Modeling Video Viewing Behaviors for Viewer State Estimation (Authors Version)

Ryo Yonetani
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto-shi
Kyoto 606-8501, Japan
yonetani@vision.kuee.kyoto-u.ac.jp

ABSTRACT

ACM, (2012). This is the authors version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in the proceeding of ACM Multimedia 2012 Doctoral Symposium (ACMMM 2012 DS).

Human gaze behaviors when watching videos reflect their cognitive states as well as characteristics of the video scenes being watched. Our goal is to establish a method to estimate the viewer states from his/her eye movements toward general videos, such as TV news and commercials. The proposed method is based on a novel model of video viewing behaviors, which takes into account structural and statistical relationships between video dynamics, gaze dynamics and viewer states. This model realizes statistical learning of gaze information while considering dynamic characteristics of video scenes to achieve viewer-state estimation. In this paper, we present an overview of the viewer-state estimation method based on the model of video-viewing behaviors, including several past work done by the author's team.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems—*Human information processing*;

General Terms

Algorithm, Experimentation

Keywords

saliency dynamics, gaze behavior, viewer state estimation

1. INTRODUCTION

Eyes are a window into the mind — eye movements reflect various cognitive states. A study on the eye movements having a long history in the field of visual psychology revealed many important findings including the similarity of eye movements toward the same images or the dissimilarity depending on human states (e.g., given tasks) [11]. These findings indicate the possibility to realize an estimation of the human states based on statistical learning of gaze information. Recently, several studies have proposed a method to estimate cognitive states and given tasks [1, 4, 10, 12, 13].

One of the difficulties in eye-movement studies derives from the fact that eye movements are affected by the scenes being watched. It is clear that spatial layouts of objects in the scenes and categories of objects affect where and what tend to be looked at. In order to control such influences from characteristics of the scenes, previous studies [1, 4, 10] mainly deal with designed contents or relatively simplified scenes. On the other hand, it still remains unclear how to realize the estimation while considering complex contents, especially general videos (e.g., TV news, commercials) that contain various types of objects and varying characteristics over time. In such case, it is crucial to understand the dynamic characteristics of the videos.

In this study, we aim to establish a method of viewer-state estimation from eye movements of viewers watching general videos. The main contribution of this work is that we propose a novel model of video-viewing behaviors, which takes into account structural and statistical relationships between videos-scene dynamics, gaze dynamics and states of video viewers (see Figure 1). Statistical learning of the relation-

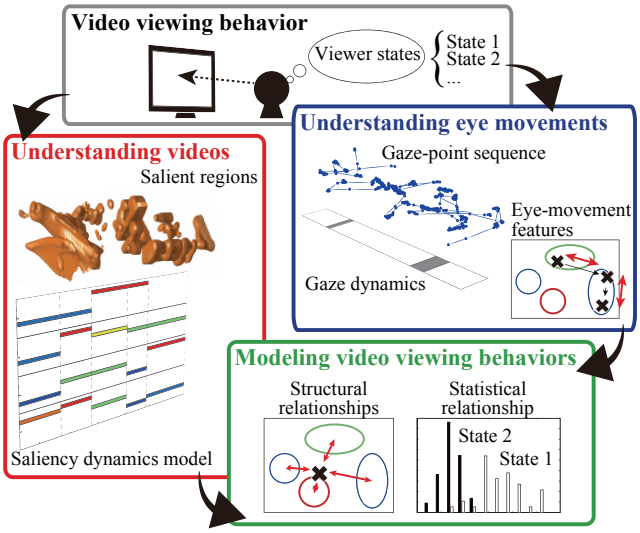

Figure 1: Modeling video-viewing behaviors.

MM'12, October 29–November 2, 2012, Nara, Japan.

ships enables us to estimate the viewer states from newly observed pairs of the video scenes and eye movements.

