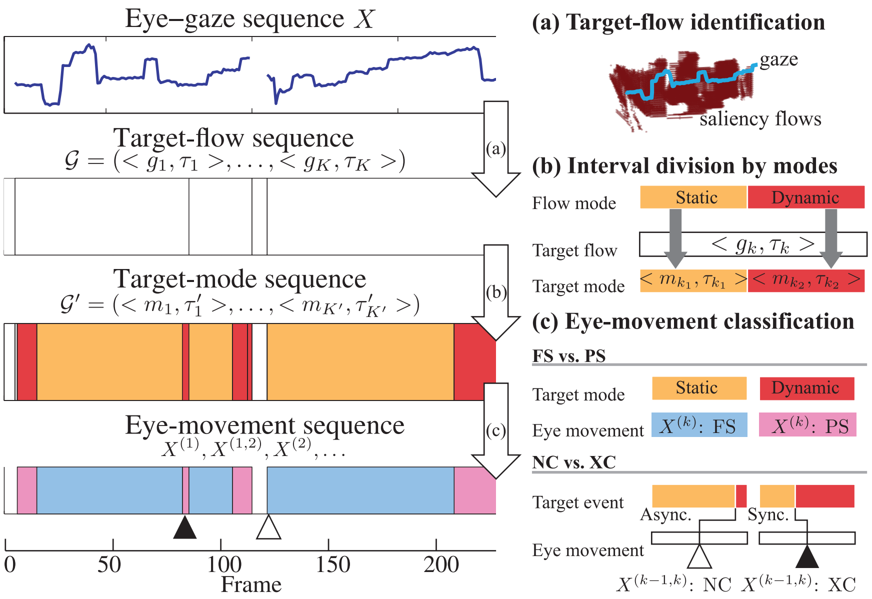


Fig. 5 Eye-movement classification. (a) Target-flow identification and initial segmentation. (b) Interval division based on target modes. (c) Eye-movement classification.

FSs and PSs take place within the interval $< m_k, \tau'_k >$ and $X^{(k)}$ can be classified to either of them. These eye movements can be classified using modes of their gaze targets; based on the target-flow mode m_k, we assign the label FS to $X^{(k)}$ if $m_k = m_\mathrm{s}$ and PS if $m_k = m_\mathrm{d}$.

On the other hand, NCs and XCs take place between two intervals $< m_{k-1}, \tau'_{k-1} >$ and $< m_k, \tau'_k >$. As mentioned in Section 3.2, these two movements are discriminated by observing whether the eye movement is synchronized with *events* or not. Specifically, we evaluate the temporal distance between the starting time of the eye movements and that of the corresponding events to discriminate the two movements. Let us assume that the most recent event from b_k occurs at $T_k (< b_k)$ in NC/XC destination flows. NCs are then discriminated from XCs using temporal distance between T_k and b_k; if $b_k - T_k$ is within the reaction time σ of exogenous saccades (generally around 0.2 sec [18]), $X^{(k-1,k)}$ can be labeled as XC, and otherwise it is labeled as NC. Finally we obtain the sequence of eye movements, $X^{(1)}, X^{(1,2)}, X^{(2)}, \ldots, X^{(K-1)}, X^{(K-1,K)}, X^{(K)}$, labeled by FS, PS, NC, or XC (see Fig. 5 (c)).

4. Mental Focus Estimation

4.1 Estimation Overview

Following the procedure in Section 3, a sequence of eye movements, consisting of fixation scan FS, pursuit scan PS, endogenous target change NC and exogenous target change XC, is obtained. This section presents the feature extraction of the eye movements and the estimation of the mental focus. We assume that the level of mental focus corresponds to the amount of attention resource to tasks, and that humans scan or select saliency flows more actively using the resource when they are in a higher level of mental focus.

