Vol.2014-CVIM-192 No.32
2014/5/16

Modeling Spatiotemporal Correlations between Video Saliency and Gaze Dynamics

Ryo Yonetani[†1,1,a)] Hiroaki Kawashima[1,b)] Takashi Matsuyama[1,c)]

Abstract: In this study, we propose a framework to describe the relationship named *spatiotemporal correlation* between video contents and human gaze dynamics. The spatiotemporal correlation consists of (1) the event-level spatiotemporal gaps between visual events in videos and gaze reactions and (2) the scene-level correlations between video scene structures and corresponding gaze dynamics. Our framework can describe this twofold relationship simply and efficiently by discovering and combining primitive spatiotemporal patterns of visually salient regions in videos and those of gaze. The effectiveness of this framework is confirmed via several practical tasks of gaze behavior analyses in real environments, attentional target identification, attentive state estimation and gaze point prediction.

1. Introduction

We humans are surrounded by a vast amount of display systems in our daily life. These systems provide visual contents involving a variety of visual events such as scene changes in movies, human actions in surveillance videos and camera motions in egocentric videos. Facing such contents, we direct our eyes to them and try to get information by design. Alternatively, eyes are sometimes directed to the contents unconsciously when eye-catching events happen such as sudden pop-ups of logos in commercial films.

Researchers have long studied visual contents and human behavior mainly in the fields of computer vision, human computer interaction (HCI), multimedia, visual psychology and neuroscience. Their interests loosely fall into two issues: analyzing visual contents themselves (*visual content analyses*) and analyzing how humans act toward the contents (*human behavior analyses*). Above all, an eye movement is one of the important modalities that strongly reflect both mental states of humans and visual events in contents. Gaze behavior analyses and their applications in real environments are indeed one of the recent trends: for example, measuring gaze-based social interactions [1], [2], estimating mental states from gaze [3], [4], [5], detecting developmental disorders [6] and gaze-based content designs [7].

Based on the aforementioned two research tides, we aim to develop a framework to describe the relationships formed by visual contents and gaze data. Within the framework, we try to describe the effects of visual events upon observers' gaze via visual content analyses. The effectiveness of the framework is then assessed via practical gaze behavior analyses in real environments.

[1] Kyoto University
[†1] Presently with The University of Tokyo
[a)] yonetani@iis.u-tokyo.ac.jp
[b)] kawashima@i.kyoto-u.ac.jp
[c)] tm@i.kyoto-u.ac.jp

1.1 Issues and our contributions

We particularly cover the situations where a single human observer is watching various videos taken in real environments, such as TV commercial films, surveillance videos and dynamic interfaces. Under such situations, let us assume that gaze data (sequences of 2-d gaze points on a screen) of the observer are obtained via gaze tracking. In addition, the videos are assumed to contain various kinds of visual events such as object translations and deformations, texture variations and scene changes, which all have the potential of affecting observed gaze dynamics. The aim of this study is to describe these effects in our framework as the relationships between video and gaze dynamics. To this end, we address the following two issues.

Issue 1: Handling diverse visual events in videos

Videos taken in real environments can display a variety of visual events in the form of spatiotemporal patterns, and at the same time, those events are given a variety of category labels, which results in diverse physics and semantics of the videos. Moreover, it is generally uncontrollable and unknown that when, where and what kinds of visual events take place in the videos. These natures of events bring difficulties when modeling them as a systematic input to eyes and analyzing their effects upon gaze dynamics.

Issue 2: Considering time-varying scene structures

Due to the diverse visual events posed above, scene structures, overall properties of video scenes consisting of various visual events, can vary over time. Thus we need to consider the following twofold relationships: (1) visual events in a scene structure affect gaze reactions to the events, and (2) scene structures affect overall gaze dynamics being observed. For example, (1) objects in motion can cause a reaction delay in pursuit gaze reactions, but (2) it depends on the types of scene structures (e.g., if they contain moving objects) that if gaze dynamics originally contains the pursuits. It is a different situation from traditional visual psychology and HCI studies that aim to clarify gaze behavior under controlled

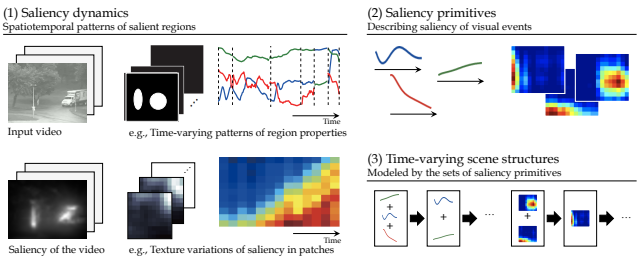

Fig. 1 Overview of saliency dynamics models. Parts of the images in this figure are contained in the dataset provided by [8].

situations, assuming a constant or limited type of scene structures and visual events to observe specific gaze dynamics. In conclusion, a novel framework is required to describe the relationships when dealing with the time-varying scene structures.

Contribution 1: saliency dynamics models

Our first contribution is to propose a model named *saliency dynamics models* that describes the effects of visual events upon gaze dynamics for Issue 1. The basic idea is to leverage the dynamic changes of visual saliency in videos for event characterizations. This idea is aimed at avoiding semantic diversity of visual events while preserving the essence when describing the relationships between video and gaze data. Specifically, we extract spatiotemporal patterns of salient regions from videos, which we refer to as saliency dynamics (Figure 1 (1)).

To describe the saliency dynamics, our models introduce a primitive spatiotemporal pattern of salient regions referred to as *saliency primitives* (Figure 1 (2)). The saliency primitives serve as a unit to describe the saliency of various events such as object translations, deformations and texture variations. Namely, they indicate how much visual events attract our attention while sacrificing why they attract the attention explained by semantic aspects. In addition, a set of the primitives can characterize overall scene structures and thus they can contribute to the description of time-varying scene structures posed in Issue 2 (Figure 1 (3)). By achieving saliency primitives from videos in a data-driven manner, we can describe scene structures efficiently for given videos.

Contribution 2: framework for spatiotemporal correlations

The second contribution is development of a novel framework to describe the relationships between video and gaze data. While scene structures of videos can be described by saliency dynamics models, we need models of gaze dynamics and the relationships as well, where the relationships involve the twofold characteristics presented in Issue 2.

To this end, we first regard gaze dynamics as sequences of primitive patterns, *gaze primitives*. Thanks to the primitive-based descriptions of scene structures and gaze dynamics, we can model the relationships between video and gaze data simply as those among primitives. Specifically, we now introduce the special term *spatiotemporal correlations* for the primitive-based descriptions of the twofold relationships posed in Issue 2. The spatiotemporal correlations consist of *event-level spatiotemporal gaps* and *scene-level correlations* of the following characteristics:

Event-level spatiotemporal gaps are temporal or spatiotemporal distances defined in a pair of saliency and gaze primitives, which aim to explain the effects from a single visual event to the corresponding gaze reaction. The spatiotemporal gaps are brought by various dynamic factors in gaze behavior, such as reaction delays and anticipation when reacting to a certain visual event.

Scene-level correlations are the combinations of modeled scene structures, i.e., sets of saliency primitives, and (possibly sequences of) gaze primitives in a certain temporal interval. Dynamic changes in the types of these correlations over time can explain the effects from time-varying scene structures to the gaze dynamics posed in Issue 2.

Consequently, the proposed framework comprises the models of scene structures, gaze dynamics and spatiotemporal correlations as summarized in Figure 2. The framework first receives video and gaze data to extract primitives (Arrows 1 in Figure 2) and exploit them for describing their event-level spatiotemporal gaps and scene-level correlations (Arrows 2 in the figure).

1.2 Effectiveness of our framework

The effectiveness of our framework, in other words, how the framework can describe actual situations and contribute to practical applications, are assessed by describing the relationships and evaluating them via practical gaze behavior analyses in real environments. More specifically, we address the following three practical tasks while gradually upgrading a variety of video contents being worked with.

Attentional target identification (Section 3) is a task of judging which objects in contents are looked at by observers. Since traditional methods mainly rely on absolute positions of gaze points, they often suffer from gaze tracking errors. We solve this by introducing the event-level spatiotemporal gaps between visual events and the gaze reactions. This section particularly adopts manually-designed contents with simple and constant scene structures as a first step.

Attentive state estimation (Section 4) is a variant of mental state estimation tasks, which aims to judge if observers concentrate on videos or not. We now broaden the variety of videos to designed but unconstrained ones (like TV commercials) and extend traditional approaches that can only work when scene structures are constant. Specifically, this section introduces saliency dynamics models to adaptively use features for estimation based on time-varying scene structures.

Gaze point prediction (Section 5) is a task of predicting where observers tend to look at from video data. Existing methods cannot deal with a particular situation for unedited and natural videos such as gaze dynamics containing delays because of fast motions of attentional targets. We finally utilize overall spatiotemporal correlations for this situation by learning the degree of event-level spatiotemporal gaps conditioned by scene-level correlations for prediction.

In the following section, we first present fundamental saliency dynamics models which will be mainly used in Sections 4 and 5.

2. Modeling of Saliency Dynamics

This section presents how various visual events and scene structures in videos can be modeled in our framework. Visual events involve semantic and physical aspects that have a different

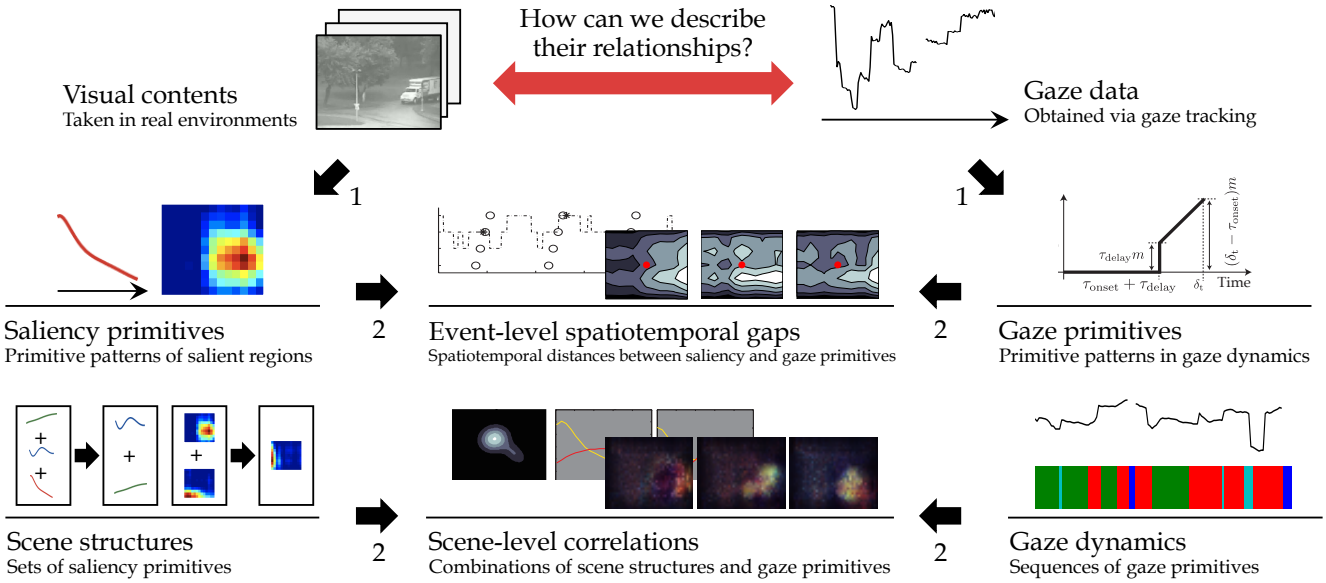

Fig. 2 Framework describing the spatiotemporal correlations between video and gaze data. Parts of the images in this figure are contained in the dataset provided by [8].

effect upon gaze dynamics. Specifically, the semantics provide "why observers watched visual events" since the observers often direct their eyes to regions with specific semantic categories. For example, human faces are known to attract our attention well, and several gaze behavior analyses take particular note of the faces in video scenes [9], [10]. On the other hand, the physical aspects are capable of explaining "how much observers are attracted to visual events". To cope with the diversity of those two aspects posed in the previous section, we particularly focus on the saliency of visual events as their physical aspects while sacrificing their semantics. Furthermore, we model them simply and efficiently by discovering their primitive spatiotemporal patterns.

2.1 Modeling based on saliency primitives

Saliency is one of the important properties of visual stimuli that attract human visual attention in a bottom-up manner. The degree of saliency is originally given by the contrast of stimuli between a certain point and its surround [11]. Our model begins with calculation of pixel-wise saliency for each frame of input videos, which is often referred to as *saliency maps* [12]. In the saliency maps, we use local regions of highly salient points as salient regions that attract observers' gaze.

Salient regions contain dynamic changes over time as various visual events occur such as object translations, deformations and texture variations. In addition to the individual variations of salient regions, the dynamics that the variations follow as well as the number of the regions also dynamically change as scenes do over time. In this study, we refer to such dynamic changes provided by salient regions as *saliency dynamics*.

