Shih, J. L. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Modeling Video Viewing Styles with Probabilistic Mode Switching

Hiroaki KAWASHIMA[a,b*], Kousuke UEKI[b] & Kei SHIMONISHI[b]
[a]*School of Social Information Science, University of Hyogo, Japan*
[b]*Graduate School of Informatics, Kyoto University, Japan*
*kawashima@sis.u-hyogo.ac.jp

Abstract: During video lectures, learners may have attentional modes such as "follow a lecturer's guide (speech and pointers)," "look ahead of spoken parts and actively check slide content," and "roughly browse a slide." The dynamic change of these modes is useful to characterize personal and/or temporary viewing styles. This paper presents a method to analyze video viewing styles through gaze behavioral data by using a probabilistic generative model with a latent mode variable. In our experiments, we show that the model can infer viewers' temporal mode patterns and successfully characterize task-dependent viewing situations.

Keywords: Eye-tracking, Video lectures, Viewing styles, Probabilistic gaze-behavior model

1. Introduction

While clickstream logs on video lectures (e.g., MOOCs) are promising sources to analyze learning behavior (Kim et al., 2014), those logs mainly contain users' intentional activities, such as page jumps, and it cannot be directly used to understand detailed learners' reaction to video content (e.g., slide information, lecture's speech, and pointing actions). Gaze data in video lectures, on the other hand, enable us to analyze learners' behavior deeply enough to infer which consisting regions attract/confuse viewers (Nguyen & Liu, 2016) and how much viewers followed a lecturer's speech (Sharma, Jermann, & Dillenbourg, 2014).

To understanding learners' performance and skills, distinctive spatio-temporal gaze patterns is often analyzed in gaze studies (Mangaroska, Sharma, Giannakos, Trætteberg, & Dillenbourg, 2018). This gives us insight that such gaze patterns partially reveal viewers' internal process, and in other words, there may exist useful "intermediate representation" between gaze patterns and human cognitive process. In Kawashima, Ueki, & Shimonishi (2019), a minimal model for viewers' attentional modes during video lectures, such as "roughly grasp slide content," "actively follow slide content," and "follow a lecturer's guide (speech and pointers)" is introduced. While these modes are not necessarily directly related to human attentional modes, inferred modes can be treated as useful intermediate representation to characterize viewers' situations. In this paper, we extend the work by introducing probabilistic mode switching model to analyze and visualize learners' viewing styles in video lectures.

2. Multi-mode Gaze Model

The model proposed in Kawashima et al. (2019) is a multi-mode generative model, where a viewer's gaze sequence is assumed to be affected by the three component submodels. We first briefly describe the model in this section and introduce its extension in Section 3 to model a variety of viewing styles.

2.1 Model

Assume that raw gaze data are converted into a sequence of areas-of-interest (AOIs), which we refer to as *gaze regions*, based on the velocity of eye movements (Salvucci & Goldberg, 2000). Suppose that $r_1, \dots, r_{k-1}, r_k, \dots, r_K$ is a gaze sequence of a viewer, where $r_k \in \{R_1, \dots, R_N\}$ is a region ID at time

step , and $R_1, \dots, R_N$ are region IDs in a slide. Note that *time* (or *time step*) k denotes an ordered number of AOI switches. Meanwhile, we use t_k to describe actual *media time* (physical time whose origin is the start time of a slide) in the video at time step k.

Mode 0 (base distribution): Attracted by salient regions such as high contrast or important terms, modeled by region distribution $P(r_k = R_i \mid m_k = 0) = a_i$,
Mode 1 (region order): Follow slide content by considering the meaning of content information, modeled by transitional probabilities $P(r_k = R_j \mid r_{k-1} = R_i, m_k = 1) = b_{ij}$,
Mode 2 (lecture's guide): Follow a lecturer's guide such as spoken words and pointers, modeled by time-dependent region distribution $P(\mathrm{r_k} = R_i \mid t_k = t, m_k = 2) = c_{it}$.

Then, the influence ratio of the above submodels (modes) is assumed to change dynamically:

$$P(r_k|\cdot) = P(m_k = 0|\cdot)P(r_k|m_k = 0) + P(m_k = 1|\cdot)P(r_k|r_{k-1}, m_k = 1) + P(m_k = 2|\cdot)P(r_k|t_k, m_k = 2),$$

where $(\cdot)$ denotes the previous gaze regions r_{k-1}, r_{k-2}, ... and timestamps $t_k, t_{k-1}, \dots$.

