Modeling Semantic Aspects of Gaze Behavior while Catalog Browsing
(Authors Version)

Erina Ishikawa
Kyoto University
Kyoto, Japan
ishikawa@vision.kuee.kyoto-u.ac.jp

ABSTRACT

ACM, 2013. This is the author's 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 proceedings of the 15th ACM on International Conference on Multimodal Interaction 2013 Doctoral Consortium (ICMI 2013 DC).

Gaze behavior is one of crucial clues to understand human mental states. The goal of this study is to build a probabilistic model that represents relationships between users' gaze behavior and user states while catalog browsing. In the proposed model, users' gaze behavior is interpreted based on semantic and spatial relationships among objects constituting displayed contents, which is referred to as *designed structures*. Moreover, a method for estimating users' mental states is also proposed based on the model to evaluate the model by measuring the performance of user state estimation. The results from preliminary experiments show that the proposed model improved estimation accuracies of user states compared to other baseline methods.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

General Terms

Algorithms, Experimentation

Keywords

eye movements, user states, topic modeling

1. INTRODUCTION

Eye movements play an important role to understand human mind such as interests and intentions. Analyses of eye movements have long been conducted in the fields of visual psychology [10] and human computer interaction [6].

ICMI '13, Dec 09-13 2013, Sydney, NSW, Australia
ACM 978-1-4503-2129-7/13/12.
http://dx.doi.org/10.1145/2522848.2532197.

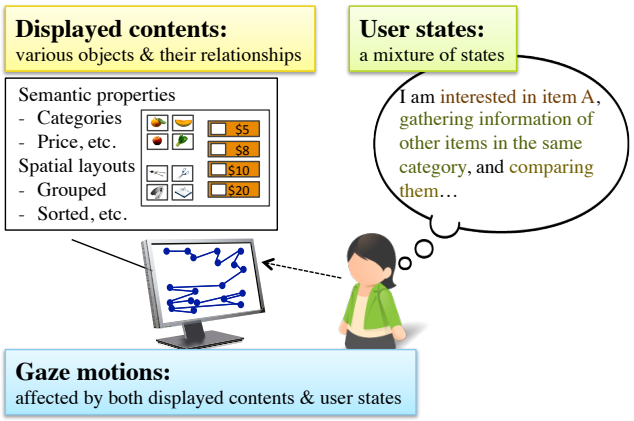

Figure 1: Gaze behavior in real environments

More recently, beyond mere analyses of observed gaze data, several models have been proposed to describe probabilistic relationships between gaze behavior and user states [1, 2]. Mathematical modeling of the gaze motions is of great interest in various applications, such as designing interfaces, evaluating usability, and estimating user states.

Previous studies have mainly focused on controlled situations, where displayed contents have only a few objects and user states can be classified into one of several predefined states. However, understanding gaze motions in uncontrolled environments is still challenging because the displayed contents may contain various objects with spatial and semantic relationships with each other. Moreover, it may not be possible to describe a user's mental state with a single classification label, as seen in Fig. 1. For the former problem, several studies aim to introduce content information into gaze behavior analyses. Yonetani et al., for instance, proposed an estimation method of viewer-states from the viewer's gaze motions while general video browsing, such as TV commercials [11]. They focused on spatial and temporal relationships between gaze dynamics and saliency dynamics of videos. However, for understanding gaze behavior while static content browsing, such as product catalogs, semantic aspects of displayed contents gain importance. In consequence, utilizing semantics of displayed contents becomes important for gaze behavior analysis.

The goal of this study is to build a probabilistic model of real-world gaze behavior while catalog browsing by employing underlying semantics of displayed contents. The main

contribution of this study is to introduce semantic and spatial relationships among objects in displayed contents for interpreting gaze behavior. Interpreted gaze behavior can describe more detailed browsing behavior of users, which leads to further understanding of user states. An estimation method of user states is also established based on the proposed model in order to evaluate the model by measuring the performance of user state estimation.

2. PROBLEMS

Suppose a user is browsing a digital catalog to select a gift for his/her friend. Items in the catalog contain various media such as images and text. The user's eye movements are observed as a sequence of gaze points on the screen by an eye tracker placed below the screen. In this situation, this study aims to understand semantic aspects of gaze behavior while catalog browsing such as "inspecting item details" or "comparing several items". Specifically, this study addresses the following three problems.