The basic concept of our models of saliency dynamics is to introduce primitive spatiotemporal patterns of salient regions as a unit. The primitive patterns, which we refer to as *saliency primitives*, describe the variations of salient regions caused by visual events. Furthermore, a set of primitives can characterize scene structures consisting of multiple events simultaneously occurring in a certain temporal interval. By modeling primitives appropriately and learning them from a set of videos, we can describe how events and scene structures affect gaze dynamics based on the models efficiently configured for the given videos.

Options for the modeling of saliency primitives

There are several options for the modeling of saliency primitives to consider what types of and how events are described.

The first option is about how to define a temporal interval to extract spatiotemporal patterns for saliency primitives: *segmentation* and *sliding windows*. The segmentation approach looks for a set of points (segmentation points) where the temporal intervals split. This approach can explicitly deal with scene change events while it has difficulty in detecting segmentation points so as not to split spatiotemporal patterns incorrectly. On the other hand, the sliding-window approach slides a fixed-length window from the beginning to the end of sequences with an overlap and conducts a certain procedure in the temporal intervals defined by each window. This approach can avoid splitting spatiotemporal patterns incorrectly thanks to the redundant representation by the overlap although we cannot deal with scene change events explicitly.

Given a temporal interval, the second option is about what kinds of features should be extracted as saliency dynamics patterns in the interval. If we want to take particular note of variations in a sole salient region, for example when we deal with events caused by distinct objects, the properties of regions such as positions, shapes and the degree of saliency can be explicitly utilized. On the other hand, when we deal with a more general variation including texture variations, the changes of a sole salient region cannot always describe the whole variations. For this case, it is effective to describe the patterns of one or more salient regions jointly and implicitly as parts of the texture variations of saliency in a certain spatiotemporal patch. We refer to these two approaches as *object-based* and *patch-based* approaches.

The third option is how to represent the extracted patterns as saliency primitives. While model-based representations like

Table 1 Differences between OSDMs and PSDMs.

	OSDM (Section 2.2)	PSDM (Section 2.3)
Applicable video types	Intentionally-designed videos	Unedited natural videos
Definition of temporal intervals	Segmentation	Sliding-windows
Features to be extracted	Properties of regions (object-based)	Textures (patch-based)
Representations of primitives	Model-based	Direct

switching linear dynamical systems [13], [14] and dynamic textures [15], [16] are aimed at representing patterns efficiently with a small number of parameters, we need to define a suitable model for the given patterns. On the other hand, direct representations by the form of vector sequences are model-free and they can deal with any kinds of patterns although they take an ingenuity to avoid diversity and noise in the patterns.

Video categories and proposed saliency dynamics models

Finally, we introduce specific saliency dynamics models based on the options presented above. For guidance to choose the options, we here introduce two categories of videos that individually tend to contain specific types of visual events.

Intentionally-desgined videos. The videos taken with a certain objective, e.g., TV commercial films and movies, are designed to attract observers' attention on intended objects (logos, products and so on), and thus the limited number of objects can be shown simultaneously in a certain temporal interval. These objects are mostly highly salient since they are designed to make their appearance distinct relative to their surrounds. In addition, they often involve frequent scene changes to give much information to observers.

Unedited natural videos. Videos recorded under uncontrolled situations without intentions do not always contain the limited number of objects with high saliency. For example, plain natural sceneries sometimes contain less objects. On contrary, surveillance videos with human crowds contain massive objects. Note that visual events are often regarded as texture variations when analyzing natural videos [16], [17]. Moreover, unedited videos have less scene changes.

These two categories of videos require different options when modeling the saliency primitives. We thus propose two models of saliency dynamics which are individually suitable for those categories (see Table 1). Specifically, we refer to the model for intentionally-designed videos as *object-based saliency dynamics models (OSDM)* and for unedited natural videos as *patch-based saliency dynamics models (PSDM)*. The OSDM is aimed at describing visual events caused by distinct objects as well as scene changes. On the other hand, the PSDM introduces the modeling of saliency primitives suitable for a greater variety of local events including texture variations. Note that they are not the unique models against the two video categories. For example, we can introduce model-based representation of primitives in the PSDM, like a family of dynamic textures [15], [16].

Notations

Let $\boldsymbol{p} = (x, y) \in \Omega$ be a 2-d point in a frame of videos, where $\Omega \subset \mathbb{R}_+^2$ is a spatial domain corresponding to the frame. We particularly use $\boldsymbol{p}_t = (x_t, y_t)$ if we specify a certain point at frame $t \in \mathbb{N}$. The saliency maps are denoted as $S : \Omega \to \mathbb{R}_+$, where the degree of saliency at point $\boldsymbol{p}$ is $S(\boldsymbol{p})$. Above all, we specify the saliency map at frame t as S_t and the local regions $\Omega' \subseteq \Omega$ of S_t as $S_{(\Omega',t)}$ (i.e., $S_t = S_{(\Omega,t)}$). Then, a sequence of saliency maps obtained from a video can be denoted as an ordered set, $\mathcal{S} = (S_1, \ldots, S_T)$, where T is the number of frames. If we introduce a local spatiotemporal patch defined as $\Omega' \times \mathcal{T}$ where $\mathcal{T} \subseteq [1, T]$, the local spatiotemporal volume in the patch is denoted as $\mathcal{S}_{\Omega' \times \mathcal{T}} = (S_{(\Omega', \min(\mathcal{T}))}, \ldots, S_{(\Omega', \max(\mathcal{T}))})$ where $\min(\mathcal{T})$ and $\max(\mathcal{T})$ are lower and upper bounds of $\mathcal{T}$, respectively.

2.2 Object-based saliency dynamics model

The OSDM is aimed at modeling saliency dynamics provided by intentionally-designed videos containing visual events from distinct objects and scene changes. A key problem to use this model is how to detect segmentation points that give reasonable intervals to model the patterns in each interval by saliency primitives accurately. In this section, we briefly introduce a formulation and model estimation of the OSDM (see [18] for detail).

2.2.1 Formulation

We first assume that videos are segmented into a sequence of K temporal intervals, $\mathcal{I} = (I_1, \ldots, I_K)$. Saliency maps in interval $I_k = [i_{k1}, i_{k2}]$, $\{S_t \mid t \in I_k\}$ individually contain C_k salient regions, where the spatiotemporal pattern of the c-th region is described by a sequence of multidimensional vectors, $\Theta_k^{(c)} = \left(\boldsymbol{\theta}_{i_{k1}}^{(c)}, \ldots, \boldsymbol{\theta}_{i_{k2}}^{(c)}\right)$. Then, the saliency dynamics in interval I_k can be represented by a set of patterns, $\boldsymbol{\Theta}_k = \{\Theta_k^{(1)} \ldots, \Theta_k^{(C_k)}\}$.

We describe each pattern $\Theta_k^{(c)}$ by a single saliency primitive modeled in a parametric manner. Since distinct objects in our videos of interests mostly behave naturally to attract our attention, the corresponding patterns of salient regions seem to follow some dynamical systems. We thus define saliency primitive $D_k^{(c)}$ identified to $\Theta_k^{(c)}$ by a first-order multivariate autoregressive model (AR model) as a family of LDSs:

$$\boldsymbol{\theta}_t^{(c)} = M_k^{(c)} \boldsymbol{\theta}_{t-1}^{(c)} + \boldsymbol{b}_k^{(c)} + \boldsymbol{v}_t, \tag{1}$$

where $M_k^{(c)}$ is a $J \times J$ transition matrix, $\boldsymbol{b}_k^{(c)}$ is a J-dimensional bias vector, $\boldsymbol{v}_t$ is a J-dimensional noise vector modeled by a Gaussian distribution $\mathcal{N}(0, Q_k^{(c)})$. Namely, saliency primitive $D_k^{(c)}$ has $M_k^{(c)}, \boldsymbol{b}_k^{(c)}, Q_k^{(c)}$ as parameters.

In terms of describing complex patterns by the switches of simple models, the OSDM is similar to the switching linear dynamical systems (SLDS) [13], [14]. Comparing to the SLDS, the OSDM introduces a set of AR models to describe dynamics in a certain interval, and thus it has an advantage in describing the situations where the number of elements (objects) providing the dynamics change over time.

2.2.2 Extraction and modeling of salient regions

When introducing the OSDM, we first need to model $\boldsymbol{\theta}_t^{(c)}$ so as to describe properties of salient regions. In order to handle their positions, shapes and the degree of saliency, we model the regions in a frame by the Gaussian mixture model (GMM). That is, each salient region is modeled by a single Gaussian component where the parameters of the Gaussian describe the properties introduced above. Let us denote a mean vector (= positions), covariance matrix (= shapes) and weight (= saliency) of the c-th Gaussian as $\boldsymbol{\mu}_t^{(c)}, \Sigma_t^{(c)}, \phi_t^{(c)}$. We estimate these param-

eters from massive samples approximating input saliency maps via an EM algorithm. Practically, we give estimated parameters at a certain frame as initial inputs in the next frame to obtain a continuous change of the estimated parameters over time. The properties of the c-th region at frame t are finally described by $\boldsymbol{\theta}_t^{(c)} = ((\boldsymbol{\mu}_t^{(c)})^{\mathrm{T}}, (\boldsymbol{\sigma}_t^{(c)})^{\mathrm{T}}, \phi_t^{(c)})^{\mathrm{T}} \in \mathbb{R}^6$ where $\boldsymbol{\sigma}_t^{(c)} \in \mathbb{R}^3$ consists of two variances and a covariance of $\Sigma_t^{(c)}$.

2.2.3 Identification of saliency primitives and segmentation

Problem settings

The OSDM introduces temporal interval sequence $\mathcal{I} = (\boldsymbol{I}_1, \ldots, \boldsymbol{I}_K)$ to deal with time-varying scene structures that characterize scene change events. Each interval contains a set of spatiotemporal patterns of salient regions, where they are supposed to be identified by saliency primitives defined in Equation (1).

Model estimation for the OSDM consists of segmentation of input saliency maps to derive appropriate interval sequence $\mathcal{I}$ and description of spatiotemporal pattern $\Theta_k^{(c)}$ by saliency primitive $D_k^{(c)}$. Segmentation $\mathcal{I}$ should be given so as to identify $D_k^{(c)}$ with small identification costs to $\Theta_k^{(c)}$ in interval I_k. On the other hand, $\mathcal{I}$ should be given preliminarily when identifying primitives to spatiotemporal patterns and evaluate costs.

To address this problem, we first generate many temporal interval candidates and select an appropriate segmentation based on identification costs of saliency primitives in each candidate. Specifically, we first generate hierarchical structures of interval candidates based on a scale-space representation of inter-frame differences of saliency maps (Figure 3 (2-1)), fit GMMs to extract salient regions and identify saliency primitives to their patterns in each interval candidate (Figure 3 (1-1) and (1-2)). Then, we evaluate segmentation points defined by two successive interval candidates based on identification costs of the primitives and derive a whole segmentation (Figure 3 (2-2)). As a consequence, we can conduct the segmentation based on the identification costs.

Identification of saliency primitives

Given a certain interval, $I_k = [i_{k1}, i_{k2}]$, our identification procedure consists of estimating the number of components for GMM and at the same time identifying saliency primitives to each spatiotemporal pattern in the sequence of saliency maps $(S_{i_{k1}}, \ldots, S_{i_{k2}})$ (Figure 3 (1)). We first set a range to the number of components, $\{\Gamma_{\min}, \Gamma_{\max}\}$ and fit $\Gamma_{\min}, \ldots, \Gamma_{\max}$-component GMMs individually to S_t via the procedure in Section 2.2.2. We describe the spatiotemporal pattern of the c-th of $\Gamma \in \{\Gamma_{\min}, \ldots, \Gamma_{\max}\}$ regions in the k-th interval as $\Theta_k^{(c,\Gamma)}$.

Let us denote the saliency primitive identified to $\Theta_k^{(c,\Gamma)}$ as $D_k^{(c,\Gamma)}$. As defined in Equation (1), $D_k^{(c,\Gamma)}$ has a set of parameters consisting of transition matrix $M_k^{(c)}$, bias vector $\boldsymbol{b}_k^{(c)}$ and error covariance matrix $Q_k^{(c)}$ (in what follows, we omit subscript Γ without loss of generality). $M_k^{(c)}$ and $\boldsymbol{b}_k^{(c)}$ can be basically estimated by minimizing a prediction error from $\boldsymbol{\theta}_{t-1}^{(c)}$ to $\boldsymbol{\theta}_t^{(c)}$. However, positions, shapes and the degree of saliency of salient regions, which are described by elements of $\Theta_k^{(c)}$, sometimes perform high correlation to each other. Thus we estimate parameters by the Ridge regression so that we can avoid a multicollinearity problem.