Eye movements have different characteristics according to their type, and thus we first extract features which indicate activeness in video viewing, from the observed eye movements in a different way for their type. Since we focus on dynamic aspects in eye movements, the features are extracted as summarizations of their dynamics. Here, the eye movements are affected not only by a target flow but by the other surrounding flows in a scene. For instance, scenes composed of multiple flows cause NC and XC whereas scenes with a single flow cause only FS or PS. In addition, when the scene includes both static and dynamic flows, PS seems to be observed more often than FS because the dynamic flows are more salient than the static ones. That is, the features can perform differently according to the type of saliency-dynamics patterns, which indicate the number and modes of existing flows.

Table 2 Saliency-dynamics patterns and observable eye movements. The detailed specifications of the patterns in the left column can be found in Table 1.

Saliency-dynamics pattern	Observable eye movements
Single Static (SS)	FS
Single Dynamic (SD)	PS
Multi Static (MS)	FS, NC, XC
Multi Static/Dynamic (MSD)	FS, PS, NC, XC
Multi Dynamic (MD)	PS, NC, XC

Therefore, we then integrate the features, which are extracted from the eye movements in a certain time window defined by the saliency-dynamics patterns, into a feature set. The observable eye movements for each of the patterns can be derived as **Table 2**, since the eye movements are categorized by their association with modes or events of saliency flows. By learning a discriminative model from the feature sets for each pattern, we can estimate the level of mental focus by switching an appropriate model to the observed eye movements based on the corresponding saliency-dynamics patterns in view.

4.2 Feature Extraction

As shown in Section 3.2, FSs and PSs are steady eye movements and they have internal dynamics. With regard to FSs, they usually contain saccades as gaze shifts. As seen in a study on the scene perception [19], the saccades are regarded as crucial features in eye movements. We suppose that such saccades occur more actively when humans are in higher level of mental focus, and introduce a stroke length e_{fs1} and a frequency e_{fs2} of saccades as features of FSs. Let us use the notation $\dot{X}^{(k)}$ for the velocity of an eye-motion pattern $X^{(k)}$. $\|\dot{X}^{(k)}\|$ is thresholded at π_v, so that N_v instances of partial temporal intervals $\{O_1, \ldots, O_{N_v}\}$ which contain saccades are detected. For each partial eye-motion pattern $X_v^{(k)} \triangleq (\boldsymbol{x}_t \mid t \in O_v)$, we calculate the stroke length of saccades and then get the average value as e_{fs1}:

$$e_{\mathrm{fs1}} = \frac{1}{N_v}\sum_{v=1}^{N_v}\frac{1}{\sqrt{a_v}}\left(\max_{(i,j)\in O_v}\left(\|X_i^{(k)} - X_j^{(k)}\|\right)\right), \qquad (4)$$

where a_v denotes the area size of a target flow at the time O_v starts. $e_{\mathrm{fs1}}^{(v)}$ is normalized by $\sqrt{a_v}$ because the length of saccades seems to depend on the target size. e_{fs2} is defined as $e_{\mathrm{fs2}} = N_v/\tau'_k$, where τ'_k denotes the duration of X_k.

Meanwhile for PSs, we consider the synchronization between eye movements and target flows. When humans track a moving object, they tend to synchronize the pursuit acceleration to expected changes of the target movement, and maintain the velocity at a constant level while no change of target velocity is expected [20]. The feature of PSs therefore contains synchronous components in the speed of eye movements. Such components lie in $\|\dot{X}_t^{(k)}\|\cos\theta_t \left(\cos\theta_t = \frac{\dot{X}_t^{(k)}\cdot\dot{C}_t^{(k)}}{\|\dot{X}_t^{(k)}\|\|\dot{C}_t^{(k)}\|},\ t \in [b_k, e_k]\right)$, an orthographically-projected component of $\dot{X}^{(k)}$ to the corresponding part of a target motion velocity denoted as $\dot{C}^{(k)}$. We introduce the feature e_{ps1} that indicates the synchronization by using the average ratio of speed between eyes and targets:

$$e_{\mathrm{ps1}} = \frac{1}{\tau'_k}\sum_{t\in[b_k,e_k]}\frac{\|\dot{X}_t^{(k)}\|\cos\theta_t}{\|\dot{C}_t^{(k)}\|}. \qquad (5)$$