The proposed method has many helpful applications. For multimodal interaction between systems and users, viewer-state estimation allows the systems to provide information to the users in a timely manner. Moreover, since viewer states toward displayed contents can be regarded as an evaluation of themselves, the proposed method has a potential to give a novel indication to human-centric content designs.

2. PROBLEM SETTING

Suppose that videos are displayed on a screen, and a human is watching the videos. Eye movements of the viewer can be observed as a sequence of gaze points on the screen by using an eye tracker. Viewer states are here assumed to be represented by one of several discrete states, such as "focusing on a certain object in a video", "paying attention to a video", and further, "favoring a video". Viewer-state estimation is a problem of classifying the viewer states based on the given video and gaze information.

Since eye movements are affected by both viewer states and dynamic characteristics of the video scenes, we assume that the eye movements or their structural relationships to the scene characteristics (e.g., a spatial relationship between gaze points and objects) are statistically conditioned by both of them. Let S and E be features of video-scene characteristics and of gaze dynamics (or the structural relationships) respectively, and let A be one of the viewer states. When E is statistically learned under the condition of both S and A, the unknown state $\hat{A}$ is estimated from a newly-observed $\hat{S}$ and $\hat{E}$ via a maximum likelihood estimation:

$$\hat{A} = \arg\max_{A} p(E = \hat{E} \mid S = \hat{S}, A).$$

Based on the above assumptions, This study tackles the following 2 topics:

(A) Understanding dynamic characteristics of videos what kinds of video characteristics should be employed, and how to model their dynamics? (Description of S)

(B) Understanding eye movements How to model gaze dynamics or their relationships to the video characteristics? (Description of E)

Our previous studies [13] and [12] have addressed those topics. The following sections overview the achievements presented in those literatures.

3. SALIENCY DYNAMICS MODEL FOR VIDEO-SCENE DESCRIPTION

Understanding dynamic changes of video characteristics is an important topic in the field of pattern recognition and computer vision. As known as a problem in generic object recognition, general videos contain objects with a diversity of categories as well as poses and positions. This diversity causes a critical problem for the viewer-state estimation when considering both video characteristics and eye movements because it is infeasible to statistically learn gaze features for each type of diverse video scenes.

To overcome this problem, we utilize saliency of videos, a characteristic of videos without its semantics, as a mean to describe video scenes. Videos have several salient regions attracting viewer gazes, which contain dynamics in their position, shape, and strength of saliency. We extract such saliency dynamics from videos, and classify them as several typical patterns to describe the video scenes. The proposed method can decrease the number of video-scene types (i.e., types of saliency dynamics) to be considered, and realize statistical learning of gaze features with regard to each type of scenes even if they originally contain diversity.

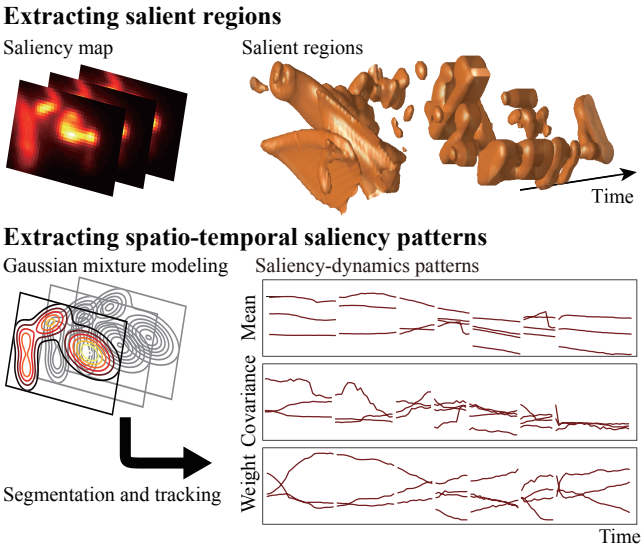

Figure 2: Spatio-temporal saliency patterns.

The rest of this section briefly introduces the model of saliency dynamics proposed in our study [13].