The probability $P(m_k = m)$, $m \in \{0, 1, 2\}$ describes the influence ratio of submodel m, whose sum is $\sum_{m=0}^{2} P(m_k = m) = 1$. Based on this model, typical situations can be described as follows: When a viewer follows the lecturer's guide (spoken words and pointers) at time *k*, the value of $P(m_k = 2)$ becomes higher than $P(m_k = 0)$ and $P(m_k = 1)$. Meanwhile, when a viewer follows the slide content actively by ignoring the lecturer's guide, $P(m_k = 1)$ becomes higher. In a time period of mind wandering, $P(m_k = 0)$ get higher due to not following either the content or the lecturer.

Note that the concept of mode $m = 2$ is closely related to the with-me-ness proposed in Sharma et al. (2014) in terms that the corresponding submodel tries to describe specific spatio-temporal points that a lecturer attracts gaze of learners.

2.2 Model Training

Suppose that a collection of gaze sequences $r_{\mathrm{seq}}^{(v)}$ $(v = 1, \dots, n)$ from n viewers of the same slide is given in a model training phase. Using this gaze data, the model parameters a_i, b_{ij}, and c_{it} is estimated based on the maximum-likelihood estimation. This estimation is not straightforward due to the hidden variables m_k, whose posterior needs to be estimated simultaneously. That is, this model training can be viewed as clustering of gaze patterns to three modes, and therefore the expectation-maximization (EM) algorithm can be utilized.

The algorithm repeats the E- and M-steps iteratively. In the E-step, the posterior $P\left(m_k^{(v)} \middle| r_{\mathrm{seq}}^{(v)}\right)$ for each time step k is computed using the model with current parameters during iterations. Then, all the parameters are updated in the M-step, where additive smoothing (in our experiments, additional 0.1 count) was used to avoid the zero-frequency problem.

3. Probabilistic Mode-Switching Model

3.1 Inference of Mode Sequences with a Probabilistic Mode Transition

Given a newly observed gaze data r_{seq}, a *mode posterior* $P(m_k = m \mid r_{seq}) = \gamma_k(m)$ $(m = 0, 1, 2, k = 0, 1, 2, \dots, K)$ can be obtained with a similar procedure as the E-step in the training phase. The posterior can be considered as the inferred ratio of the influence of each submodel. An inferred sequence of $\gamma_k(m)$ contains useful information of when and which mode the viewer took during the video viewing situations. For example, the value of $\gamma_k(2)$ is expected to be large for time k when the viewer focuses on the lecturer's talk (guide information).

To statistically analyze inferred mode sequences, we do not directly use the patterns of the sequences but utilize the temporal structure behind the patterns. Specifically, we extend the model

described in the previous section by introducing mode switching model with the following *mode-transition probability:*

$$P(m_k = q | m_{k-1} = p) = A_{pq},$$

where $p, q \in \{0, 1, 2\}$. Note that this extension corresponds to the model of state transition of hidden Markov models (HMMs), while its states (modes) and emission probabilities are specifically designed to describe viewers' situations as explained in Section 2.1.

3.2 Model Training

Now the extended model has a parameter set $\{a_i, b_{ij}, c_{it}, A_{pq}\}$. Since the goal of this study is to characterize video viewing styles, we introduce an assumption that these parameters can be divided into two: viewer-independent (shared) parameters $\theta_{\text{shared}} = \{a_i, b_{ij}, c_{it}\}$ and viewer-dependent parameters $\{A_{pq}^{(v)}\}$. By using the shared parameters, modes and their switching can be analyzed in the common space. Note that we add superscript (v) to explicitly denote that the mode-transition probabilities depend on viewer v.

For better convergence of the model parameters, we also divide the training into two steps. In the first step, viewer-independent parameters θ_{shared} are estimated without mode switching model (introduced in Section 2). Then, viewer-dependent parameters $\{A_{pq}^{(v)}\}$ are estimated for each of viewer v to characterize the structure of viewer v's mode transition. In this second step, all the viewer-independent parameters are fixed. As for initial mode probabilities $P(m_1)$, we used equal probabilities 0.5 for $m_1 = 0$, 2, and 0 for $m_1 = 1$.

3.3 Comparing Mode-Switching Structures

Once each viewer's mode-switching behavior is encoded by the model above, dissimilarity between two models can be introduced. This dissimilarity can be considered as a pseudo distance between two viewing patterns. Since the extended model is analogous to HMMs, we here use a distance measure proposed in Juang & Rabiner (1985), which utilizes Kullback-Leibler (KL) divergence.