Once we estimate the parameters, we can generate spatiotemporal pattern $\hat{\Theta}_k^{(c)} = (\hat{\theta}_{i_{k1}}^{(c)}, \ldots, \hat{\theta}_{i_{k2}}^{(c)})$ from given initial value $\boldsymbol{\theta}_{i_{k1}}^{(c)}$ of original pattern $\Theta_k^{(c)} = (\theta_{i_{k1}}^{(c)}, \ldots, \theta_{i_{k2}}^{(c)})$. We then calculate error covariance matrix $Q_k^{(c)}$ by modeling the distribution of errors between original and generated patterns, $\theta_t^{(c)} - \hat{\theta}_t^{(c)}$ by a normal distribution: $\theta_t^{(c)} - \hat{\theta}_t^{(c)} \sim \mathcal{N}(0, Q_k^{(c)})$. In addition, we can calculate a negative log likelihood (NLL) score $h_k^{(c)}$ by evaluating the errors, $h_k^{(c)} = -\sum_{t=i_{k1}}^{i_{k2}} \log P(\theta_t^{(c)} - \hat{\theta}_t^{(c)}; 0, Q_k^{(c)})$.

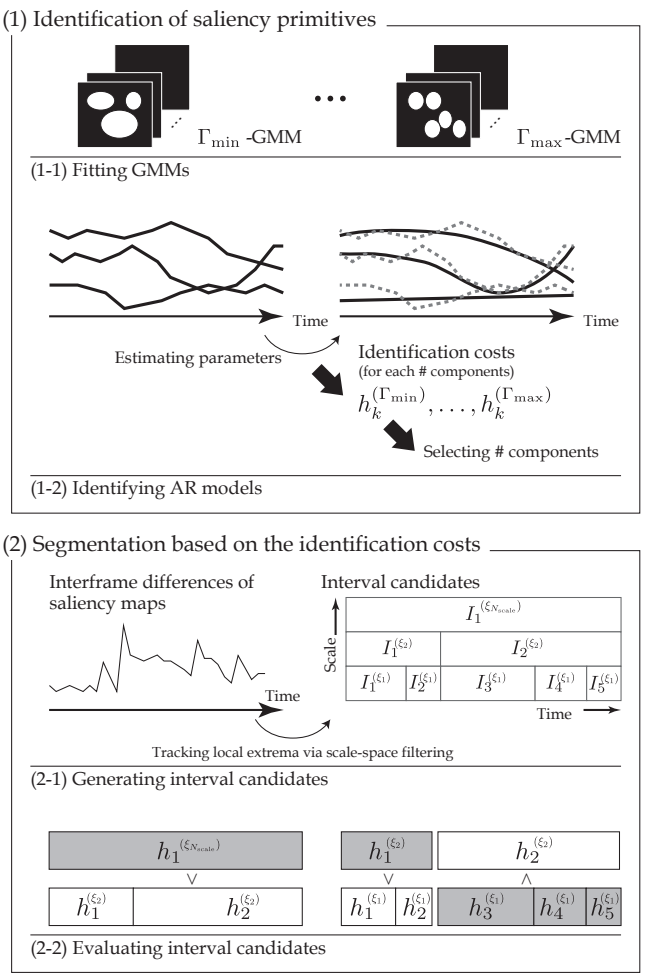

Fig. 3 Estimation algorithm for the object-based saliency dynamics models.

As a result of the procedure above, we have a set of saliency primitives $\{D_k^{(c,\Gamma)} \mid c = 1 \ldots, \Gamma\}$ and corresponding NLL scores $\{h_k^{(c,\Gamma)} \mid c = 1 \ldots, \Gamma\}$ for each $\Gamma \in \{\Gamma_{\min}, \ldots, \Gamma_{\max}\}$. To determine the number of components that is the most suitable for introducing saliency primitives at the k-th interval, we first evaluate the worst fit of primitives for each Γ, $h_k^{(\Gamma)} = \max\{h_k^{(c,\Gamma)} \mid c = 1 \ldots, \Gamma\}$. As Γ increases from $\Gamma_{\min}$ to $\Gamma_{\max}$, NLL score $h_k^{(\Gamma)}$ decreases until the fitness of saliency primitives becomes sufficiently good. We thus define $\hat{\Gamma}_k$ as the point where the NLL scores stop decreasing. Finally, we obtain primitive set $\{D_k^{(1)} \ldots D_k^{(\hat{\Gamma}_k)}\}$ from spatiotemporal patterns in temporal interval I_k, $\boldsymbol{\Theta}_k = \{\Theta_k^{(1)}, \ldots, \Theta_k^{(\hat{\Gamma}_k)}\}$, where the identification cost of primitives is given as $h_k = h_k^{(\hat{\Gamma}_k)}$.

Segmentation based on the scale-space analysis

Video segmentation is a well-known problem to detect scene change events in visual content analyses as reviewed in [19]. Our segmentation technique presented below is aimed at detecting the scene changes with the object to describe saliency dynamics patterns in each interval accurately by a set of saliency primitives.

A key idea is to generate multiple interval candidates base on the scale-space analysis [20] and evaluate the segmentation points

between successive interval candidates based on the identification costs of primitives (Figure 3 (2-1)). Specifically, we first calculate difference $f_t \in \mathbb{R}$ between successive saliency maps S_{t-1}, S_t to obtain sequence $\boldsymbol{f} = (f_1, \ldots, f_T)$ as an input. We then convolve a series of Gaussian functions with smoothing scales $\{\xi_1, \ldots, \xi_{N_{\text{scale}}}\}$ ($\xi_{n-1} < \xi_n$), let's say $Gauss^{(\xi_n)}$, to sequence $\boldsymbol{f}$ and obtain a scale-space representation $\boldsymbol{f}^{(\xi_n)} = \boldsymbol{f} * Gauss^{(\xi_n)}$, where $*$ denotes a convolution operation. By tracking local extreme points in a set of outputs $\{\boldsymbol{f}^{(\xi_1)}, \ldots, \boldsymbol{f}^{(\xi_{N_{\text{scale}}})}\}$ with changing the smoothing scales from $\xi_{N_{\text{scale}}}$ to ξ_1, a hierarchical structure of the points is obtained. For simplicity of discussions, we set $\{\xi_1, \ldots, \xi_{N_{\text{scale}}}\}$ so as to obtain new local extreme points for every scale variation $\xi_n \to \xi_{n-1}$. In addition, we set the maximum scale $\xi_{N_{\text{scale}}}$ so as not to contain any local extreme point.

Given a local extreme point at certain scale ξ_n, we look for the corresponding point at scale ξ_1 by tracking the point from ξ_n to ξ_1 and use the point as one of the segmentation points at ξ_n. Then, we deal with segments defined by successive segmentation points as interval candidates. We denote the interval candidates generated at scale ξ_n as $\hat{\mathcal{I}}^{(\xi_n)} = (\hat{I}_1^{(\xi_n)}, \ldots, \hat{I}_{K_{\xi_n}}^{(\xi_n)})$. For each interval, a set of saliency primitives are identified with spatiotemporal patterns in $\hat{I}_k^{(\xi_n)}$ and identification cost $h_k^{(\xi_n)}$ is given to the interval based on the procedure presented so far.

After obtaining identification costs for all the interval candidates, we can evaluate segmentation points (Figure 3 (2-2)). Let us introduce a subsequence of $\hat{\mathcal{I}}^{(\xi_{n-1})}$ at scale ξ_{n-1}, $\hat{\mathcal{I}}^{(\xi_{n-1})}|_{(j,j+l)} = (I_j^{(\xi_{n-1})}, \ldots, I_{j+l}^{(\xi_{n-1})})$, which defined in the same interval as candidate interval $\hat{I}_k^{(\xi_n)}$ at scale ξ_n. In the segmentation, we choose one of $\hat{I}_k^{(\xi_n)}$ and $\hat{\mathcal{I}}^{(\xi_{n-1})}|_{(j,j+l)}$ based on the identification costs (Figure 3 (2-2)). Specifically, we split the interval if $h_k^{(\xi_n)} \geq \sum_{j'=j}^{j+l} h_{j'}^{(\xi_{n-1})}$. By recursively conducting the judgements from $\xi_{N_{\text{scale}}}$ to ξ_1, we can obtain an appropriate segmentation to describe spatiotemporal patterns with saliency primitives.

2.2.4 Example

As an example, we analyzed 12 TV commercial films of 15 sec length stored at 30 fps. These videos are designed to contain several distinct objects generating various visual events and scene changes over time. As for an input saliency map, we adopted the graph-based visual saliency [21], where the features include luminance, color, edge orientations and motions. In addition, we assumed there were only several objects in each frame of the videos and empirically set $\Gamma_{\min} = 1, \Gamma_{\max} = 8$. Under these settings, the number of intervals, K, was estimated at $11 \leq K \leq 19$ for any video (mean: 15.7, SD: 2.2). The number of primitives (i.e., salient regions) in each interval, C_k, was estimated at $2 \leq C_k \leq 5$ for any scene (mean: 2.8, SD: 0.7).

An example of segmentation results is depicted in Figure 4. Although many peaks were found in the 1st row, the final segmentation in the 3rd row contained several scene change events such as the 4th to 5th, 8th to 9th, 11th to 12th, 14th to 15th and 15th to 16th intervals. In addition, a new appearance of objects also contribute to the switches of scene structures such as the 1st to 2nd and the 12th to 13th intervals.

Extracted spatiotemporal patterns of salient regions corresponding to Figure 4 are shown in the left of Figure 5. Obviously,

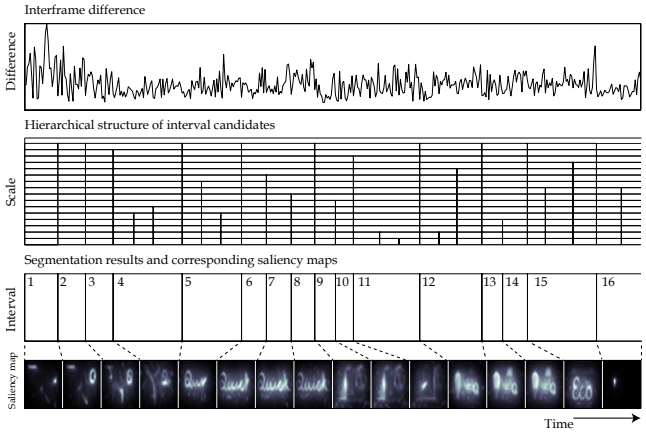

Fig. 4 Example of segmentation results. 1st row: a sequence of interframe differences. 2nd row: hierarchical structures of interval candidates generated by the scale-space analysis of the interframe differences in the 1st row. 3rd row: segmentation results consisting of the selected intervals from the candidates in the 2nd row. the images below depict saliency maps at the beginning frame of each interval. The images used in this figure was provided by courtesy of Panasonic Corporation.

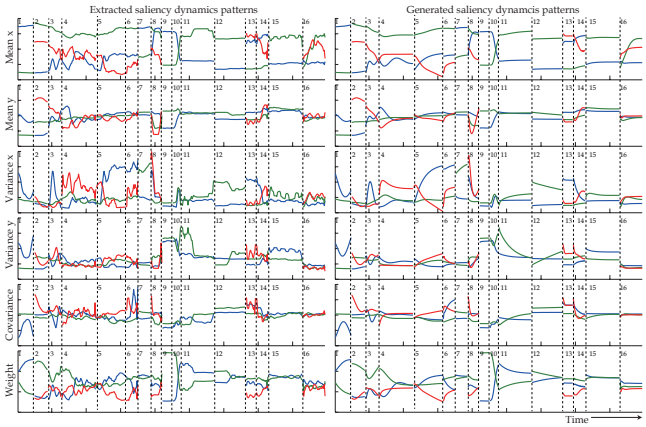

Fig. 5 Extracted spatiotemporal patterns of salient regions (left of the figure) and generated patterns from the identified saliency primitives (right). Lines of the same colors in each interval between left and right of the figure indicate the same saliency primitive. Each row shows different properties of regions.

the extracted patterns contain large noises. One of the reasons is the definition of saliency; saliency maps are generally obtained frame by frame and represent the degree of saliency at each point in a frame, and thus the point in the same object can obtain different saliency if the surrounding objects generate visual events. Another reason is the instability in the fitting of GMMs. When salient regions are too large to model by a single Gaussian component, the proposed model introduce several components to represent the regions such as the 5th, 6th and 7th intervals, which sometimes makes the fitting unstable.

The right of Figure 5 depicts the generated patterns from identified saliency primitives. Note that this result finally describes the time-varying scene structures modeled by a set of saliency primitives. Regardless of the noisy inputs explained above, saliency primitives allow us to deal with underlying primitive patterns in the extracted spatiotemporal patterns since the identification includes the estimation of noise variance. Since the primitives contain translations, deformations (resizes) and saliency variations

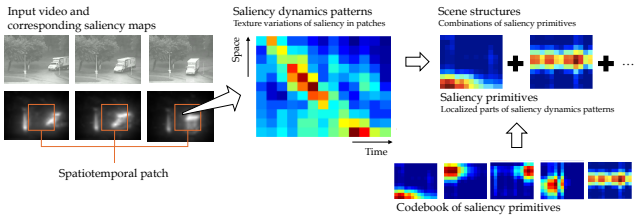

Fig. 6 Overview of the patch-based saliency dynamics model. Parts of the images in this figure are contained in the dataset provided by [8].