PSs contain saccadic components as well, and we suppose that such components mainly lie in the rest of information $\|\dot{X}_t^{(k)}\|\sin\theta_t$. The feature e_{ps2} which includes the saccadic components is given by the following equation:

$$e_{\mathrm{ps2}} = \frac{1}{\tau'_k}\sum_{t\in[b_k,e_k]}\left|\|\dot{X}_t^{(k)}\|\sin\theta_t\right|. \qquad (6)$$

NCs and XCs are transitional eye movements and therefore contain no internal dynamics in themselves. Following the saccade evaluation above, we focus on the occurrence frequency of the NCs. Given that K instances of NCs occur during the interval L defined by a single saliency-dynamics pattern consisting of F flows, the feature e_{nc} is defined as $e_{\mathrm{nc}} = K/(L \cdot F)$. With regard to XCs, we calculate the reaction time between the starting time of XCs and that of events which associate with the XCs, since exogenous saccades are featured by the synchronization with the events. Given that K instances of XCs occur with the reaction time r_k ($r_k = b_k - T_k$ in Section 3.3) in a pattern, the feature e_{xc} is defined as $\sum_{k=1}^{K} r_k/K$.

4.3 Integration and Estimation

The next step is the integration of features within saliency-dynamics patterns; multiple features obtained from the observable eye movements are combined into one feature set for each pattern. In what follows, $\boldsymbol{e}_p = [e_1, \ldots, e_N]$ represents the feature set for a saliency-dynamics pattern p, the elements of which denote features derived from eye movements observed in the pattern. For instance, if p is MS, the observable eye movements are FS, NC, and XC (see Table 2). So $\boldsymbol{e}_p$ can be represented as $\boldsymbol{e}_p = [e_{\mathrm{fs1}}, e_{\mathrm{fs2}}, e_{\mathrm{nc}}, e_{\mathrm{xc}}]$. Each element of $\boldsymbol{e}_p$ is here normalized into a range of $[0, 1]$. Notice that every type of eye movements can be observed more than once during a saliency-dynamics pattern. Therefore, we first calculate a feature from each eye movement, and then give the average value per each type of feature to the component of $\boldsymbol{e}_p$.

Here we assume that the different levels of mental focus R are described as $\{R_1, \ldots, R_N\}$. The estimation of mental focus is then formulated as a problem to estimate the state $\hat{R} \in \{R_1, \ldots, R_N\}$ which brings to a maximum a posterior probability from a new observation $\boldsymbol{e}_p^*$, such as

$$\hat{R} = \arg\max_{R} P(R|\boldsymbol{e}_p = \boldsymbol{e}_p^*) \propto \arg\max_{R} P(\boldsymbol{e}_p = \boldsymbol{e}_p^*|R)P(R). \qquad (7)$$

Furthermore, in this paper we assume $P(R)$ as a constant, and transform the equation above into $\arg\max_R P(\boldsymbol{e}_p = \boldsymbol{e}_p^*|R)$. We build a discriminative model of mental focus levels from feature sets derived from training data, differently for saliency-dynamics patterns. And for the estimation, we switch the model according to the observed saliency-dynamics patterns.

5. Experiments

5.1 Experimental Setup

We conducted some experiments and estimated the level of mental focus. In these experiments, we aim to discriminate two levels of mental focus: high and low, as a relatively-simplified evaluation. 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 attention, and therefore are expected to include some obvious saliency flows.

Environments and conditions

A subject sat in front of a screen [*2], and an eye tracker [*3] was installed below the screen. The eye-tracking accuracy was, on average, around 0.7°. The spatial distance between the subject and the screen was around 1000 mm, and in these settings eye movements could be observed during experiments.

Since the mental focus specifies an attentional state to video-viewing tasks, we adopt the following two conditions in order to control the level of the mental focus in the experiments:

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

Condition 2 (low level of mental focus) A subject watches a video, and besides he/she does a mental calculation while watching.