Extraction of saliency patterns.

Video frames often contain several salient regions, the number of which can vary over time. Each salient region individually contains dynamics in its position, shape and strength of saliency. The saliency dynamics which we aim to model is the one described by such spatio-temporal patterns of multiple salient regions.

In order to model the dynamics of region and its texture (i.e., saliency distribution) simultaneously, we introduce a GMM to describe the multiple salient regions contained in a video frame. This modeling sacrifices the representation of detailed contours and textures of regions. However, the GMM allows us to describe locations, approximate shapes, and strength of saliency of the regions by means, covariance and weights of the components, respectively. Figure 2 depicts an example of salient regions and their spatio-temporal patterns. We employ a model of visual attention [6] to obtain salient regions. The concrete procedure for extraction and parameterization of saliency patterns is presented in [13].

Modeling saliency dynamics.

Granted that we obtain saliency patterns from videos, it still remains unclear how to model them to describe the saliency dynamics of the video scene, S. Considering the utilization of scenes for a learning-based approach to the estimation, it is desirable that saliency dynamics are classified into a finite number of typical patterns to decrease the number of scene types enough to introduce statistical learning.

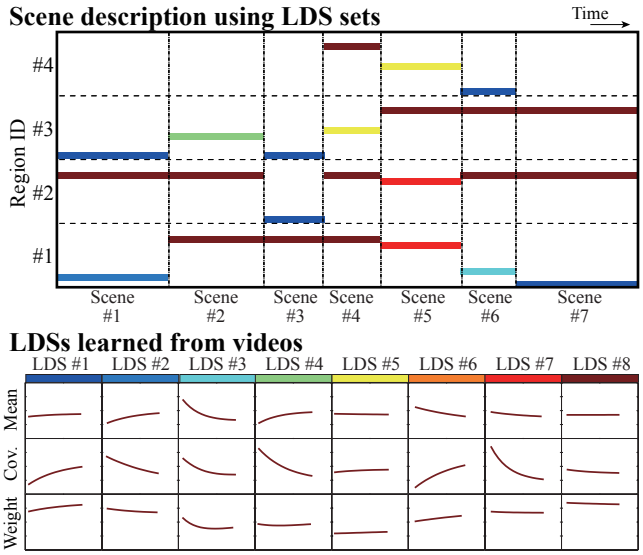

Figure 3: Applying the proposed saliency-dynamics model to the patterns described in Figure 2.

In [13], we propose a novel saliency dynamics model (SDM), which describes the dynamical changes of salient regions including the change of the number of regions by the switching of multiple linear dynamical systems (LDS). It is based on a switching linear dynamical system (SLDS), which is an efficient tool to represent complex human motion [3, 8, 9]. The SLDS models complex dynamics as switching between simpler dynamics, where each of the time evolutions of dynamics is formulated by an LDS. The SDM applies the SLDS to each dynamics pattern of salient regions. That is, multiple SLDSs are used for multiple salient regions, where each of the saliency patterns is modeled by a different SLDS.

By applying the SDM, the number of salient regions that simultaneously exist in a certain interval (i.e., the interval of a scene) is described by the number of LDSs identified in the interval. The saliency dynamics in the scene, S, are described by a set of LDSs. Figure 3 demonstrates an application of the SDM to saliency patterns presented in Figure 2.

4. MODELING GAZE DYNAMICS AND ITS RELATIONSHIP TO VIDEO DYNAMICS

The basic concept behind our method of viewer-state estimation is that different gaze dynamics can be observed depending on viewer states. When video-scene dynamics are given, the difference in gaze dynamics can be represented using the structural relationship between gaze dynamics and video dynamics. This section introduces several approaches to extract features from gaze dynamics and from the structural relationship for the viewer-state estimation.

Modeling gaze dynamics.