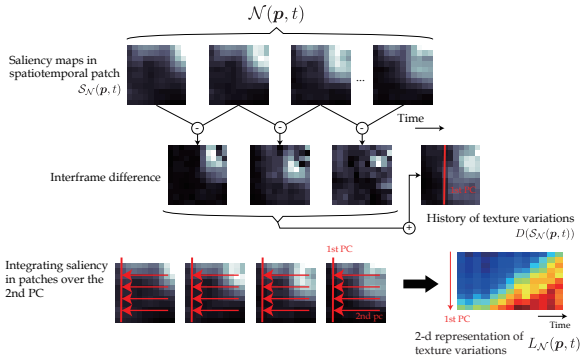

Fig. 7 Extracting texture variations of saliency maps.

of salient regions, they are capable of describing visual events caused by distinct objects.

2.3 Patch-based saliency dynamics model

Next, we present the patch-based saliency dynamics model (PSDM), which takes a great advantage when dealing with unedited natural videos such as surveillance videos. Among the options adopted in this model, the direct representation of saliency primitives allows us to deal with complex variations caused by a variety of visual events. However, we need an efficient and robust modeling to cope with the diversity and noise in the saliency dynamics. To this end, we introduce a codebook of saliency primitives, where the primitives describe localized parts of saliency dynamics in a direct manner, like in the right of Figure 6. By statistically learning the codebook from videos so that each primitive describes the parts frequently appearing the videos, we can achieve the efficiency as well as the robustness when describing saliency dynamics.

In the following subsections, We briefly present a method to extract texture variations of saliency maps in a spatiotemporal patch as saliency dynamics and learning method of the codebook (see [22] for detail).

2.3.1 Extracting texture variations of saliency maps

Let us denote a spatiotemporal patch around $(\boldsymbol{p},t)=(x,y,t)$ as

$$\mathcal{N}(\boldsymbol{p},t) := \Omega_{(\delta_\mathrm{x},\delta_\mathrm{y})}\times\mathcal{T}_{\delta_\mathrm{t}},$$
$$\Omega_{(\delta_\mathrm{x},\delta_\mathrm{y})} \subseteq [x-\delta_\mathrm{x},x+\delta_\mathrm{x}]\times[y-\delta_\mathrm{y},y+\delta_\mathrm{y}],\ \mathcal{T}_{\delta_\mathrm{t}}\subseteq[t-\delta_\mathrm{t},t+\delta_\mathrm{t}],$$

where $\delta_\mathrm{x},\delta_\mathrm{y},\delta_\mathrm{t}$ define the size of patch. Although we can essentially define δ_x and δ_y independently, in what follows we use the same size $\delta_\mathrm{x}=\delta_\mathrm{y}=\delta_\mathrm{s}$ and denote the spatial patch as $\Omega_{\delta_\mathrm{s}}$ for simplicity. Then, a spatiotemporal volume of saliency maps cropped by the patch is denoted as follows:

$$S_{\mathcal{N}(\boldsymbol{p},t)}=\left(S_{(\Omega_{\delta_\mathrm{s}},\min(\mathcal{T}_{\delta_\mathrm{t}}))},\ldots,S_{(\Omega_{\delta_\mathrm{s}},\max(\mathcal{T}_{\delta_\mathrm{t}}))}\right).$$

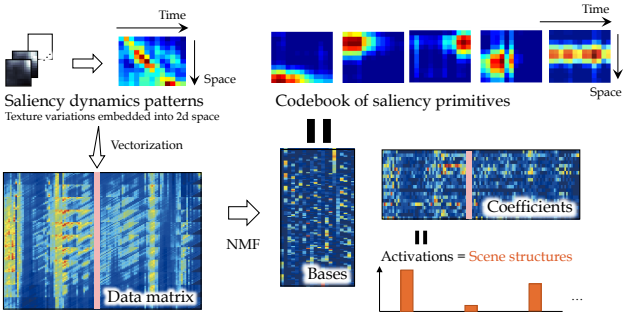

Fig. 8 Learning a codebook of saliency primitives.

$S_{\mathcal{N}(\boldsymbol{p},t)}$ contains the texture variations of saliency maps in a spatiotemporal patch, which is regarded as saliency dynamics in this model. If $\Omega_{\delta_\mathrm{s}}=\Omega$, $S_{\mathcal{N}(\boldsymbol{p},t)}$ leads to the description of overall scene structures like the OSDM. Otherwise, i.e., $\Omega_{\delta_\mathrm{s}}\subset\Omega$, we can describe local scene structures in a given patch.

To avoid diversity of the saliency dynamics, we particularly focus on their amplitude when extracting the texture variations. In other words, we introduce an orientation-invariant description for the extracted texture variations. Specifically, we first look for an axis in a spatial domain to describe the amplitude of texture variations the best. We calculate the absolute inter-frame differences in $\mathcal{S}_{\mathcal{N}(\boldsymbol{p},t)}$ and sum them up over time to obtain the history of variations. We then approximate the history by massive samples and apply the principal component analysis to obtain two principal components in the spatial domain, where the first principal component describes an orientation of the maximum variation of the history. Finally, we sum up the degrees of saliency over the second principal component for every frame to get the 2-d representation of the texture variations $L_{\mathcal{N}(\boldsymbol{p},t)}$ (see also Figure 7).

2.3.2 Learning a codebook of saliency primitives

Given many samples of saliency dynamics extracted in the procedure above, we learn a codebook consisting of saliency primitives that describe localized parts of the patterns. Since the saliency dynamics characterize scene structures consisting of multiple visual events, they can contain the mixture of several dynamics. For this reason, a standard model of dynamic textures that introduces a single LDS for each spatiotemporal patch such as [15] is not always appropriate for our situations. Instead, mixture models such as [16] can describe such dynamics with a set of sub-models. The PSDM that utilizes saliency primitives to describe parts of the dynamics can be regarded as the latter approach. In what follows, we aim to learn a codebook of saliency primitives effectively via matrix factorization.

Let us denote a flatten vector of saliency dynamics $L_{\mathcal{N}(\boldsymbol{p},t)}$ as $l_{\mathcal{N}(\boldsymbol{p},t)}\in\mathbb{R}_+^J$ where $J=(2\delta_\mathrm{s}+1)\cdot(2\delta_\mathrm{t}+1)$. We introduce a codebook consisting of N saliency primitives, $\mathcal{D}=\{D_1,\ldots,D_N\}$, where $D_n\in\mathbb{R}_+^J$ is the flatten vector of primitive patterns defined in the same spatiotemporal domain as $\boldsymbol{l}_{\mathcal{N}(\boldsymbol{p})}$. Then, $\boldsymbol{l}_{\mathcal{N}(\boldsymbol{p},t)}$ can be described with $\boldsymbol{w}(\boldsymbol{p},t)=(w_1,\ldots,w_N)^\mathrm{T}\in\mathbb{R}_+^N$, where w_n is the degree of activation for primitive D_n (i.e, how strongly D_n appears).

To learn codebook $\mathcal{D}$, we adopt a non-negative matrix factorization (NMF) [23] (see Figure 8). NMF plays an effective role in face analysis [23], music transcription [24], document clustering [25], etc. It decomposes a non-negative ma-

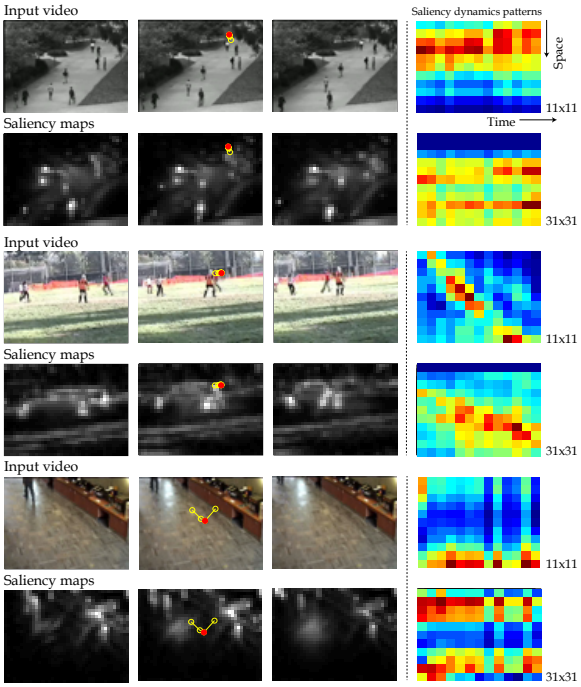

Fig. 9 Examples of saliency dynamics in spatiotemporal patches of different sizes. The patterns in the 4th column are extracted at gaze points of a single observer, which is denoted as the red points in the 2nd column of input images and saliency maps. Parts of the images in this figure are contained in the dataset provided by [8], [28], [29].

trix into two non-negative factors, where one factor consists of localized and structured bases and the other has activation coefficients like Figure 8. Let us introduce a $J \times N_{\mathrm{sp}}$ data matrix containing N_{sp} samples of saliency dynamics patterns, $\mathcal{L} = (\boldsymbol{l}_{\mathcal{N}(\boldsymbol{p}_1,t_1)}, \ldots, \boldsymbol{l}_{\mathcal{N}(\boldsymbol{p}_{N_{\mathrm{sp}}},t_{N_{\mathrm{sp}}})})$. Then, NMF derives the two factors as $\mathcal{L} = \bar{\mathcal{D}}W + \mathcal{E}$, where $\bar{\mathcal{D}} = (D_1, \ldots, D_N)$, a $J \times N$ basis matrix, represents a sequence of saliency primitives (that is, the codebook $\mathcal{D}$), $W = (\boldsymbol{w}(\boldsymbol{p}_1, t_1), \ldots, \boldsymbol{w}(\boldsymbol{p}_{N_{\mathrm{sp}}}, t_{N_{\mathrm{sp}}}))$, an $N \times N_{\mathrm{sp}}$ coefficient matrix, is corresponding activations, and $\mathcal{E}$ is a residual. We estimate $\bar{\mathcal{D}}$ and W by adopting multiplicative update rules [26].

2.3.3 Examples

We employed ASCMN database [27] as examples of unedited natural videos, which contained 24 videos consisting of surveillance videos, videos of human crowds, etc. The Itti's model [12] was adopted to obtain input saliency maps where the features include the luminance, color and orientation. We particularly focused on the local scene structures as a unique product of the PSDM compared to the OSDM, and investigated several sizes of patches: $(\delta_{\mathrm{x}}, \delta_{\mathrm{y}}, \delta_{\mathrm{t}}) = (5\ \text{pixel}, 5\ \text{pixel}, 0.4\ \text{sec})$, and $(15\ \text{pixel}, 15\ \text{pixel}, 0.4\ \text{sec})$, where the videos were first resized into 80×60 pixel resolution. Note that the spatial sizes of patches were 11×11 pixel and 31×31 pixel in the above settings. The size of codebook N was empirically set to $N = 20$.

Figure 9 depicts selected examples of extracted saliency dynamics as well as corresponding videos and saliency maps. These patterns were extracted at the point where a single observer looked at, and describing local scene structures in a spatiotemporal patch. The points of gaze, the red points in the 2nd column of the figure, are located at the center point of the patterns in the 4th column due to the definition of $\mathcal{N}(\boldsymbol{p}, t)$. These examples demonstrate that the points of gaze are not always directed to the most salient locations in a spatiotemporal patch, such as 5th and 6th rows in Figure 9. In other words, there are sometimes spatiotemporal gaps between saliency and gaze dynamics. The yellow points in the 2nd row of Figure 9 describe gaze scanpaths around the red gaze points, which indicate the large gaze motions can provide large spatiotemporal gaps. In this way, the local scene structures modeled by the PSDM can contribute to the analyses of the event-level spatiotemporal gaps. We will revisit these phenomena in Section 5.

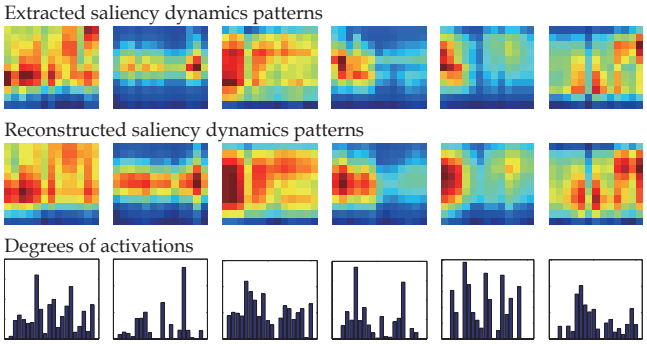

Fig. 10 Extracted and reconstructed patterns of saliency dynamics and the degrees of activations.

Figure 10 shows the comparison between extracted patterns and reconstructed ones from learned primitives as well as the degrees of activations for each pattern. As discussed in Section 2.2.4, the variations of saliency cannot always be continuous. However, several discontinuities were smoothed in the reconstructed patterns as shown in the right of figure. It indicates that our model can derive brief patterns while avoiding noises.