For each condition, subjects were asked to orient their gaze to a screen as far as possible. They carried out the tasks in the following sequence: video group A (six out of all the videos)—Condition 1, video group B (the other six videos)—Condition 2,

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

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

video group B—Condition 1, and video group A—Condition 2.

Preprocessing and parameter setting

Eye-gaze data were acquired by the eye tracker at 30 Hz. As preprocessing, we applied a median filter with 0.5 sec window to the data in order to suppress spontaneous noises and to interpolate defects by eye blinks [*4]. The remaining defects of eye-gaze data caused by eyelid closures constituted 23.6% of the total sequences. For the saliency extraction, the parameter π_s in Section 2.1 was empirically defined as 0.1. Saliency areas then covered 31.5% regions of the total video frames, and the ratio of the state in which subjects looked at any saliency areas was 88.7%. For the saliency-dynamics analysis, the parameters π_c, w_d, w_s, w_p (see Section 2.3) were defined as $\pi_c = 2°/\mathrm{s}$, $w_d = 2°$ based on the size of the human central visual field, $w_s = 0.1$ sec and $w_p = 0.5$ sec to avoid creating short fragments of static flows and saliency-dynamics patterns. The reaction time of exogenous saccades δ (see Section 3.3) was defined as 0.2 sec by following Ref. [18], and the threshold for saccade speed π_v (see Section 4.2) was defined as 8°/sec to avoid detecting fixations incorrectly.

5.2 Results and Discussions

Eye movement analysis

We aggregate feature sets for each saliency-dynamics pattern observed in the experiments. Since we used two conditions in the experiments to control the levels of the mental focus, the obtained data consists of two classes (high/low mental focus). Namely, the obtained feature sets constitute two conditional distributions, conditioned by the level of the mental focus. In order to verify the separability of these two classes in terms of linear discrimination, we first apply the linear discriminant analysis (LDA) to the obtained data set, and calculate partial F-Values to figure out which elements of the feature sets contribute.

Figures 6, **7**, **8**, **9** and **10** describe the relative frequency distributions of the data after LDA projection, in order to visualize the separability. **Table 3** shows partial F-Values for each saliency-dynamics pattern. We can find that the importance of features varies according to the saliency-dynamics patterns, and these results also suggest the following aspects of gaze behaviors:

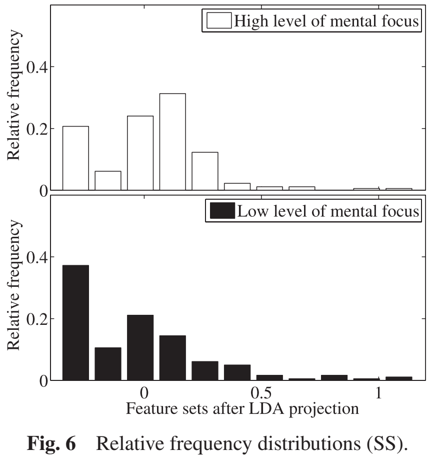


Fig. 6 Relative frequency distributions (SS).

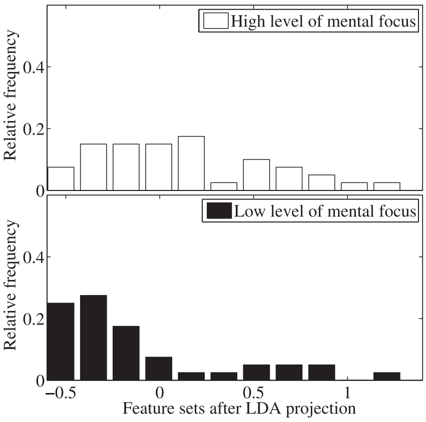


Fig. 7 Relative frequency distributions (SD).