(A) Modeling contents.

Semantic information of displayed content is an essential clue to understand the meanings of gaze motions. For instance, the switch of gaze targets among different media describing the same item can mean "inspection of the item", while the switch of gaze targets between different items can mean "comparison of the items". To employ such semantics, previous studies mainly consider semantic properties of objects and associate them with object regions [7, 3]. The annotation requires to define both object regions and labels of the properties manually. However, as displayed contents become complex and diverse, it becomes more difficult to determine appropriate labels, e.g., users can regard several object regions as a single meta-object regions, and semantic properties associated with the meta-objects can be different from those of component-objects. Thus, the following problem should be considered: *how can semantic information of contents be represented for understanding gaze motions?*

(B) Modeling gaze motions.

When we analyze gaze behavior while catalog browsing, it is important to consider temporal information of gaze motions, such as the order of objects being looked at. For instance, fixation sequence of the form X-Y-X ... can be used to identify pair comparisons while subjects are making a choice [8]. To analyze temporal information of a user's gaze, previous studies apply N-gram analysis to a sequence of gaze targets [7], however, acquiring a sequence of gaze targets eventually discards smaller gaze motions in each object region. Small motions are also important to understand browsing behaviors; for example, they indicate how users examine items. To leverage both small and large gaze motions, gaze data should be analyzed in multiple scales. Therefore, this study addresses the following problem: *how can gaze motions be represented employing both temporal information and multiple scale information?*

(C) Modeling user states.

To achieve the modeling of relationships between gaze motions and user states, we need to consider how to represent user states. Previous studies mainly assume user states as one of several discrete states defined preliminarily, such as "being interested in item A". However, user states in uncontrolled conditions can be more complex, such as "being interested in item A and B, and comparing them". Since possible user states depend on situations (e.g., contents, tasks), it is difficult to define them in a top-down manner. The problem here is: *how can such various user states be dealt with?*

3. APPROACH

The overview of the proposed model of gaze behavior is shown in Fig. 2. The model assumes a user state as a mixture of states (Sec. 3.3), where each state is associated with a frequency distribution of specific gaze motion patterns. Then, the gaze-pattern distributions are generated based on the user state (Sec. 3.2). The gaze patterns are described based on *designed structures*, which represent spatial and semantic relationships among objects (Sec. 3.1).

3.1 Designed structures of displayed contents

Items constituting digital catalog usually have various semantic properties. As the number of the items increases, the semantic properties can be more and more diverse, which causes difficulty when defining and annotating appropriate semantic labels to object regions. To overcome this problem, instead of using the semantic properties themselves, the model exploits the semantic relationships among objects, such as "describing the same item" and "describing items in different categories". The relationships gain importance when we aim to understand user states while browsing digital catalog, e.g., "examining a item" and "comparing items".

To take such relationships into account, I propose structures that describe semantic relationships of objects and their spatial layout jointly. Particularly, the semantic aspects in the 'designed structures' are modeled by a directed graph (*content graph*), where the nodes and edges of the graph describe various entities and their semantic relations, respectively (see Fig. 2 (A)). For instance, consider a catalog content consisting of some items which can be classified into several categories. In this case, a category (e.g. *Fruit*) is a property of an item (e.g. *Apple*), and the edge from the item node to its category node describes an *is-a* semantic relation. Additionally, items in the digital catalog are usually described by various media, such as images and text. The media occupy spatial regions on the screen, which I refer to as *unit regions*. Each unit region node is connected to its item node by the *describe* semantic relation. The unit regions are used to associate gaze points with the semantic properties in the following section.