3. Attentional Target Identification Using Temporal Synchronizations

This section and following two sections are aimed at assessing the effectiveness of our framework by describing spatiotemporal correlations and evaluating them via practical gaze behavior analyses in real environments. Although the saliency dynamics models introduced in Section 2 allow us to handle various visual events and time-varying scene structures using saliency primitives, this section first adopts manually-designed videos with a constant scene structure and saliency primitives given preliminarily. Thanks to this simplification, we can concentrate on the evaluation of spatiotemporal correlations.

3.1 Event-level spatiotemporal gaps for attentional target identification

In this section, we particularly address event-level spatiotemporal gaps and investigate how they appear in actual gaze behavior. Imagine the situations where we are browsing dynamic contents with visual events like Figure 11. In the example, three items generate visual events (object translations) in a certain temporal interval. When we examine one of them (the center one in the example), a reaction to the events will appear in our gaze dynamics almost at the same time. This is a temporal synchronization between visual events and gaze reactions, and we aim to describe it by the event-level spatiotemporal gaps in our frame-

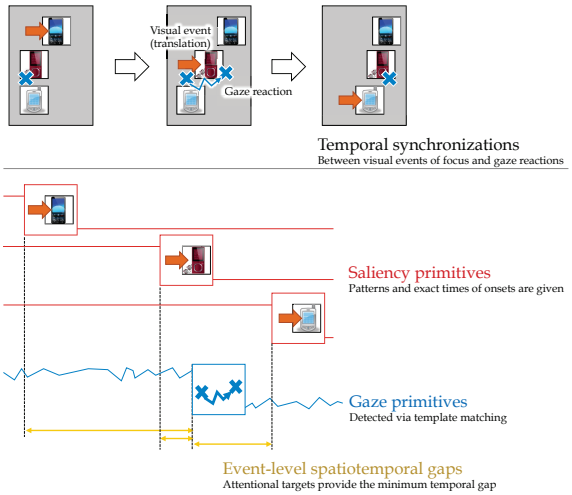

Fig. 11 Describing event-level spatiotemporal gaps. The temporal distances between saliency and gaze primitives represent the temporal synchronizations between visual events and gaze reactions.

work. Specifically, suppose first that the spatiotemporal patterns caused by the visual events are represented by saliency primitives and the patterns of primitives as well as the exact times that the primitives appear are given. Then, we detect gaze primitives corresponding to the reactions by matching the template reflecting the patterns of primitives. Finally, we can calculate temporal distances between the onset times of saliency and gaze primitives as a descriptor of temporal synchronizations (see Figure 11).

We leverage this synchronization for the task of identifying attentional targets from visual contents with several distinct objects (attentional target identification). Intuitively, the most naive approach is to use the spatial locational relationships between the objects and the points of gaze; given regions of objects, we can identify targets by judging which object region is the closest to the points of gaze. However, this approach is not always effective when gaze tracking systems involve a large measurement error. To solve this problem, we propose a identification method based on the temporal synchronizations, which we refer to as the *Gaze Probing* [30]. The Gaze Probing regards the objects with saliency primitives that provide the minimum spatiotemporal gaps to reactions as attentional targets. Since gaze tracking errors affect the only template matching to detect gaze primitives of reactions, we can achieve a robust identification by designing saliency primitives and templates appropriately.

3.2 Overview of the Gaze Probing

Let us denote a set of objects in dynamic contents as $\{O_c | c = 1, ..., C\}$. These objects are supposed to be distinguished from each other so as to be easily tracked by observers, while they are possibly overlapped to each other or out of frame temporarily. We denote properties of the c-th object region as $\boldsymbol{\theta}_t^{(c)} \in \mathbb{R}_+^J$ and their spatiotemporal pattern as $\Theta^{(c)} = (\boldsymbol{\theta}_1^{(c)}, \ldots, \boldsymbol{\theta}_T^{(c)})$.

For now we do not adopt saliency dynamics models to discover saliency primitives and instead manually design and embed them into object motions. We denote designed primitives as $D = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{\delta_t})$. Then, we embed multiple instances of D in $\Theta^{(c)}$, where the i-th onset of primitives is located at $t_i^{(c)}$. That is, $\Theta^{(c)}$ is partially defined as $\boldsymbol{\theta}_t^{(c)} = \boldsymbol{d}_{t-t_i^{(c)}+1} \quad \left(t_i^{(c)} \leq t \leq t_i^{(c)} + \delta_t - 1\right)$. Note that the remaining parts of $\Theta^{(c)}$ can be interpolated arbitrarily so as not to obtain more saliency than the primitives.

The Gaze Probing measures temporal synchronizations between visual events and gaze reactions as the event-level spatiotemporal gaps (specifically, temporal distances) between the onsets of designed primitives and those of gaze primitives detected from gaze data. To investigate the temporal synchronizations clearly, we first design overall scene structures so that all the primitives embedded in multiple objects must have temporally different onsets to each other with an enough margin. Namely, for arbitrary pairs of objects $O_c, O_{c'}$ $(c \neq c')$ and pairs of IDs i and i', $t_i^{(c)}, t_i^{(c')}, t_{i'}^{(c)}$ and $t_{i'}^{(c')}$ must satisfy $\left|t_i^{(c)} - t_{i'}^{(c)}\right| \geq \varepsilon$ and $\left|t_i^{(c)} - t_i^{(c')}\right| \geq \varepsilon$, where the minimum margin, ε, should be large enough to distinguish it from a reaction delay. Such scene structures allows us to discriminate the designed primitives in synchronization from those provided by the others.

Once we detect the onset of gaze primitives corresponding to gaze reactions at frame T_{react}, we can calculate the temporal distances between the onsets as event-level spatiotemporal gaps. Specifically, we introduce an evaluation score for each instance of designed primitives such as $V_i^{(c)} = |T_{\text{react}} - t_i^{(c)}|$. Then, target $O_{\hat{c}}$ can be identified as $\hat{c} = \arg\min_c V_i^{(c)}$. Practically, we set threshold ϵ_{th} to $V_i^{(c)}$ in order to avoid irrelevant synchronizations. Namely, if $V_i^{(c)}$ is larger than the threshold, we regard the corresponding reaction as false positive detection.

3.3 Example

We implemented a simple dynamic content depicted in Figure 12 to evaluate the effectiveness of the Gaze Probing. In this content, each object has a saliency primitive named *onset of horizontal scrolls* illustrated in Figure 13 (1), which can be well reflected in gaze dynamics and thus easy for us to detect the onset of corresponding gaze primitives of reactions, T_{react}, by matching the template shown in Figure 13 (2). Note that this design contributes to the robustness to gaze tracking errors obviously, since the template matching needs not use vertical gaze positions which contain larger errors than horizontal in many cases [31], [32].

In our experiment, each object displayed specific items (cellular phones and their description) and six participants were asked to choose the most interesting item from the objects for 60 sec. Figure 14 demonstrates an example of gaze data when looking at the content in Figure 12. Although the gaze dynamics observed in the experiment (above of the figure) were sometimes suffered from gaze tracking errors, we can still detect gaze reactions (* marks in the bottom of the figure) that synchronize with embedded primitives being looked at (o marks on the dashed line). From the overall gaze data of 360 sec collected from all the participants, 56 gaze primitives were detected (93.3% accuracy). Precision, the ratio of correct identification of attentional targets and the number of overall detected gaze primitives, was 76.8% while the baseline that identifies targets by comparing the spatial distance between object locations and the point of gaze marked 41.9%.

Consequently, we can identify attentional targets accurately from the event-level spatiotemporal gaps between designed saliency primitives and detected gaze primitives of reaction. We

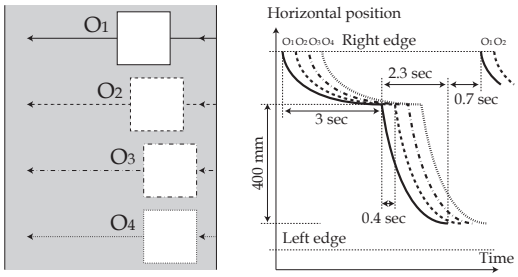

Fig. 12 Designed content with saliency primitives. This figure is a part of author's publication [30] copyrighted by Human Interface Society Japan.

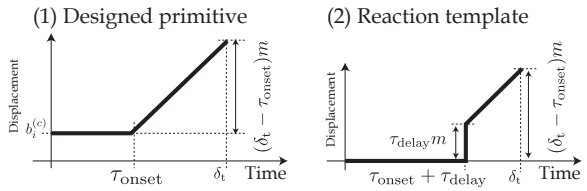

Fig. 13 Saliency primitive and the corresponding reaction template.

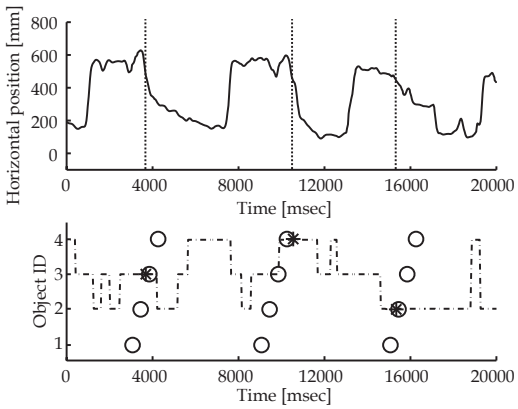

Fig. 14 Examples of gaze data and identification results. Above: gaze data (solid line) and reactions (dot line), below: designed primitives (o), reactions (*) and the verified object (dashed line). This figure is a part of author's publication [30] copyrighted by Human Interface Society Japan.

can introduce arbitrary content designs as long as they contain characteristic saliency primitives. [30] discusses the requirements of saliency primitive designs and confirms the effectiveness of the Gaze Probing using another designed content.

4. Attentive State Estimation based on Video Scene Structures

This section is aimed at assessing our framework under situations where observers are watching intentionally-designed videos such as TV commercial films. Namely, the videos we use here involve time-varying scene structures due to various types of visual events including scene changes. Thus, we now introduce the OSDM and try to handle visual events and scene structures with help from saliency primitives discovered by the model.

4.1 Feature extraction for scene-level correlations

Within our framework, we particularly focus on scene-level correlations between scene structures and gaze dynamics, which is the other aspect of spatiotemporal correlations that was not addressed in the previous section. The aim here is to describe how the scene-level correlations can be characterized differently depending on the time-varying types of scene structures. To this end, we first classify saliency primitives and gaze primitives into several types based on their spatiotemporal patterns. Then, the types of scene structures can be featured by the combinations of saliency primitive types. In addition, we refer to this information when extracting features that describe scene-level correlations effectively as follows:

Gaze-based feature extraction focuses on how specific types of gaze primitives can be characterized when looking at a certain type of saliency primitives.

Saliency-based feature extraction examines which types of saliency primitives originally tend to be looked at in a certain type of scene structures.

As for a task of gaze behavior analyses, this section addresses attentive state estimation that classifies if observers concentrate on displayed videos or not. The proposed descriptions of scene-level correlations and scene structures are effectively utilized for this task as below. First, we train discriminative models of attentive states with features of scene correlations for each type of scene structures. Then, given new video and gaze data, we adaptively apply the trained models based on the identified types of scene structures. It enables us to estimate attentive states when watching videos while considering time-varying scene structures.

4.2 Overview of the proposed method

4.2.1 Formulation

As for the basis of our attentive state estimation, we follow a traditional approach to mental state estimation based on a supervised learning framework such as [5]. It begins with extraction of features from gaze data such as frequencies of saccades and durations of fixations. At the same time, feature samples in a training dataset are given one of the several labels indicating discrete mental states. Then, the mental state estimation is formulated as a problem of learning a discriminative model for these labels.

Let us introduce gaze data $X = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_T)$. We denote feature vectors extracted from X as $\boldsymbol{\varphi}(\boldsymbol{X}) \in \mathbb{R}^{N_{\text{feat}}}$, where N_{feat} is the number of features. In addition, we consider discrete labels $A \in \{A_1, \ldots, A_{N_{\text{state}}}\}$, where N_{state} is the number of mental states. The estimation can be then formulated as a classification problem based on the posterior probability of A with observation $\boldsymbol{\varphi}(\boldsymbol{X})$:

$$\hat{A} = \arg\max_{A} P\left(A \mid \boldsymbol{\varphi}(X)\right). \tag{2}$$

Different from traditional approaches that work only when scene structures are given or constant, ours can deal with the situations where the scene structures dynamically change in an uncontrolled manner. Specifically, the proposed method incorporates scene structures and scene-level correlations derived by the OSDM in Section 2.2 into the formulation in Equation (2). Let us assume that X is split into $(X_1 \ldots, X_K)$ based on scene segmentation $\mathcal{I} = (I_1, \ldots, I_K)$. Each interval I_k has a set of spatiotemporal patterns of salient regions, $\boldsymbol{\Theta}_k = \{\Theta_k^{(1)}, \ldots, \Theta_k^{(C_k)}\}$, where $\boldsymbol{\Theta}_k^{(c)}$ is modeled by saliency primitive $D_k^{(c)}$. We classify $D_k^{(c)}$ into several types and describe the types of scene structures by the combinations of types identified to the primitives. Specifically, let us consider a set of possible saliency primitive types $\mathcal{W} = \{w_1, \ldots, w_N\}$.