Gaze dynamics can be modeled by several primitive eye movements such as a fixation, a pursuit, and a saccade. Moreover, there are some observable types of eye movements depending on the characteristics of video-scene dynamics. For instance, fixations and pursuits can be observed when humans scan static and dynamic salient regions, respectively. In addition, saccadic eye movements tend to be observed when multiple regions exist in a frame. These eye movements are requested to be evaluated in a different way as presented in [1, 7] because they have different characteristics for their type.

Thus in [12], we first define the observable types of eye movements for several video-scene types identified empirically as presented in Table 1, and then extract primitive gaze features (e.g., fixation duration lengths, eye-motion speeds in pursuit, saccade frequencies. See [12] for the detail) for each of the eye-movement types. Gaze features for each type of the video-scenes are obtained by aggregating the primitive features based on the observed types of eye movements.

Table 1: Video-scene types and corresponding types of observable eye movements.

Video scenes	Observable eye movements
Single static region	Fixation
Single dynamic region	Pursuit
Multiple static regions	Fixation, saccade
Multiple static & dynamic regions	Fixation, pursuit, saccade
Multiple dynamic regions	Pursuit, saccade

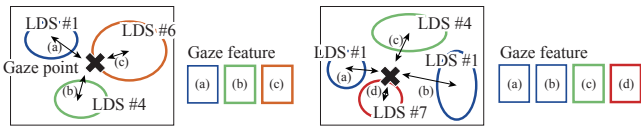

Figure 4: Spatial structures between gaze points and salient regions characterized by LDSs.

Modeling spatial structure.

The structural relationship between gaze dynamics and video dynamics can also be utilized as features to classify viewer states. Our previous study [13] presents a gaze feature consisting of the spatial structure between gaze points and salient regions (Figure 4). Since each region is characterized by LDSs after introducing the SDM, this gaze feature represents "what types of dynamics tend to be focused on".

5. EXPERIMENTS AND FUTURE WORK

5.1 Attentive-state estimation

To verify the proposed method, we conducted some experiments to estimate attentive states of viewers, which indicate how strong viewers pay attention to videos. In the experiments, we aimed to classify two levels of attentiveness: highly attentive to or distracted from videos, as a relatively-simplified evaluation. 10 subjects took part in the experiments, and 12 TV commercial videos were employed.

Experimental setup.

A subject is sitting in front of a screen, and an eye tracker is installed below the screen. We adopt the following two conditions in order to control the attentiveness:

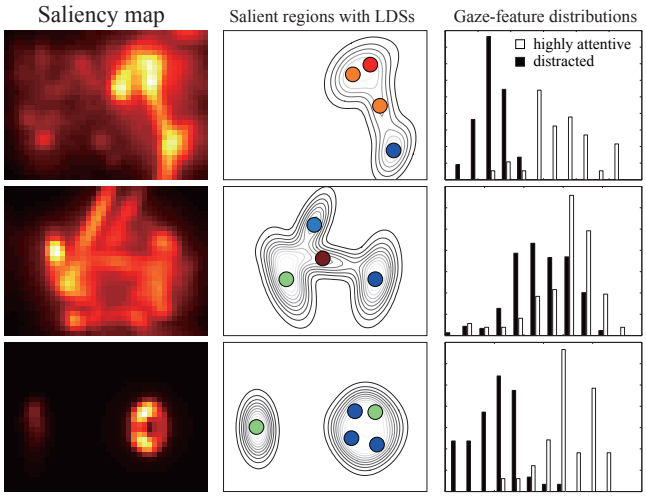

Figure 5: Example results. The color of circles in the 2nd column corresponds to the ID of LDSs in Figure 3. Gaze-feature distributions in the 3rd column are projected into a discriminant space for visualization. The white and black bars correspond to highly attentive and distracted states, respectively.

Table 2: Estimation accuracies.

Method	Duration [13]	[12]	[13]
Accuracy (%)	59.3	78.2	**80.6**

Condition 1 A subject watches a video and receives a simple interview after that.

Condition 2 A subject watches a video while doing a mental calculation.