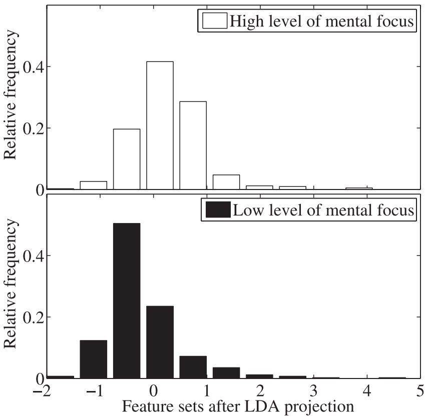


Fig. 8 Relative frequency distributions (MS).

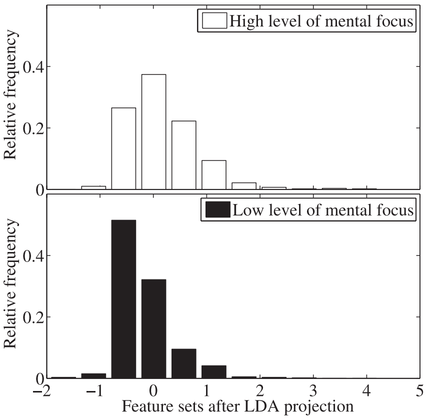


Fig. 9 Relative frequency distributions (MSD).

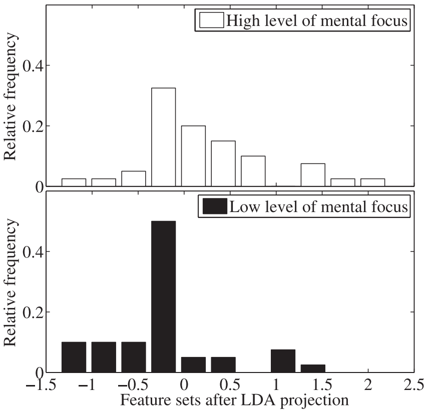


Fig. 10 Relative frequency distributions (MD).

[*4] Since we employ sequences of eye-gaze points on a screen as eye movements, we do not regard the eye blinks especially in this paper, though some studies report on the effectiveness of eye blinks to estimate attention resource [21].

Table 3 Partial F-Values for saliency-dynamics patterns.

Pattern	Parital F-Values
SS	$e_{\rm fs1}$: 1.23, $e_{\rm fs2}$: 3.84
SD	$e_{\rm ps1}$: 0.69, $e_{\rm ps2}$: 3.68
MS	$e_{\rm fs1}$: 34.36, $e_{\rm fs2}$: 41.10 $e_{\rm nc}$: 2.11, $e_{\rm xc}$: 4.11
MSD	$e_{\rm fs1}$: 20.71, $e_{\rm fs2}$: 24.47 $e_{\rm ps1}$: 10.07, $e_{\rm ps2}$: 21.99 $e_{\rm nc}$: 2.43, $e_{\rm xc}$: 1.51
MD	$e_{\rm ps1}$: 0.26, $e_{\rm ps2}$: 5.32 $e_{\rm nc}$: 2.43, $e_{\rm xc}$: 2.04

- The level of mental focus mainly affects saccades which internally occur in FS/PS. We can find the above aspect from the partial F-Value of $e_{\rm fs1}$, $e_{\rm fs2}$ and $e_{\rm ps2}$. Subjects seem to scan target flows actively rather than change them when they are in the high level of the mental focus, since all the NC/XC contributions are much smaller than the others.
- Synchronization in the speed between eye movements and target flows is not affected much by the level of mental focus. That is, subjects basically tend to pursue dynamic saliency flows at any level of mental focus. That is because the partial F-Value of $e_{\rm ps1}$ is relatively smaller than that of $e_{\rm fs1}$ and $e_{\rm fs2}$ in MSD, as well as than that of $e_{\rm nc}$ and $e_{\rm xc}$ in MD.

We can find that the two relative frequency distributions are somewhat separated. Still, some of them are clearly impossible to separate linearly. Therefore, for the estimation, we employ the non-linear discrimination that directly estimates the conditional probability distributions.