Let $\theta^{(v_1)} = \theta_{\text{shared}} \cap \{A_{pq}^{(v_1)}\}$ and $\theta^{(v_2)} = \theta_{\text{shared}} \cap \{A_{pq}^{(v_2)}\}$ be the parameter sets of viewer v_1 and v_2, respectively. Then, the divergence-related value can be computed by using log likelihood:

$$D(v_1 || v_2) = \frac{1}{K^{(v_1)}} \left[\log P\left(r_{\text{seq}}^{(v_1)} \middle| \theta^{(v_1)}\right) - \log P\left(r_{\text{seq}}^{(v_1)} \middle| \theta^{(v_2)}\right) \right],$$

where $K^{(v_1)}$ is the length of gaze sequence $r_{\text{seq}}^{(v_1)}$. Considering the non-symmetric property of KL divergence, the pseudo distance can be defined as the following average:

$$\text{Dist}(v_1, v_2) = (D(v_1 || v_2) + D(v_2 || v_1))/2.$$

4. Experiments

In this experiment, we verify the proposed model in terms of its capability of characterizing and visualizing video viewing styles. To focus on the evaluation of the proposed model itself, we conducted laboratory experiments with designed settings. Specifically, we prepared not only (a) a normal video-viewing situation but two additional artificial situations: (b) with a sub task and (c) with an edited content (static slides with no sound). Since the ground truth of viewers' internal states cannot be obtained, we considered that these designed situations highly bias their internal states and simulate some extreme situations such as mind wandering (Mills, Bixler, Wang, & D'Mello, 2016; Hutt et al., 2017) or ignoring a lecturer's guide. Self-reporting or think aloud protocol is another option to obtain ground truth of internal process, but we do not take this option to avoid an additional task affecting gaze behavior.

4.1 Experimental Settings

33 university students were recruited to conduct the lab-setting research. Each of the students was explained a summary of our research and signed an informed consent form upon the arrival. The explanation of the research objective was abstract enough to avoid affecting their gaze behavior.

4.2 Tasks and Content

Video-viewing tasks were assigned to each of the 33 participants. They were asked to (i) watch a video on a monitor and (ii) answer several written questions related to the content of the watched video. The post-questions were prepared to make each participant concentrated enough on the video content and to measure the degree they could follow the content. Confidence of each answer were also collected in five-point Likert scale, and the prior-knowledge of each question was also asked to verify that most of the participants did not know the question-related content before the experiment.

As described in the beginning of this section, the participants were divided randomly into the following three groups (11 participants each):

(a) Group 1 (normal): No additional task was assigned.
(b) Group 2 (sub task): Additional task of mental calculation (repeatedly subtract 3 from 1000) was assigned from slide 2.
(c) Group 3 (no guide): No additional task but an edited video (a sequence of static slides with no sound) was displayed from the middle (slide 3-); the length of the presence of each slide was also edited according to the density of the content.

As for the sub task, participants were asked to vocalize each result of the mental calculation, but the vocalized values were not checked. The confidence of the answers to the post-questions was low enough in most of the questions, and the scores of Group 2 were lower than Group 1 for all the questions.

The topic of the content was “inferential statistics” from JMOOC gacco (https://gacco.org/), and the length of the video content, consisting of 4 slides, was about 10 minutes. The screen of the video consisted of a slide and a lecturer. The slide was partially and temporarily overlaid by the lecturer's arm and pointers. In this experiment, the 3rd slide was used, which contained only texts.

4.3 Data acquisition

Tobii X120 eye tracker was used to measure participants' gaze points on a screen with sampling rate 60 Hz. We here used this lower sampling rate because we focus on analyzing fixations rather than saccadic eye movements. Each participant was asked to sit in front of the screen where chin rest was used to reduce measurement noise as much as possible. While the used device was robust against head movements to some extent, we decided to use this setting to focus on the verification of the proposed model and algorithm. In the identification step of fixations, lack of data less than 150 ms was considered as instantaneous noise and interpolated using surrounding data.

While automatic region segmentation is possible to define regions, in this experiment, we manually segment regions based on the meaning of words for the sake of avoiding region segmentation errors. Region IDs were basically numbered from top left to bottom right, which roughly coincided with the order of reading the slide content.

4.4 Results

The two-step training described in Section 3 was applied to the acquired gaze sequences on the slide 3: The shared parameters θ_{shared} were estimated by normal-task data in the first step, and mode-transition probabilities were obtained for each of the viewers in Group 1-3 in the second step. During the second step, each viewer’s mode posterior sequence was also computed.

Figure 1 shows examples of mode sequences from each of the three tasks (Group 1-3). It can be seen that mode 1 and mode 2 are dominant in (a) normal task and (b) sub task while only mode 1 is dominant in (c) no guide. Compared to Group 1 (normal task), mode 0 appears more frequently in Group 2 (sub task). These trends can be seen consistently in most of other viewers’ sequences.