The condition 1 and 2 correspond to the highly attentive and distracted states, respectively. Details about the experiment are presented in [12, 13].

Results and discussions.

Table 2 presents estimation accuracies from the literature [12, 13], with the baseline method that employs a gaze duration feature describing how long subjects look at salient regions. In addition, Figure 5 depicts some illustrative results consisting of saliency maps of input frames, salient regions characterized by the LDSs, and distributions of gaze features describing the spatial structures as proposed in Section 4. These results demonstrate that the proposed methods can work well even when the baseline has no clear discrimination of the two attentiveness levels.

As discussed in Section 2, the problem in this study consists of two topics: modeling of video and gaze dynamics (or the structural relationship between the gazes and videos). The SDM proposed in [13] focuses on the former problem and provides a detailed classification of saliency dynamics for video-scene description. By using the model in conjunction with spatial-structure features, it enables us to handle what types of dynamics tend to be focused on, while they are invisible in the other method [12].

On the other hand, the gaze features employed for the method in [12] reflect the dynamic characteristics of eye movements whereas the feature in [13] is obtained frame-wisely. Experimental results in [12] have revealed the following findings on eye movements and attentiveness:

- The levels of attentiveness affect saccades within target regions. Specifically, subjects seem to scan targets actively rather than change them when they are in the higher level of attentiveness.
- Synchronization in the speed between eyes and targets is not affected much by attentiveness. Subjects tend to pursue dynamic regions at any level of attentiveness.

Such dynamic aspects in eye movements are reported to play a crucial role to analyze human states [5].

Obviously, the two approaches are complementary, and they can be unified by introducing gaze features in [12] for video scenes obtained by the SDM in [13].

5.2 Future work

Our video-viewing behavior model focuses on structural and statistical relationships between saliency dynamics in videos, gaze dynamics and viewer states. Currently, any semantic information in the videos is invisible to our saliency dynamics model, which is closely related to human minds. As recent saliency models have tried to introduce some semantic concepts [2], future work will improve our model by handling such semantic concepts for better understanding of human video-viewing behaviors.

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

6. REFERENCES

[1] R. Bednarik, H. Vrzakova, and M. Hradis. What do you want to do next : A novel approach for intent prediction in gaze-based interaction. In *ETRA*, pages 83–90, 2012.

[2] A. Borji. Boosting bottom-up and top-down visual features for saliency detection. In *CVPR*, pages 1–8, 2012.

[3] C. Bregler. Learning and recognizing human dynamics in video sequences. In *CVPR*, pages 568–574, 1997.

[4] S. Eivazi and R. Bednarik. Predicting problem-solving behavior and performance levels from visual attention data. In *IUI*, pages 9–16, 2011.

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

[6] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 20(11):1254–1259, 1998.

[7] R. J. K. Jacob and K. S. Karn. Commentary on Section 4. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye: cognitive and applied aspects of eye movement research*, pages 573–605. Elsevier Science, 2003.

[8] Y. Li, T. Wang, and H. Shum. Motion texture: a two-level statistical model for character motion synthesis. *ToG*, 21(3):465–472, 2002.

[9] B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *TPAMI*, 22(9):1016–1034, 2000.

[10] J. Simola, J. Salojärvi, and I. Kojo. Using hidden Markov model to uncover processing states from eye

movements in information search tasks. *Cognitive Systems Research*, 9(4):237–251, 2008.
[11] A. Yarbus. Eye movements and vision. *Plenum*, 1967.
[12] R. Yonetani, H. Kawashima, T. Hirayama, and T. Matsuyama. Mental focus analysis using the spatio-temporal correlation between visual saliency and eye movements. *Journal of Information Processing*, 20(1):267–276, 2012.
[13] R. Yonetani, H. Kawashima, and T. Matsuyama. Multi-mode saliency dynamics model for analyzing gaze and attention. In *ETRA*, pages 115–122, 2012.