Learning Aspects of Interest from Gaze

Kei Shimonishi, Hiroaki Kawashima, Ryo Yonetani, Erina Ishikaw, Takashi Matsuyama

Abstract

This is a auther version of "Learning Aspects of Interest from Gaze"(ACM International Conference on Multimodal Interaction (ICMI2013) 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction, pp.41-43, Sydney, Australia, 2013.12.13.)

This paper presents a probabilistic framework to model the gaze generative process when a user is browsing a content consisting of multiple regions. The model enables us to learn multiple aspects of interest from gaze data, to represent and estimate user's interest as a mixture of aspects, and to predict gaze behavior in a unified framework. We recorded gaze data of subjects when they were browsing a digital pictorial book, and confirmed the effectiveness of the proposed model in terms of predicting the gaze target.

category H.1.2User/Machine SystemsHuman factors

keywords Probabilistic model; eye movement; recommendation

1 Introduction

Analysis of eye movements has long been studied in the fields of human computer interaction and vision psychology. One of the challenging issues in the fields is the estimation of latent user states including interests [2, 3, 4] and intentions [1, 6] from observed eye movements. The underlying approach of these studies is to extract various gaze features, such as fixation duration and saccade length, and to associate them with discrete user-state labels in a supervised learning fashion. Thus, we need to assume what kinds of states users are likely to become, and often give the labels of states in a top-down manner. However, this assumption is not always appropriate when we apply the estimation techniques to interactive systems. In particular, recommender systems require evaluation scores (the degree of interest) to the items being looked at, as well as those never being looked at (or never displayed in a content). For this case, the degree of

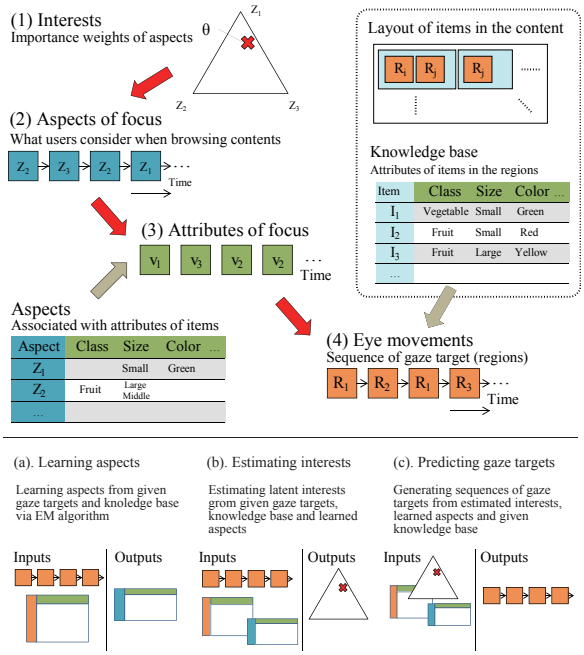

Figure 1: Overview of the proposed model

interest toward a displayed item [2, 4] (e.g., "a *carrot*") is not enough. Instead, the aspects of interest that they place importance (e.g., they are looking for " *healthy*" foods) are more essential in order to estimate the degree of interest toward items never being looked at. Namely, we need to know why (from which viewpoint) these items are looked at rather than which items are of interest.

However, the aspects of user's concern depend on situations (e.g., contents, tasks), and it is often hard to define them in a top-down manner. We therefore try to represent these aspects (viewpoints) indirectly via attributes of items. That is, we associate aspects with attributes, such as specification and appearance (e.g., vegetable, red), and learn this association to find aspects in a data-driven fashion. Besides, user's interest is modeled as a vector ($\boldsymbol{\theta}$ in Fig. 1 (1)) whose components describe the user's importance to the aspects.

The main contribution of this study is to introduce a probabilistic generative model that describes the process of content browsing by modeling user's interests. Namely, we assume that a user focuses on one of the aspects that reflect his/her la-

tent interests ((1) to (2) in Fig. 1), then chooses an attribute value related to the aspect ((2) to (3)), and finally looks at a certain item with the attribute value ((3) to (4)). The proposed model enables us to (a) learn aspects to be considered from gaze data, (b) estimate latent interests from newly observed gaze data using the trained aspects, and (c) predict (recommend) new items that match the estimated interests, in a unified framework.

2 A generative model of gaze

2.1 Content regions and attributes

Suppose that a user is browsing a digital catalog content displaying several items (Fig. 2 shows the environment we used in the experiment). Since an item in the content can be represented by several media, we define *unit regions* R_n $(n = 1, ..., N)$, where each unit region corresponds to the 2-d area on a screen that contains, for example, an image or a text description. As shown in Fig. 2 (bottom), several unit regions constitute an item region, i.e., one item can be presented by several unit regions. We assume that unit regions do not overlap each other.

It is natural to assume that a user browsing a content examines and compares *attributes* of items. Since the attributes are presented through images or described by texts, we associate the attributes with the unit regions. For example, a text-type unit region provides descriptive attributes such as the "category" and "size" of items, and an image-type unit region conveys appearance attributes such as "color" and "texture". In particular, we assume that the relation table between attributes and items (or unit regions) is provided from the knowledge base (see also Fig 1) consisting of a common set of P attributes, where p-th attribute can take Q_p possible values. We therefore denote a set of all the attribute values as $\mathcal{V} := \{V_1, ..., V_Q\}$, where $Q = \sum_p Q_p$ the total number of the attribute values.