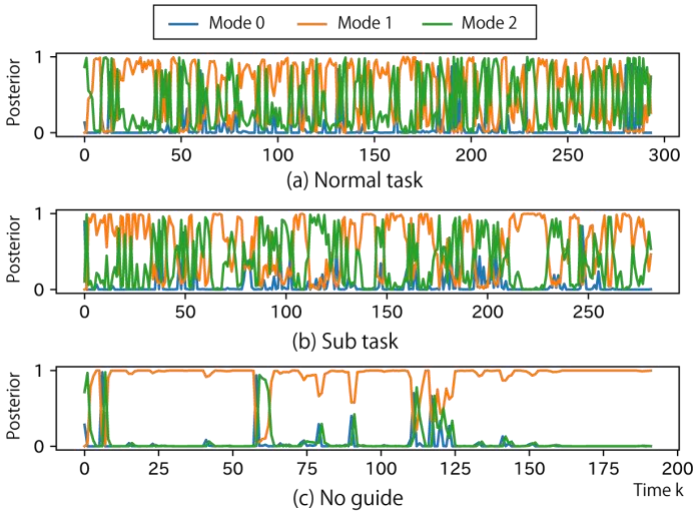

Figure 1. Examples of mode-posterior sequences in the three situations with (a) normal task (Group 1), (b) sub task (Group 2), and (c) no guide (Group 3).

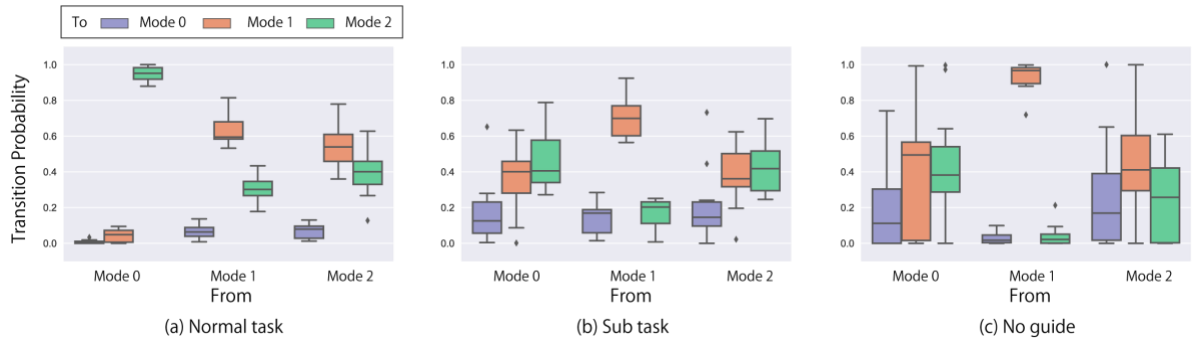

Figure 2. Statistics of mode-transition probabilities estimated through the switching-mode model in the three situations with (a) normal task (Group 1), (b) sub task (Group 2), and (c) no guide (Group 3).

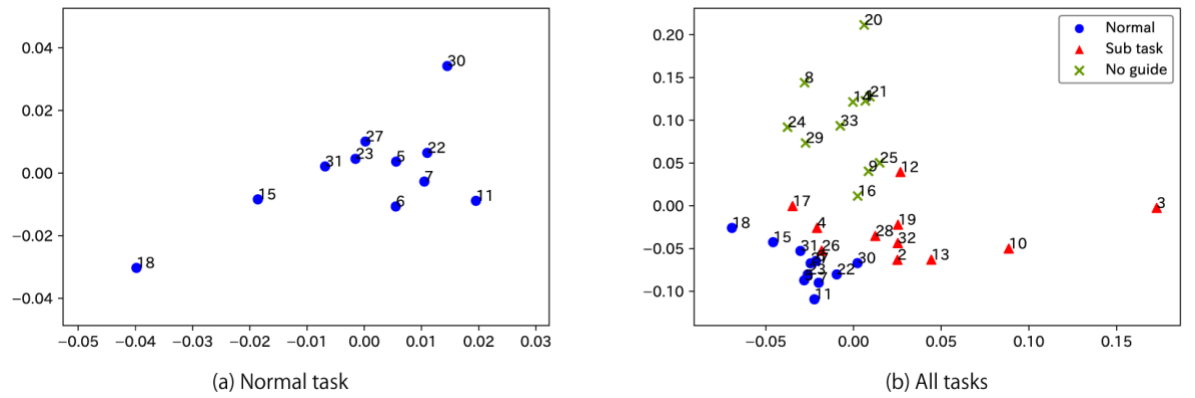

Figure 3. Results of multi-dimensional scaling of dissimilarities of trained models (the numbers correspond to viewers' IDs): (a) dissimilarity structure among normal-task (Group 1) viewers are depicted in 2-d space. (b) dissimilarity structure of all viewers in Group 1-3 are visualized in 2-d space. Colors of marker corresponds to the groups (assigned tasks).