Given a scene structure modeled by a set saliency primitives in the k-th interval, $\mathcal{D}_k = \{D_k^{(1)}, \ldots, D_k^{(C_k)}\}$, we first classify $D_k^{(c)}$ into one of several types, which is denoted as $W_k^{(c)} \in \mathcal{W}$. Then, the type of scene structures at the k-th interval is modeled as a vector consisting of histogram counts of $W_k = \{W_k^{(1)}, \ldots, W_k^{(C_k)}\}$, which is denoted as $hist(W_k)$. Finally, we use $\mathcal{D}_k$ and $hist(W_k)$ to modify Equation (2) as follows:

$$\hat{A}_k = \arg\max_A P\left(A \mid \boldsymbol{\varphi}(X_k, \mathcal{D}_k), hist(W_k)\right),$$

where $\boldsymbol{\varphi}(X_k, \mathcal{D}_k) \in \mathbb{R}^{N_{\text{feat}}}$ is a feature vector describing scene-level correlations between gaze dynamics X_k and scene structures $\mathcal{D}_k$. This formulation describes an adaptive estimation of attentive states based on time-varying types of scene structures, $hist(W_k)$.

4.2.2 Gaze-based feature extraction

Gaze-based feature extraction aims to describe the characteristics of scene-level correlations with the object of "how specific types of gaze primitives can be characterized when looking at a certain type of saliency primitives". As assumed in the previous section as well, gaze primitives basically reflect spatiotemporal patterns of saliency primitives being focused on. Thus, we identify the types of gaze primitives based on those of saliency primitives of focus. To this end, we first classify the types of saliency primitives so that different types of gaze primitives can be observed according to the types of saliency primitives. Then, we extract different features for the types of gaze primitives.

Specifically, we first classify saliency primitives into two types, static and dynamic, based on their average translation speed via k-mean clustering. Then, gaze primitives when looking at static and dynamic saliency primitives are labeled fixations and pursuits, respectively. Moreover, we regard the changes of saliency primitives being looked at as saccades.

Once we identify the types of gaze primitives, unique features are extracted for each of them (see [33] for detail). First, fixations contain internal gaze shifts to scan objects. We suppose that such shifts occur more actively when observers are in a higher level of attentiveness, and thus we introduce the size and the frequency of the shifts as features. As for features of pursuits, we extract the synchronization of speeds between gaze shifts and the motions of salient regions. When humans track a moving object, they tend to synchronize the pursuit acceleration to the expected changes of target motions and maintain the velocity at a constant level as long as the target velocity is not expected to change [34]. In addition to the features presented above, we introduce the frequency of saccades as features.

Finally, extracted features are aggregated into a vector based on the types of saliency primitives contained in an observed scene structure. Namely, we add features for fixations and pursuits if scene structures contain static and dynamic primitives, respectively. In addition, we add features for saccades if multiple primitives exist in the scenes.

4.2.3 Saliency-based feature extraction

The saliency-based feature extraction aims to describe "which types of saliency primitives originally tend to be looked at in a certain type of scene structures". In other words, we investigate what types of visual events tend to be looked at in the light of saliency. The saliency primitives achieved by fitting the OSDM indicate various types of visual events such as translations, resizes and the variations of saliency. Since it is difficult to introduce prior knowledge on which types of visual events frequently appear in videos and furthermore how much they tend to attract eyes, we introduce the classification of primitives that preserves all the properties as far as possible. Then, we define the saliency-based feature of scene-level correlations.

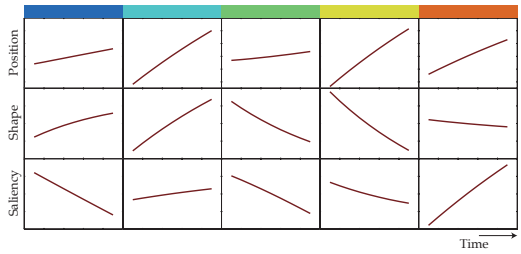

Fig. 15 Example of representative primitives for each type.

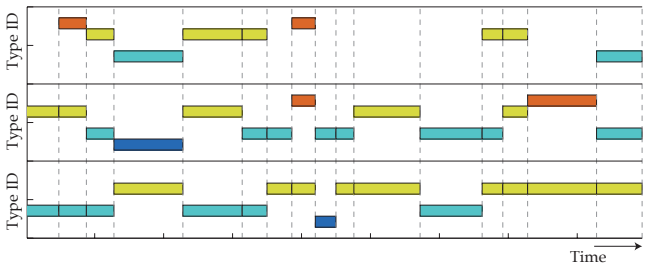

Fig. 16 Example of identified saliency primitive types corresponding to Figure 5. The colors and vertical positions of each rectangle describe the ID of identified types, where the colors correspond to Figure 15. The combination of types in each interval (split by dotted lines) defines the types of scene structures.

We classify saliency primitives via a hierarchical clustering. As a feature of the primitives, we first generate a fixed-length spatiotemporal pattern from the primitives and apply principal component analysis to obtain the variations of positions, shapes and saliency. A dissimilarity between two saliency primitives utilized for clustering is then defined as the sum of the correlations of those variations. As a result of the clustering based on the dissimilarity, we can visualize representative primitives for each type by identifying a single saliency primitive from spatiotemporal patterns of the same type. Figure 15 shows an example of representative primitives when the number of types N is set to $N = 5$. These primitives describe various visual events defined by the combinations of translations, resizes and variations of saliency. In addition, Figure 16 depicts selected identification results of saliency primitive types corresponding to Figure 5, where the colors correspond to the types in Figure 15. Although representative primitives in Figure 15 do not always describe the original primitives in Figure 5 accurately when N is small, we can still classify scene structures into several types based on the combination of saliency primitive types in a data-driven manner.

Once we classify the types of saliency primitives, we can leverage those information for feature extraction from gaze data. Specifically, we utilize spatial locational relationships between saliency primitives and gaze points to learn the types of saliency primitives that tend to be looked at in a soft-assignment fashion. As depicted in Figure 17, the spatial relationships are given as a

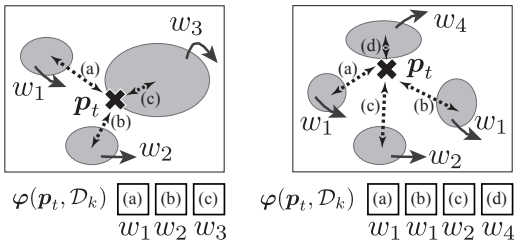

Fig. 17 Extracting spatial locational relationships between saliency primitives and gaze points for features.

Table 2 Estimation results

Method	Baseline	M_G	M_S
Accuracy [%]	66.4	70.2	78.7
Coverage [%]	100	100	52.2

set of distances between the locations of saliency primitives and those of gaze, where the primitives are aligned based on an ascending order of their types.

4.3 Experiments

Experimental setups

We recorded gaze data of ten participants during watching 12 TV commercial films in several conditions of attentiveness. Since these videos contained frequent scene changes, participants' gaze was basically expected to concentrate on salient regions that attracted their exogenous attention, although some endogenous actions like examinations and switches of attentional targets were likely to occur. In the experiments, we asked the participants to follow the instructions below so that they were able to freely watch videos as far as possible in high/low level of attentiveness.

Task 1 (high level of attentiveness) : Please watch a video and answer the questionnaire to evaluate how much you liked the video on a seven-point scale.

Task 2 (low level of attentiveness) : Please watch a video while doing the following calculation task at the same time; please keep on subtracting 7 from 1000 and report answers (1000, 993, ...) to the experimenter.

Tasks 1 and 2 corresponded to high and low attentive states, respectively. Above all, Task 2 made participants conduct a secondary task (i.e., the calculation) to decrease the attention resource to the video-viewing task. See [18] for more details.

As for evaluation measures, we adopted *accuracy*: the ratio of intervals given a correct estimate of attentive states and *coverage*: the ratio of intervals characterized by types of scene structures observed more than one time during experiments. The coverage becomes high when unseen videos contain previously trained scene types, indicating the effectiveness of method in terms of generalization capability. We compared gaze-based and saliency-based features (M_G and M_S). In addition, we also employed a baseline method that utilizes all the gaze-based features without any distinction of the types of saliency and gaze primitives.

Results and discussions

Table 2 described the scores of all the methods. These results demonstrate the effectiveness of utilizing scene-level correlations in terms of predicting attentive states. Particularly, the saliency-based feature extraction works better if we can assume all the

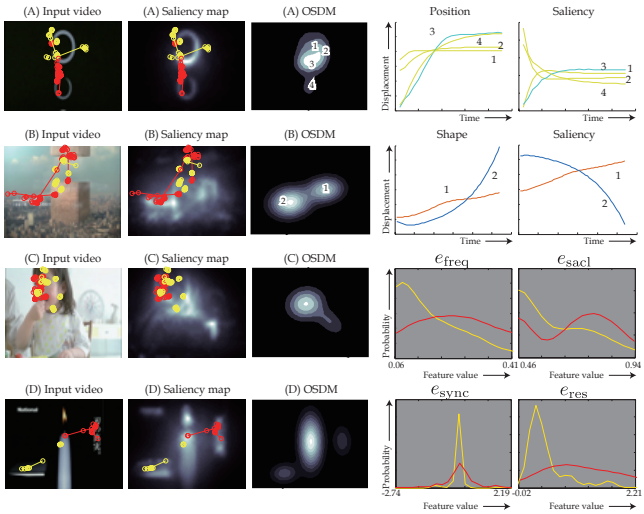

Fig. 18 Estimation results. 1st column: input videos, 2nd column: saliency maps, and 3rd column: fitting results of the OSDM. The red and yellow points indicate subsequences of gaze points (gaze points at ±3 frames) for all the participants under high and low attentive states, respectively. In Examples (A) and (B), the 4th and 5th columns depict selective properties of saliency primitives shown in the titles, where the numbers from the 3rd to 5th columns indicate the ID of saliency primitives. In addition, the color of lines are the ID of the primitive types described in Figure 15. In Examples (C) and (D), the 4th and 5th columns describe the estimated probability distributions for the selected gaze features shown in the titles, where the color of lines correspond to the points of gaze in the 1st and 2nd columns. The images used in this figure were provided by courtesy of Panasonic Corporation.

videos are given and trained preliminary, while the gaze-based extraction has the advantage of being applicable to unseen videos.

Figure 18 depicts selected examples of estimation results. In the 1st and 2nd columns, color points show subsequences of gaze points (gaze points at ±3 frames) for all the participants, where red and yellow show high and low attentive states, respectively. The 3rd column contains fitting results of the OSDM. When participants looked at different regions for the levels of attentiveness, the saliency-based features work effectively as shown in Examples (A) and (B). The 4th and 5th columns of these examples depict selected properties of saliency primitives where the color of lines shows the types of the primitives described in Figure 15. In Example (A), gaze points under the high level of attentiveness (red) concentrated on the 3rd and 4th saliency primitives. These primitives correspond to the appearance event of an object with a large translation, while the other primitives being looked at under the low level of attentiveness describe smaller translation events. Example (B) has two saliency primitives in its scene structure, and the distributions of gaze points differ for the levels of attentiveness. The 2nd region, which tended to be looked at more frequently when participants were in the high level of attentiveness, corresponds to a text caption with visual events of losing saliency due to an appearance event of a new object from the top of frame. Although the semantic meaning of region (i.e., text caption) is invisible in the proposed method, we can capture the tendency of gaze behavior from the viewpoint that what types of saliency primitives are attracting eyes.

Examples (C) and (D) show the estimated probability distri-

butions of several gaze-based features that contributed to the estimation. Since gaze points in Example (C) concentrated on a face regardless of attentive states, it is difficult to introduce the saliency-based features for this situation. However, there were differences in the frequency of gaze shifts when fixating the face and that of saccades as shown in the 4th and 5th columns of the example. Specifically, participants tended to provide gaze shifts and saccades more frequently when they were highly attentive. Example (D) describes another tendency of gaze behavior when pursuing objects with translation events. As shown in the 4th and 5th columns, participants tended to pursue moving targets with a more constant ratio of speeds and directions in when they were in the low level of attentiveness. Alternatively, participants tended to examine objects with translation events more actively when they were highly attentive to videos.

In conclusion, the proposed framework leverages saliency to describe how observers watch videos depending on the level of attentiveness while considering time-varying scene structures of the videos. The experiments demonstrate that our framework successfully works when observers' gaze is mostly exogenous, and it is our future work to evaluate the framework with more semantically complex videos that cause endogenous gaze actions.

5. Gaze Point Prediction from Spatiotemporal Correlations

In the previous two sections, we focused on event-level spatiotemporal gaps and scene-level correlations separately based on the proposed framework. While we analyzed the spatiotemporal gaps provided by a single type of gaze primitives in Section 3, the degree of gaps can vary depending on the types of gaze primitives; in Figure 9, a large gap occurred particularly when gaze shifted larger. Moreover, the gaps are also affected by the types of visual events and furthermore, the time-varying scene structures like Section 4. For example, sudden motions of objects among many static objects can provide a large reaction delay. Consequently, the event-level spatiotemporal gaps can be affected by the scene-level correlations consisting of scene structures and gaze dynamics. The aim of this section is to describe the overall spatiotemporal correlations of the aforementioned characteristics based on the proposed framework.