User's eye movements are observed as a sequence of gaze points on the screen. Let us use the term "session" to denote one trial of capturing a continuous sequence of gaze. Suppose that, in the learning phase, we have a set of M sessions of gaze data, $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$. Eye movements in m-th session are described as a sequence of regions being looked at, $X_m = (r_{m1}, \ldots, r_{mT_m})$, where $r_{mt} \in \mathcal{R} := \{R_1, \ldots, R_N\}$ and $r_{mt} \neq r_{m(t+1)}$. The basic approach toward user state estimation is to extract features that characterize the eye movements, such as fixation duration [4]. However, in this study, we adopt the frequency distribution indicating how many times each item is looked at, since the comparison of items/regions is particularly important in the situation we consider. Let us denote the frequency distribution of unit regions being looked at by

$$\boldsymbol{g}_m = (g_{m1}, \ldots, g_{mN}),$$

where $g_{mn} \in \mathbb{N}$ denotes the number of times that the user looked at unit region R_n during session S_m.

2.2 A probabilistic model of content browsing

As introduced in Sec. 1, we here assume that one item can be viewed from different *aspects* and that the aspects are associated with the attributes of items. Let us denote a set of K aspects as $\mathcal{Z} := \{Z_1, \ldots, Z_K\}$. We model interests as a K-dimensional parameter vector $\boldsymbol{\theta} \in [0, 1]^K$, where the k-th component, θ_k, is the probability that users choose the aspect Z_k (i.e., $P(Z_k) = \theta_k$ and $\sum_k \theta_k = 1$).

Here, we assume the following content-browsing process, which generates a sequence of regions from user's interest (see also Fig 1). Let $\boldsymbol{\theta}(S_m)$ be the interest the user has in session S_m. For simplicity, we assume $\boldsymbol{\theta}(S_m)$ does not change in the session. At each time t during the session, the user focuses on aspect $z_t \in \mathcal{Z}$, where z_t is determined according to the probability distribution $P(z_t; \boldsymbol{\theta}(S_m))$. Then, attribute value $v_t \in \mathcal{V}$ related to z_t is focused on. We assume that this process is determined by $P(v_t|z_t)$, whose parameters are given by the conditional probability table $\mathcal{P}_{VZ} := \{P(V_q \mid Z_k)\}$. Finally, region $r_t \in \mathcal{R}$ is looked at, where r_t is chosen from a set of unit regions with the attributes value v_t. The last process is determined by the parameters, $\mathcal{P}_{RV} := \{P(R_n \mid V_q)\}$, of $P(r_t|v_t)$.

Hence, the probability that the user looks at region R_n at time t becomes

$$P(r_t = R_n|S_m) = \sum_{V_q \in \mathcal{V}} \sum_{Z_k \in \mathcal{Z}} P(r_t = R_n, v_t = V_q, z_t = Z_k|S_m), \tag{2.1}$$

where the joint probability can be calculated by using

$$P(r_t, v_t, z_t|S_m) = P(r_t|v_t)P(v_t|z_t)P(z_t|S_m).$$

Finally, the probability of X_m, the sequence of regions being looked at, is calculated as follows:

$$P(X_m|S_m) = \prod_{t=1}^{T_m} P(r_{mt} \mid S_m) = \prod_{R_n} P(R_n \mid S_m)^{g_{mn}}. \tag{2.2}$$

As for learning, the parameters $\mathcal{P}_{VZ}$ and $\boldsymbol{\theta}(S_m)$ can be estimated via the expectation-maximization (EM) algorithm given a set of gaze data, $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_M\}$, and $\mathcal{P}_{RV}$.

Note that user's interest $\hat{\boldsymbol{\theta}}$ in a new session can be estimated from corresponding gaze data once parameter $\mathcal{P}_{VZ}$ is learned. This is a key to predict which items the user is interested in (i.e., which items will be looked at next). That is, given estimated $\hat{\boldsymbol{\theta}}$ in the new session, the prediction can be done by calculating the distribution of regions by Eq. (2.1) using given $\mathcal{P}_{RV}$ and learned $\mathcal{P}_{VZ}$ with $\hat{\boldsymbol{\theta}}$.

2.3 Remarks on the proposed model

The proposed model can be seen as the extension of the probabilistic latent semantic analysis (pLSA), used also in recommender systems [5]. However, the key of the proposed framework is that we can introduce a variety of gaze-related structures into the model through the design parameter $\mathcal{P}_{RV}$: the probability distribution that the regions are looked at given a focused attribute.

As mentioned in Sec. 2.1, the information of knowledge base provides an association between items (or unit regions) and attributes, and therefore serves as the basis of designing $\mathcal{P}_{RV}$. Indeed, as a simple implementation, we define $P(R_n \mid V_q)$ by the inverse of the number of regions which have attribute V_q in the content; that is, every region with V_q has an equal probability, and the other regions without V_q have zero probability.