Mental focus estimation

Following Eq. (7), we estimate the level of mental focus. 24 data for each subject (totally 240 data), which consist of two levels of mental focus per each video, are obtained. We apply leave-one-out cross validation method to obtain the estimation accuracy for each saliency-dynamics pattern. Namely, we remove one of the 240 data to learn a conditional probability distribution using the rest of the data, and test the removed data to be classified correctly. To interpolate the obtained distributions, here we apply the additive smoothing with the empirically-defined smoothing parameter $\alpha = 0.0001$ to them. We iteratively change the feature set to remove so that we test all the data, and obtain the average accuracy per saliency-dynamics pattern.

Table 4 shows estimation accuracies. The accuracies are obtained for all the saliency-dynamics patterns and their average. Here we employ two estimation baselines. One utilizes gaze durations toward saliency areas as a feature of eye movements (see “Duration” in Table 4). We calculate temporal durations while subjects look at each saliency area, and get the average ratios between the durations and temporal intervals defined by saliency-dynamics patterns. From sequences of saliency-dynamics patterns, we obtain conditional probability distributions of the duration ratios for each level of mental focus, and utilize them for the estimation. The other utilizes PERCLOS [4] as a feature of eye movements (see “PERCLOS” in Table 4). In fact we simply calculate the average ratios between temporal intervals of eye-gaze data defects and those of saliency-dynamics patterns, instead of the duration ratios shown above.

We can confirm that estimations of all the patterns perform more accurately than the baselines. The proposed method switches different discriminative models for observed types of saliency-dynamics patterns. The results suggest that we can utilize better features in adapting to the changes of the saliency-dynamics patterns than uniform features employed in the baseline methods. Another difference of the proposed method from the baselines is that it focuses on the internal eye-movement dynamics in saliency areas. That is, the proposed method investigates not only “what subjects look at,” but also “how subjects look at,” and thus seems to enhance the accuracies.

Table 4 Estimation accuracies. SS, SD, MS, MSD and MD show the results for each saliency-dynamics pattern. Average shows the average accuracy of all the patterns. Duration and PERCLOS show results of the baseline methods.

Baselines		Proposed method					
Duration	PERCLOS	SS	SD	MS	MSD	MD	Average
53.8%	65.0%	66.9%	75.0%	78.8%	81.5%	76.3%	78.2%

From a viewpoint of improving the robustness of the method, one of the approaches is to vote estimation results within longer intervals. The estimation accuracy by integrating the results conducted in a single video interval (15 sec) is 93.3% on average. Besides, this paper conducts classification of two levels of mental focus as a relatively-simplified evaluation, but in principle, we can estimate multi-levels of the mental focus in the same fashion.

Comparing the two baselines, PERCLOS employs gazes of both saliency and non-saliency areas whereas Duration employs those of only saliency areas. The difference in their accuracy suggests that gazes of the non-saliency areas also contribute the discrimination of the levels of mental focus. For the discrimination of saliency and non-saliency areas, we simply set a fixed threshold for all the saliency maps. Therefore, the saliency extraction employed here has a potential risk of regarding actual gaze targets as non-saliency areas. To overcome the problem, one possible way is to improve the thresholding method such as to select thresholds adaptively for each saliency map with consideration of its saliency-value histogram.

Also, the videos we employed in the experiments contain some clear saliency areas, and therefore eye gaze seems to have a tendency to focus on those areas. However, to apply the proposed method to other kinds of videos, it is difficult to assume that clear saliency areas always exist. With regard to scenes with no saliencies such as plain natural sceneries or scenes filled with saliencies such as crowds, the proposed method with a simple saliency extraction originally has a problem with the specification of objects to be looked at. For improvement of the saliency extraction, we can introduce some heuristics in conjunction with the saliency map; for instance, employing the saliency map with object detection such as face detection [22], [23] can enhance the saliencies of actual objects to be looked at, and therefore it can be helpful to specify them.