5.1 Spatiotemporal correlations for gaze point prediction

As a practical situation where gaze behavior exhibits the spatiotemporal correlations, we assume a free-viewing of a more variety of videos than previous sections, including unedited natural ones such as surveillance videos. Since those videos do not always contain distinct objects that can be easily followed by observers nor frequent scene changes, eyes can be sometimes directed to irrelevant locations. Figure 19 depicts an example of the above situation. Although the video displays a bus that can be a salient region, several gaze points (depicted as red points) could not follow it and provided a gap since the bus contained a fast translation event from the left to the right of frames. Moreover, several points remained the right of the frame even if the bus has disappeared. Obviously, this example shows the spatiotemporal correlations: an event-level spatiotemporal gap reflecting

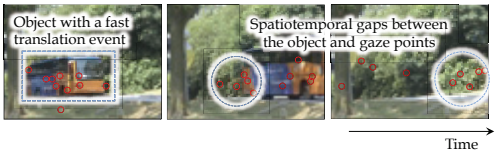

Fig. 19 Example of spatiotemporal correlations when watching videos. Parts of the photos in this figure are contained in the dataset provided by [29]. Red points indicate ground truths of gaze points, where each point corresponds to one individual observer in [27].

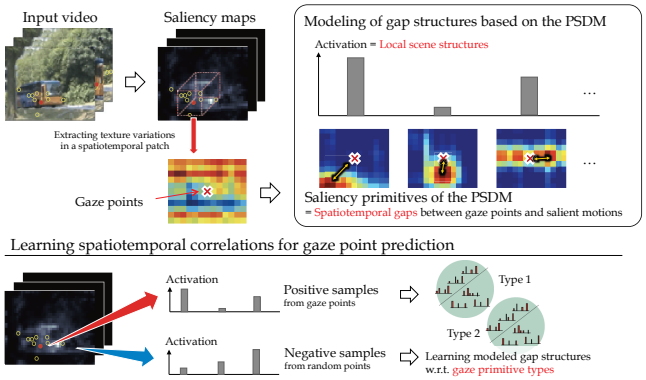

Fig. 20 Describing spatiotemporal correlations based on the gap structure model. Parts of the photos in this figure are contained in the dataset provided by [29].

a scene-level correlation consisting of specific saliency and gaze dynamics (i.e., the fast translation event and reaction pursuit).

Towards the description of overall spatiotemporal correlations, we first introduce a model to describe the relationships between spatiotemporal gaps and scene structures affecting the gaps, which we refer to as *gap structures*. Specifically, we leverage saliency primitives of the patch-based saliency dynamics model (PSDM) to describe both gaps and scene structures jointly (see Figure 20). Then, we statistically learn the modeled gap structures around the points of gaze for each type of gaze primitives so that we can involve their scene-level correlations with gaze dynamics. Intuitively, we learn the gap structures for fixations, pursuits and saccades individually. Finally, the learned relationships between gap structures and gaze primitive types describe overall spatiotemporal correlations consisting of event-level spatiotemporal gaps and scene-level correlations.

We leverage the proposed description for the task of gaze point prediction from videos. While we follow traditional learning-based saliency maps (LBSM) that just predict if a certain point tends to be looked at in a learning fashion (e.g., [35], [36], [37]), the proposed method is novel in terms of (1) predicting gaze while considering its spatiotemporal gaps such as reaction delays and (2) predicting gaze while considering the type of gaze primitives.

5.2 Overview of the proposed method

5.2.1 Formulation

The gaze point prediction is a task to predict where observers tend to look in each frame of videos. More specifically, we are to generate a prediction map where each pixel-value indicates the degree of gaze-point existence. The LBSM originally involves a supervised learning framework with a set of saliency-related features and gaze data as a training dataset. Let us denote a point

of gaze in a dataset as $\boldsymbol{p} \in \mathbb{R}^2$ and features extracted from $\boldsymbol{p}$ as $\boldsymbol{\varphi}(\boldsymbol{p}) \in \mathbb{R}^{N_{\text{feat}}}$, where N_{feat} is the number of features. The LBSM aims to provide the degree of gaze-point existence at all the pixels as a continuous value, $B(\boldsymbol{p}) \in \mathbb{R}$. Namely, it predicts where observers tend to look in a map form for each video frame. We refer to the map as a *gaze-prediction map* to distinguish it from saliency maps. Since videos contain multiple frames, the final output is a sequence of gaze-prediction maps.

As for a model of $B(\boldsymbol{p})$, we introduce the following function:

$$B(\boldsymbol{p}) = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\varphi}(\boldsymbol{p}), \tag{3}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{N_{\text{feat}}}$ is parameters of the model. We estimate $\boldsymbol{\beta}$ in a discriminative model [35], [36], [37]; $\boldsymbol{p}$ in the training dataset is given a label consisting of $\{1, -1\}$, where 1 is the positive label indicating the point tends to be looked at and -1 corresponds to a negative label for the little probability of being looked at. Then, $\boldsymbol{\beta}$ can be trained as parameters of discriminant function $B'(\boldsymbol{p})$ of the following form: $B'(\boldsymbol{p}) = \mathrm{sgn}(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\varphi}(\boldsymbol{p}) + \beta_0)$, where $\beta_0 \in \mathbb{R}$ is a bias term. Positive samples are often collected from the points of gaze in a training dataset. On the other hand, the negatives can be practically collected from random points.

5.2.2 Introducing Spatiotemporal Correlations

In this study, we introduce a model of gap structures that describes the relationships between event-level spatiotemporal gaps and scene structures. We exploit the modeled gap structures for feature description $\boldsymbol{\varphi}(\boldsymbol{p})$ in Equation (3) to predict gaze while considering time-varying scene structures in videos as well as spatiotemporal gaps that can appear when looking at specific visual events. In addition, by learning the model with respect to each type of gaze primitives, we can introduce the scene-level correlations between scene structures and gaze dynamics.

Modeling gap structures

We first introduce the assumption that the degree of gaze-point existence is particularly affected by visual events around the points of gaze. Such an assumption can be often seen in traditional studies on saliency maps such as [12] that refer to local center-surround contrasts of visual stimuli. On that basis, we consider a local scene structure defined in a certain spatiotemporal patch when introducing gap structures.

Specifically, gap structures indicate what types of salient motions can be observed in a local scene structure around gaze points and how much spatiotemporal gaps appear against those motions. As a bottom-up approach to their modeling, we utilize saliency primitives of the PSDM presented in Section 2.3. In the PSDM, saliency primitives in codebook $\mathcal{D} = \{D_1, \ldots, D_N\}$ describe localized texture variations of saliency in a spatiotemporal patch. In other words, they indicate motion patterns and relative positions of salient regions. Then, given gaze point $\boldsymbol{p}$ as a center point of the patch, activation vector $\boldsymbol{w}(\boldsymbol{p}) = (w_1, \ldots, w_N)^{\mathrm{T}}$ describes local scene structures around gaze points while each primitive contains spatiotemporal distances between the gaze points and salient motions (in what follows, we omit subscription t from the original definition of $\boldsymbol{w}(\boldsymbol{p}, t)$ and $\mathcal{N}(\boldsymbol{p}, t)$ without loss of generality). As for gaze point prediction, we utilize the activation vector $\boldsymbol{w}(\boldsymbol{p})$ as feature vector $\boldsymbol{\varphi}(\boldsymbol{p})$. Namely, the estimation of $\boldsymbol{\beta}$ can be regarded as a problem of finding specific types of saliency primitives from codebook $\mathcal{D} = \{D_1, \ldots, D_N\}$ which have different tendencies in their appearances between the points of gaze and random points.

Incorporating the types of gaze primitives

To involve scene-level correlations, we learn the modeled gap structures with respect to each type of gaze primitives. Then, we calculate gaze prediction maps for all the types of gaze primitives, which individually indicate where observers tend to look with a certain gaze primitive type. The obtained maps of each primitive type are finally integrated into single gaze-prediction maps. As a simple approach, we introduce the assumption that each type of gaze primitives can be observed with equal probability, independently and identically for spatial and temporal directions.

Specifically, let us first denote the types of gaze primitives as $\mathcal{E} = \{e_1, \ldots, e_{N_{\text{etype}}}\}$, where N_{etype} is the number of the types. By identifying gaze primitive types to each gaze point, $\boldsymbol{p}$ is given a label $g(\boldsymbol{p}) \in \mathcal{E}$ if $\boldsymbol{p}$ is a point of gaze and otherwise it is given a negative label. We then train the models with respect to each type of gaze primitives from positive samples with label e_w and negative samples collected from random points to estimate parameters $\boldsymbol{\beta}_{e_1}, \ldots, \boldsymbol{\beta}_{e_{N_{\text{etype}}}}$. As a result, the degree of gaze-point existence with gaze primitive type e_w is evaluated as $B_{e_w}(\boldsymbol{p}) = \boldsymbol{\beta}_{e_w}^{\mathrm{T}} \boldsymbol{\varphi}(\boldsymbol{p})$. Finally, we integrate model outputs over e_w to obtain the degree of gaze point existence: $B_E(\boldsymbol{p}) = \frac{1}{N_{\text{etype}}} \sum_{w=1}^{N_{\text{etype}}} B_{e_w}(\boldsymbol{p})$.

Identification of gaze primitive types

In Section 4, we identified types of gaze primitives based on observed types of saliency primitives since we assumed overt attention was basically oriented to salient regions. However, this assumption is not always appropriate for the current situation since observers can look at irrelevant locations unconsciously. We therefore take a bottom-up approach to identify the types of gaze primitives. Specifically, we classify the types based on the motion speeds of gaze shifts in each gaze primitive obtained by a sliding window approach. As a result, the types of gaze primitives can be associated with biological definitions of eye movement types: e.g., fixations, pursuits and saccades.

5.3 Experiments

Experimental setups

In experiments, we adopted several combinations of public datasets ([28] and [27]) and saliency maps (IT: Itti's model [12], RC: Cheng's model [38] and TR: Torralba's model in [35]) to investigate if the effectiveness can be consistent regardless of videos and input saliency (see [22] for detail). In order to evaluate a generalization capability on videos, we conducted a leave-one-out scheme by splitting data based on video IDs. As for an evaluation measure, we introduced the normalized scanpath saliency (NSS) [39] that first normalized prediction maps and evaluated the degree of saliency (the degree of gaze point existence in this study) at given gaze points.

In order to assess the effectiveness of (1) gap structures and (2) those learned with respect to each type of gaze primitives, we here tested original saliency maps ($\mathrm{M}_{\mathrm{ORIG}}$), gap structures learned without distinction of gaze primitive types (M_{BU}) and those learned for each gaze primitive type ($\mathrm{M}_{\mathrm{BU+E}}$). As an alternative naive approach to compensate the spatiotemporal gaps be-

Table 3 Average NSS scores over videos.

		M_{ORIG}	M_{SM}	M_{SM+E}	M_{BU}	M_{BU+E}
CRCNS	IT	0.752	0.859	0.847	1.135	**1.208**
	RC	0.927	1.002	1.021	1.152	**1.212**
	TR	0.742	0.858	0.886	1.100	**1.152**
ASCMN	IT	0.623	0.745	0.741	0.876	**0.900**
	RC	0.603	0.659	0.651	0.765	**0.775**
	TR	0.388	0.465	0.466	0.774	**0.817**

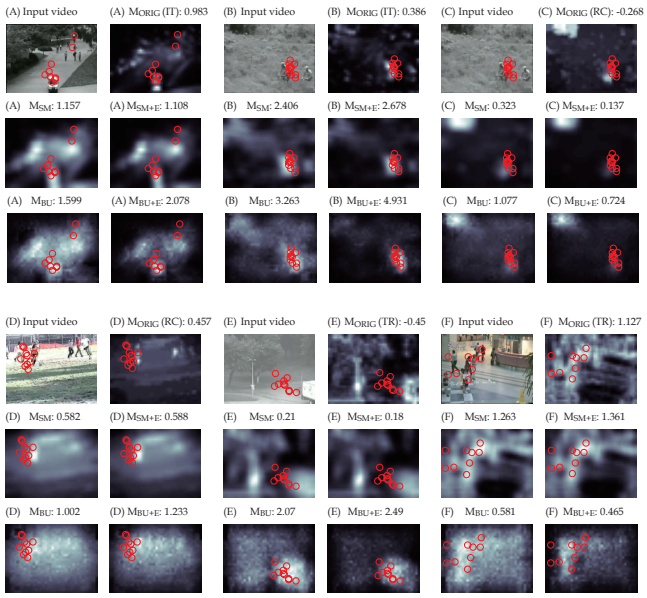

Fig. 21 Qualitative results and corresponding NSS scores averaged over observers in a frame. Luminance indicates the degree of gaze-point existence. Red points indicate a set of gaze points, where each point corresponds to an individual observer in [27]. Parts of the photos in this figure are contained in the dataset provided by [8], [28], [29].

tween saliency primitives and gaze, we introduced another baseline method that just smoothes input saliency maps where the parameter for smoothing was tuned via cross validation regardless of gaze primitive types (M_{SM}) and for each type of gaze primitives like the proposed method (M_{SM+E}). As for the types of gaze primitives, we empirically set N_{etype} to 4.