To analyze the difference of viewing styles in Group 1-3, we plot the statistics of parameter distributions of mode-transition probabilities $\{A_{pq}\}$ in Figure 2. In normal-task group (Group 1), there were less transitions to mode 0. In addition, the probability of self-loop of mode 0, which corresponds to the trend of duration length of the mode, also took very small values. This can be interpreted that in normal-task situations, viewers tended to take mode 1 or mode 2 most of the time by actively following

slide content and/or the lecturer's guide compared to sub-task situations. On the other hand, as can be seen in Figure 2 (c), Group 3 (no guide) had high self-loop probability of mode 1, which means they mostly took mode 1 (follows slide content). This is natural since there was no lecturer's guide (e.g., speech).

To demonstrate how the model can be used to discriminate video viewing styles, we visualized the structure of dissimilarities of trained models from each of the viewers. Figure 3 (a) shows the plot of Group 1 structures. Here, multi-dimensional scaling (MDS) was used to visualize the dissimilarity structure in the 2-d Euclidean space. From this figure, we can find which viewer's behavior was different from the others. For example, viewer 18, plotted far from the remaining, had longer duration of mode 0 compared to the others. Figure 3 (b) is also the visualization with MDS but all the viewers' models were used. Here, the difference of the tasks can be clearly observed in this visualization.

5. Conclusion

This paper proposed a method to characterize and visualize video viewing styles during video lectures using a probabilistic mode-switching model. The model consists of three designated submodels to extract and interpret key features in video viewing gaze behaviors, while it has similar structure as HMMs. Through the analysis of mode-transition probabilities and visualization, the proposed model successfully distinguishes the difference of gaze behavior in different tasks. While the assigned tasks were artificially designed in this study, we believe that the proposed technique can be used to find "in which period" each learner has non-typical behavior. Then, this information may help us to design personalized feedback to assist learners by providing information for self-assessment or further study and also to improve learning materials. Another interesting question is how the extracted viewing styles are related to leaners' performance, which should be investigated in the future work with larger data.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP19H04226, JST PRESTO Grant Number JPMJPR14D1, and the Telecommunications Advancement Foundation.

References

Juang, B. H. & Rabiner. L. R. (1985). A Probabilistic Distance Measure for Hidden Markov Models. AT & T Technical Journal, 64 (2), 391–408.

Kawashima, H., Ueki, K., & Shimonishi, K. (2019). Model-Based Analysis of Gaze Data During Video Lectures. 3rd Multimodal Learning Analytics Across Spaces (CrossMMLA) Workshop, the Companion Proceedings of LAK19, 526–533.

Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller., R. C. (2014). Understanding In-Video Dropouts and Interaction Peaks in Online Lecture Videos. International Conference on Learning @ Scale (L@S), 31–40.

Mangaroska, K., Sharma, K., Giannakos, M., Trætteberg, H. & Dillenbourg, P. (2018). Gaze Insights into Debugging Behavior Using Learner-Centred Analysis. ACM International Conference on Learning Analytics & Knowledge (LAK), 350–359.

Mills, C., Bixler, R., Wang, X., & D'Mello, S. K. (2016). Automatic Gaze-Based Detection of Mind Wandering during Narrative Film Comprehension. 9th International Conference on Educational Data Mining, 30–37.

Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E., & D'Mello, S. K. (2017). Gaze-based Detection of Mind Wandering during Lecture Viewing. 10th International Conference on Educational Data Mining, 226–231.

Nguyen, C. & Liu, F. (2016). Gaze-based Notetaking for Learning from Lecture Videos. Paper presented at CHI Conference on Human Factors in Computing Systems, 2093–2097.

Salvucci, D. D. & Goldberg, J. H. (2000). Identifying Fixations and Saccades in Eye-Tracking Protocols. Symposium on Eye Tracking Research & Applications (ETRA), 71–78.

Sharma, K., Jermann, P., & Dillenbourg, P. (2014). "With-me-ness": A Gaze-Measure for Students' Attention in MOOCs. International Conference of the Learning Sciences (ICLS), 1017–1022.