6. Conclusions

We proposed to analyze the spatio-temporal correlation of dynamics between the visual saliency and eye movements for mental-focus estimation. Experimental results reveal that this correlation can be a meaningful clue for the mental focus, and the proposed method performs accurately to discriminate two levels of them. This study introduces an analysis of eye-movement dynamics by categorizing them from the aspect of visual-saliency dynamics in view, in order to estimate the mental focus. On the other hand, we can also utilize this correlation in the form of the detailed analysis of the visual-saliency dynamics from the aspect of the eye-movement dynamics. That is, extraction and modeling of the visual-saliency dynamics allow taking actual eye movements and mental focus into account, as well as introducing heuristics shown in Section 5.2. Besides, we can confirm the effectiveness and the necessity of the proposed analysis, but its sufficiency has yet to be revealed. Future work will seek to generalize the analysis of the spatio-temporal correlation between visual saliency and eye movements, and to apply the proposed method to more general video-viewing scenes.

Acknowledgments This work is in part supported by Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contract of 18049046.

Reference

[1] D'Orazio, T.T., Leo, M., Guaragnella, G. and Distante, A.: A Visual Approach for Driver Inattention Detection, *Pattern Recognition*, Vol.40, No.8, pp.2341–2355 (2007). Part Special Issue on Visual Information Processing.

[2] Doshi, A. and Trivedi, M.: Investigating the relationships between gaze patterns, dynamic vehicle surround analysis, and driver intentions, *IEEE Intelligent Vehicles Symposium*, Vol.10, No.3, pp.887–892 (2009).

[3] Fletcher, L. and Zelinsky, A.: Driver Inattention Detection based on Eye Gaze-Road Event Correlation, *International Journal on Robotics Research*, Vol.28, No.6, pp.774–801 (2009).

[4] Wierwille, W., Ellsworth, L., Wreggit, S., Fairbanks, R. and Kirn, C.: Research on vehicle-based driver status/performance monitoring: Development, validation, and refinement of algorithms for detection of driver drowsiness, *National Highway Traffic Safety Administration Final Report: DOT HS*, Vol.808, p.247 (1994).

[5] Ji, Q., Lan, P. and Looney, C.: A probabilistic framework for modeling and real-time monitoring human fatigue, *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans*, Vol.36, No.5, pp.862–875 (2006).

[6] Qvarfordt, P. and Zhai, S.: Conversing with the user based on eye-gaze patterns, *Proc. SIGCHI Conference on Human Factors in Computing Systems, CHI '05*, pp.221–230, ACM, New York, NY, USA (2005).

[7] Nakano, Y. and Ishii, R.: Estimating User's Engagement from Eye-gaze Behaviors in Human-Agent Conversations, *Proc. International Conference on Intelligent User Interfaces* (*IUI2010*) (2010).

[8] Hirayama, T., Dodane, J.-B., Kawashima, H. and Matsuyama, T.: Estimates of User Interest Using Timing Structures between Proactive Content-display Updates and Eye Movements, *IEICE Trans. on Information and Systems*, Vol.E-93D, No.6, pp.1470–1478 (2010).

[9] Aoki, H. and Ito, K.: Analysis of Cognitive Processes during Viewing of Television Commercials Based on Semantic Structure of Scenes and Eye Movement Data, *Journal of Japan Industrial Management Association*, Vol.52, No.2, pp.101–116 (2001).

[10] Itti, L., Koch, C. and Niebur, E.: A Model of Saliency-based Visual Attention for Rapid Scene Analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.20, No.11, pp.1254–1259 (1998).

[11] Foulsham, T. and Underwood, G.: What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition, *Journal of Vision*, Vol.8, No.2, pp.1–17 (2008).

[12] Böhme, M., Dorr, M., Krause, C., Martinetz, T. and Barth, E.: Eye movement predictions on natural videos, *Neurocomputing*, Vol.69, No.16–18, pp.1996–2004 (2006).

[13] Walther, D. and Koch, C.: Modeling attention to salient proto-objects, *Neural networks: The official journal of the International Neural Network Society*, Vol.19, No.9, pp.1395–407 (2006).