In addition, albeit beyond the scope of this paper, one can extend the model further by taking into account the effect of spatial layout design (e.g., specific position attracts gaze more) and the gaze dynamics (e.g., modeling of temporal gaze patterns using $P(r_t|r_{t-1}, v_t)$ instead of $P(r_t|v_t)$).

3 Experiments

In order to evaluate the proposed model, we first recorded gaze data of subjects browsing a displayed content. Then, we trained the proposed model using the captured data, and evaluated the accuracy of predicting gaze targets from estimated interests.

A fish pictorial book with 15 tiled items (fish types) was prepared for a content. Each item region consisted of a pair of a text description and an image of a fish. The content was displayed on a PC monitor (Fig. 2). Each text region had the description of fish: biological categories, habitats, and sizes; on the other hand, each image presented the entire body of a fish. For appearance attributes, we extracted the histograms of hue and saturation from the images. In addition, we used the bag of features (BoFs) of the fish (foreground) as a feature of textures. Each of the features were clustered into six discrete values via K-means.

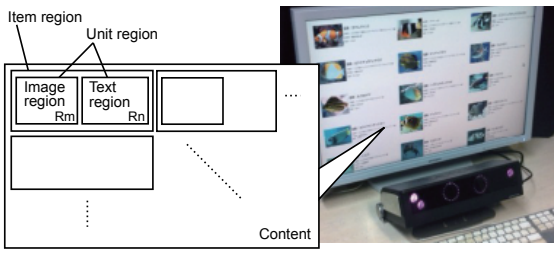

Figure 2: Experimental environment.

Six subjects took part in the experiments. They were asked to sit in front of the monitor, and their gaze data were captured by the Tobii X120 Eye Tracker (sampling rate: 120Hz) installed below the screen. Each subject was first asked to browse the displayed content and to know what items were displayed on the screen. This phase was prepared to separate the two different stages: watching new items and comparing the items. Then, in each session, the subjects conducted one of three tasks: The subject was asked to choose an item (fish) he/she wanted to eat (task 1), wanted to have for a pet (task 2), or wanted to know well (task 3). These tasks were designed so that the subjects could browse the content from different viewpoints (various aspects). Note that the number of aspects used in the model learning was not identical to that of tasks.

3.1 Evaluation and results

As introduced in Sec. 1, our motivation is to predict items a user will be interested in by observing the user's current gaze behavior. We therefore evaluated the method in terms of the prediction accuracy for the gaze behavior in the latter part of the session when the first half (or more) was observed. Specifically, in each session, we estimated the latent interest of the user by using the first x [%] of the session and estimated the probability distribution for the regions in the content by Eq.(2.1). Here, this distribution serves as the "prediction" of the remaining $100 - x$ [%] of the session. We then calculated the likelihood score of the remaining data with this estimated distribution by Eq.(2.2). The likelihood score in each session was normalized by the length of the session data, and the final score was obtained by averaging the results of all the sessions. Before this evaluation, we trained $\mathcal{P}_{VZ}$ from all the session data, $\mathcal{S}$, using the EM algorithm. The number of the aspects was chosen to $K = 10$ empirically (arround the half of the number of ses-

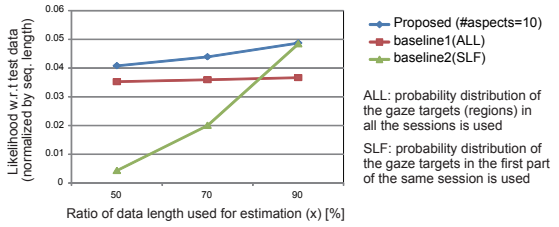

Figure 3: Prediction of the last $100 - x$ [%] of each session. The first x [%] period is used to estimate the distribution of regions in the remaining period.

sions, $M = 18$).

Figure 3 shows the result of the likelihood score with $x = 50, 70, 90$ [%]. As for the baselines, we used two types of probability distributions for the prediction. In baseline 1 the distribution of gaze targets (regions) in all the training data was used, while in baseline 2 the distribution of gaze targets in the first x [%] of the same session was used. Since the proposed model predicts gaze behavior through the estimated interest, we observe that the gaze regions in the latter period including those never looked at in the first part were successfully predicted, while the baseline 2 fails when x was small. Although the experiment and the model are still preliminary, this is a key feature for recommender systems that can find user's preferred items by interactively presenting new items.

4 Conclusion

This paper proposed a probabilistic generative model of gaze behavior, which learns and estimates user's interests from observed gaze data. For future work, we are extending the model to incorporate spatio-temporal structures such as the dynamics of gaze and interests, and also evaluating the method with a large amount of data to verify the effectiveness for recommender systems.

ACKNOWLEDGMENTS. This work was supported in part by SCAT Technology Research Foundation, Japan.

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] B. Brandherm, H. Prendinger, and M. Ishizuka. Interest estimation based on dynamic bayesian networks for visual attentive presentation agents. In *ICMI*, pages 346–349, 2007.

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

[4] 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.

[5] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, (1):89–115, 2004.

[6] 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.