5.3.1 Results and discussions

Table 3 shows NSS scores for all the conditions. These results demonstrated the effectiveness of M_{BU}, M_{BU+E} compared to the baseline methods (M_{ORIG}, M_{SM}, M_{SM+E}). Although the NSS scores of the baseline methods had a variation with regard to the saliency maps, the scores of our methods were very competitive. It indicates the independence of our models to input saliency maps to describe gap structures. Comparing methods with or without the consideration of gaze primitive types, the proposed method M_{BU+E} only performed improvements from M_{BU} while M_{SM+E} from M_{SM} shows slight differences.

Figure 21 depicts qualitative results of gaze-prediction maps and NSS scores. These results demonstrate the following characteristics of the proposed method:

Proposed method vs. baseline methods In Examples (A) and (E), a car was running out of the frame, and most of observers were trying to pursue it. In addition, Example (D) shows the situation where observers pursue the player's running. Obviously there are gaps between the points of gaze and the targets in both examples, and thus baseline methods providing a large degree of gaze-point existence at the targets get low NSS scores. On the other hand, our method incorporates such gaps and provide a large degree of gaze-point existence where the points of gaze exist and succeeded in significantly improving the NSS scores.

Differences in saliency maps Examples (B) and (C) depict the comparison of different saliency maps, IT and RC. RC is a method to look for small superpixels that contain a rare color, and thus the baseline methods show high responses at the black regions in the top-left of a frame. That brings the significant differences in NSS scores not only in the baselines but in the proposed method, although averaged scores in Table 3 show small differences among saliency maps in the proposed method.

Failure cases In Example (F), the original saliency map was able to capture the points of gaze precisely. Even such cases, the proposed method tries to consider a spatiotemporal gap since there are many samples with gaps in training datasets, which sometimes provides large degree of gaze-point existence at inappropriate locations and decreases a score.

Comparing M_{BU+E} with M_{BU}, highlighted regions indicating a large degree of gaze-point existence are more sparse in M_{BU+E} as shown in Examples (A), (B), (E), (F) in Figure 21. We can observe such results when one of the prediction scores for a certain type of gaze primitives is particularly high. With regard to Example (E), Figure 22 visualizes each of model outputs by the difference of color. In the 3rd row of the figure, we gave each pixel a 3-d value $(B_{e_1}(\boldsymbol{p}), 0.5(B_{e_2}(\boldsymbol{p}) + B_{e_3}(\boldsymbol{p})), B_{e_4}(\boldsymbol{p}))$ in an RGB order where each of them roughly corresponds to fixations, pursuits and saccades. When there was a target in motion, model outputs of pursuits (green) became much higher than the others, which made final outputs more sparse. In addition, there was also a small probability of observing saccades (blue). For example, saccades can be found when trying to attend the target (points at the left side of the frame in the 3rd column) or escaping from the target (those at the bottom-right in the 4th column).

Finally, this study introduce a simple assumption that gaze primitive types can appear with equal probability, independently and identically for spatial and temporal directions. On the other hand, gaze primitive types at a certain point can be statistically conditioned by those at its spatiotemporal neighborhood. As for this problem, one of the promising approaches is to introduce state-space models such as [40]. Then, we can dynamically select models to be used based on the gaze primitive types which are likely to appear.

6. Conclusions

In this study, we presented a novel framework to describe the spatiotemporal correlation between video and gaze data. The proposed framework can contribute to the analyses of various gaze behavior in real environments including but hopefully not limited to the situations where humans watch videos although cognitive and neurological reasoning is invisible. Future work will seek to extend the framework to more interactive situations such as human robot interaction, driving and human conversations.

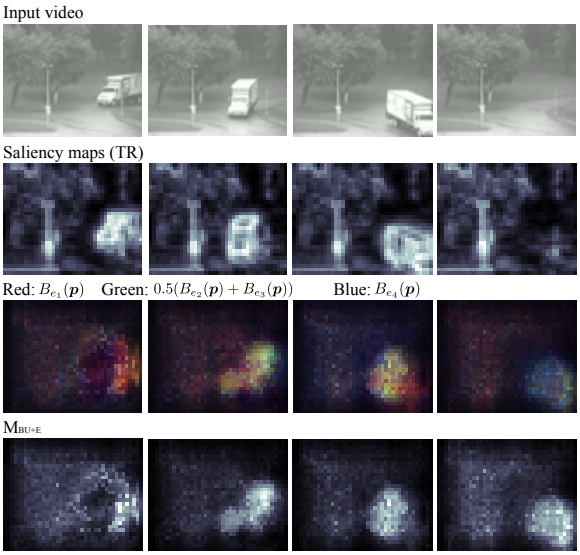

Fig. 22 Differences in outputs of the proposed method trained for each type of gaze primitives. In the 3rd row, red points show a large degree of gaze-point existence for $B_{e_1}(\boldsymbol{p})$, green for $0.5(B_{e_2}(\boldsymbol{p}) + B_{e_3}(\boldsymbol{p}))$ and blue for $B_{e_4}(\boldsymbol{p})$. Parts of the photos in this figure are contained in the dataset provided by [8]. This figure is a part of author's publication [22] copyrighted by Association for Computing Machinery.

Acknowledgment

This work is in part supported by Grant-in-Aid for Scientific Research under the contract of 24·5573.

References

[1] Park, H. S., Jain, E. and Sheikh, Y.: 3D Gaze Concurrences From Head-mounted Cameras, *NIPS* (2012).
[2] Fathi, A., Hodgins, J. K. and Rehg, J. M.: Social Interactions: A First-person Perspective, *CVPR* (2012).
[3] Simola, J., Salojärvi, J. and Kojo, I.: Using Hidden Markov Model to Uncover Processing States from Eye Movements in Information Search Tasks, *Cognitive Systems Research*, Vol. 9, No. 4, pp. 237–251 (2008).
[4] 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).
[5] Bednarik, R., Vrzakova, H. and Hradis, M.: What Do You Want to Do Next : A Novel Approach for Intent Prediction in Gaze-based Interaction, *ETRA*, pp. 83–90 (2012).
[6] Tseng, P.-H., Cameron, I. G., Pari, G., Reynolds, J. N., Munoz, D. P. and Itti, L.: High-throughput Classification of Clinical Populations from Natural Viewing Eye Movements, *Journal of Neurology*, Vol. 260, No. 1, pp. 275–284 (2013).
[7] Simonin, J., Kieffer, S. and Carbonell, N.: Effects of Display Layout on Gaze Activity During Visual Search, *HCI*, Vol. 3585, pp. 1054–1057 (2005).
[8] Mahadevan, V. and Vasconcelos, N.: Spatiotemporal Saliency in Dynamic Scenes, *TPAMI*, Vol. 32, No. 1, pp. 171–177 (2010).
[9] Cerf, M., Harel, J., Einhäuser, W. and Koch, C.: Predicting Human Gaze Using Low-Level Saliency Combined with Face Detection, *NIPS*, pp. 1–8 (2007).
[10] Subramanian, R., Yanulevskaya, V. and Sebe, N.: Can Computers Learn from Humans to See Better?: Inferring Scene Semantics from Viewers' Eye Movements, *ACMMM*, pp. 33–42 (2011).
[11] Kimura, A., Yonetani, R. and Hirayama, T.: Computational Models of Human Visual Attention and Their Implementations: A Survey, *IEICE Trans. on Information and Systems*, Vol. E96-D, No. 3, pp. 562–578 (2013).
[12] Itti, L., Koch, C. and Niebur, E.: A Model of Saliency-based Visual Attention for Rapid Scene Analysis, *TPAMI*, Vol. 20, No. 11, pp. 1254–1259 (1998).
[13] Bregler, C.: Learning and Recognizing Human Dynamics in Video Sequences, *CVPR*, pp. 568–574 (1997).
[14] North, B., Blake, A., Isard, M. and Rittscher, J.: Learning and Classification of Complex Dynamics, *TPAMI*, Vol. 22, No. 9, pp. 1016–1034 (2000).
[15] Doretto, G., Chiuso, A., Wu, Y.-N. and Soatto, S.: Dynamic Textures, *IJCV*, Vol. 51, No. 2, pp. 91–109 (2003).
[16] Chan, A. and Vasconcelos, N.: Modeling, Clustering, and Segmenting Video with Mixtures of Dynamic Textures, *TPAMI*, Vol. 30, No. 5, pp. 909–926 (2008).
[17] Zhan, B., Monekosso, D. N., Remagnino, P., Velastin, S. A. and Xu, L.-Q.: Crowd Analysis: A Survey, *Machine Vision and Applications*, Vol. 19, No. 5-6, pp. 345–357 (2008).
[18] Yonetani, R., Kawashima, H., , Kato, T. and Matsuyama, T.: Modeling Saliency Dynamics for Viewer State Estimation (in Japanese), *IEICE Trans. on Information and Systems*, Vol. J96-D, No. 8, pp. 1675–1687 (2013).
[19] Cotsaces, C., Nikolaidis, N. and Pitas, I.: Video Shot Detection and Condensed Representation: A Review, *IEEE Signal Processing Magazine*, Vol. 23, No. 2, pp. 28–37 (2006).
[20] Witkin, A. P.: Scale-space Filtering, *IJCAI*, pp. 1019–1022 (1983).
[21] Harel, J., Koch, C. and Perona, P.: Graph-based Visual Saliency, *NIPS*, Vol. 19, pp. 545–552 (2007).
[22] Yonetani, R., Kawashima, H. and Matsuyama, T.: Predicting Where We Look from Spatiotemporal Gaps, *ICMI* (2013).
[23] Lee, D. and Seung, H.: Learning the Parts of Objects by Non-negative Matrix Factorization, *Nature*, Vol. 401, No. 6755, pp. 788–791 (1999).
[24] Smaragdis, P. and Brown, J.: Non-negative Matrix Factorization for Polyphonic Music Transcription, *WASPAA* (2003).
[25] Xu, W., Liu, X. and Gong, Y.: Document Clustering Based on Non-negative Matrix Factorization, *SIGIR* (2003).
[26] Lee, D. and Seung, H.: Algorithms for Non-negative Matrix Factorization, *NIPS* (2001).
[27] Riche, N., Mancas, M. and Culibrk, D.: Dynamic Saliency Models and Human Attention: A Comparative Study on Videos, *ACCV* (2012).
[28] Itti, L. and Baldi, P.: Bayesian Surprise Attracts Human Attention, *Vision Research*, Vol. 49, No. 10, pp. 1295–1306 (2009).
[29] Li, L., Huang, W., Gu, I. Y. and Tian, Q.: Statistical Modeling of Complex Backgrounds for Foreground Object Detection, *IEEE Trans. on Image Processing*, Vol. 13, No. 11, pp. 1459–1472 (2004).
[30] Yonetani, R., Kawashima, H., Hirayama, T. and Matsuyama, T.: Gaze Probing: Event-based Estimation of Objects Being Focused on (in Japanese), *The Transaction of Human Interface Society*, Vol. 12, No. 3, pp. 125–135 (2010).
[31] Zhu, Z. and Ji, Q.: Eye Gaze Tracking under Natural Head Movements, *CVPR*, Vol. 1, pp. 918–923 (2005).
[32] Chen, J., Tong, Y., Gray, W. and Ji, Q.: A Robust 3D Eye Gaze Tracking System Using Noise Reduction, *ETRA*, pp. 189–196 (2008).
[33] Yonetani, R., Kawashima, H., Hirayama, T. and Matsuyama, T.: Mental Focus Analysis Using the Spatio-temporal Correlation between Visual Saliency and Eye Movements, *Journal of Information Processing*, Vol. 52, No. 12 (2012).
[34] 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).
[35] Judd, T., Ehinger, K., Durand, F. and Torralba, A.: Learning to Predict Where Humans Look, *ICCV* (2009).
[36] Kienzle, W., Franz, M. O., Schölkoph, B. and Wichmann, F. A.: Center-surround Patterns Emerge as Optimal Predictors for Human Saccade Targets, *Journal of Vision*, Vol. 9, No. 5, pp. 1–15 (2009).
[37] Borji, A.: Boosting Bottom-up and Top-down Visual Features for Saliency Estimation, *CVPR* (2012).
[38] Cheng, M., Zhang, G., Mitra, N., Huang, X. and Hu, S.: Global Contrast Based Salient Region Detection, *CVPR* (2011).
[39] Parkhurst, D., Law, K. and Niebur, E.: Modeling the Role of Salience in the Allocation of Overt Visual Attention, *Vision Research*, Vol. 42, No. 1, pp. 107–123 (2002).
[40] Pang, D., Kimura, A., Takeuchi, T., Yamato, J. and Kashino, K.: A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network, *ICME* (2008).