[14] Ma, Y.-F. and Zhang, H.-J.: A model of motion attention for video skimming, *Proc. 2002 International Conference on Image Processing*, Vol.1, pp.I–129–I–132 (2002).

[15] Kahneman, D.: *Attention and effort*, Prentice Hall (1973).

[16] Atkinson, R. and Shiffrin, R.: Human memory: A proposed system and its control processes, *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol.2, pp.89–195 (1968).

[17] Jones, G.M. and Milsum, J.H.: Spatial and Dynamic Aspects of Visual Fixation, *IEEE Trans. on Biomedical Engineering*, Vol.BME-12, No.2, pp.54–62 (1965).

[18] Saslow, M.G.: Latency for Saccadic Eye Movement, *Journal of the Optical Society of America*, Vol.57, No.8, pp.1030–1033 (1967).

[19] Henderson, J.M.: Human gaze control during real-world scene perception, *Trends in Cognitive Sciences*, Vol.7, No.11, pp.498–504 (2003).

[20] Becker, W. and Fuchs, A.F.: Prediction in the oculomotor system: Smooth pursuit during transient disappearance of a visual target, *Experimental Brain Research*, Vol.57, pp.562–575 (1985).

[21] Stern, J., Walrath, L. and Goldstein, R.: The Endogenous Eyeblink, *Psychophysiology*, Vol.21, No.1, pp.22–23 (1984).

[22] Sugano, Y., Matsushita, Y. and Sato, Y.: Calibration-free gaze sensing using saliency maps, *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR2010*), pp.2667–2674, IEEE (2010).

[23] Cerf, M., Harel, J., Einhauser, W. and Koch, C.: Predicting human gaze using low-level saliency combined with face detection, *Advances in Neural Information Processing Systems*, Vol.20, pp.241–248 (2008).

Ryo Yonetani received his B.E. degree in electrical and electronic engineering and M.S. degree in informatics from Kyoto University, Japan in 2009 and 2011, respectively. He is currently a Ph.D. candidate at the Graduate School of Informatics, Kyoto University. His research interests include human-computer interaction and human vision. He received the IBM Best Student Paper Award at the International Conference on Pattern Recognition '10. He is a student member of the Institute of Electronics, Information, and Communication Engineers Japan.

Hiroaki Kawashima received his M.S. and Ph.D. in informatics from Kyoto University, Japan in 2001 and 2007. He is currently a lecturer at the Graduate School of Informatics, Kyoto University, a JSPS Postdoctoral Fellow for Research Abroad, and a visiting researcher at the School of Electrical and Computer Engineering, Georgia Institute of Technology. His research interests include time-varying pattern recognition, human-computer interaction, hybrid systems, and networked control. He is a member of IEICE, the Human Interface Society of Japan, and the IEEE Computer Society.

Takatsugu Hirayama received his M.E. and Ph.D. from Osaka University, Japan, in 2002 and 2005, respectively. From 2005 to 2011, he was a research assistant professor in the Graduate School of Informatics, Kyoto University. He is currently a research assistant professor in the Graduate School of Information Science, Nagoya University. His research interests include computer vision, human vision, human communication, and human-computer interaction. He is a member of IEICE and the Human Interface Society of Japan.

Takashi Matsuyama received his B. Eng., M. Eng., and D. Eng. degrees in electrical engineering from Kyoto University, Japan, in 1974, 1976, and 1980, respectively. He is currently a professor in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. His research interests include computer vision, 3D video, human-computer interaction, and smart energy management. He wrote about 100 papers and books including two research monographs, A Structural Analysis of Complex Aerial Photographs, PLENUM, 1980 and SIGMA: A Knowledge-Based Aerial Image Understanding System, PLENUM, 1990. He won eleven best paper awards from Japanese and international academic societies including the Marr Prize at ICCV '95. He is on the editorial board of the Pattern Recognition Journal. He was awarded Fellowships from the International Association for Pattern Recognition, IPSJ, and IEICE.