3.2 Gaze motion patterns

When we analyze catalog browsing behavior, important eye movements are saccades and fixation. The saccades are a especially crucial clue to understanding browsing behavior since they indicate attentional shifts from one object to another. Moreover, there are various lengths of saccade strokes, and they describe viewing strategies of users. For instance, when a user is obtaining the detailed information of a particular item, small saccades inside the item region can be observed, while gaze positions jump across the wide area of a screen when the user is acquiring the outline of contents. Thus, in the proposed model, the saccades are first detected in multiple scales for representation of gaze motions. Specifically, a sequence of gaze-motion speeds is

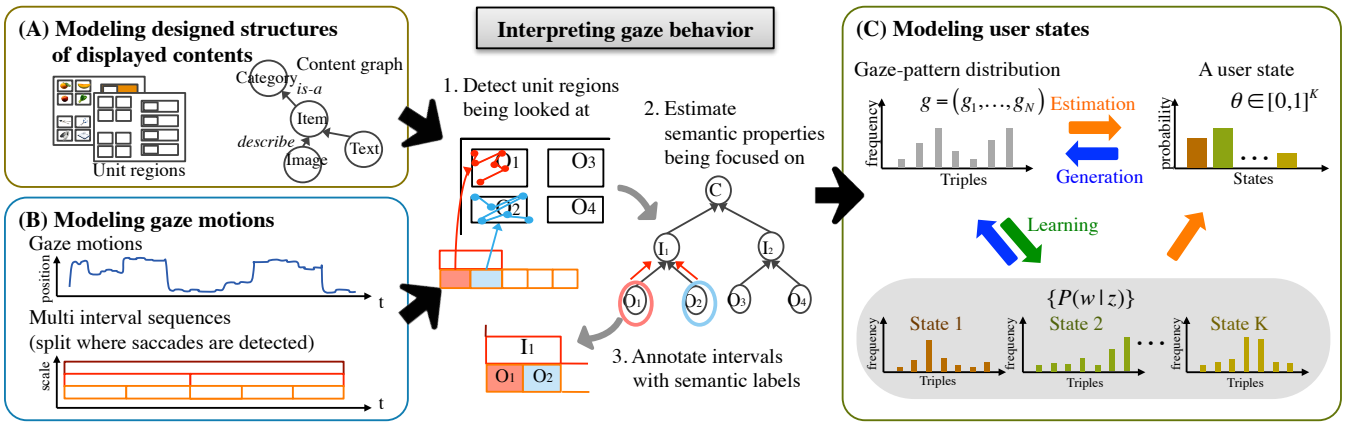

Figure 2: Overview of the proposed model

obtained from observed gaze points, and then a scale-space filter [9] is applied to the sequence. After the saccade detection, the hierarchical structure of interval sequences is obtained automatically (see Fig. 2 (B)).

Gaze motions are interpreted by associating the intervals with the semantic relations by referring to semantic properties of the content graph. For the designed structures, this model assumes unit regions to be placed close to each other if they share some semantic relations, i.e., the spatial layouts of the unit regions are assumed to correspond to the semantic relations. Under the assumptions, gaze motions from one unit region to another, which are observed as a gaze-point sequence on a screen, can be associated with the semantic relations among unit regions. The procedures for interpreting gaze behavior is depicted in Fig. 2 (middle row). The first step is to identify the unit regions being looked at for each interval in the finest scale. They can be easily identified by referring to the gaze points in the intervals. Then, the intervals in coarser scales contain a set of the unit regions. Since unit-region nodes are linked to the nodes with higher semantic properties in the content graph (e.g., an apple image is linked to *Apple* and *Fruit*), the semantic properties in the coarser-scale intervals are finally annotated as the one corresponding to the node which links to all the regions with the shortest paths.

As a result of the interpretation procedures, interval sequences with multiple scales, G, will be obtained. The elements (intervals) of G contain node labels of the content graph, i.e., unit regions or the higher semantic properties. Here, we can consider two analyses of the multi-scale interval sequences. The analysis along the time axis captures the temporal changes of the node labels being focused on, while the analysis along the scales captures the relations of the node labels in different scales. In [5], to take both time and scale into account, the model simply extracts the triplet patterns that consist of two temporally successive intervals and the interval in the next-coarse scale that contains the two intervals. The triplets describe both temporal changes of semantic properties and their hierarchical relationships. Here, let us denote a set of N possible triplets as $\mathcal{W}=\{w_1,\ldots,w_N\}$. All triplets are extracted from the multi-scale interval sequences, G, and the frequency of w_n, g_n, is used as a feature vector $\boldsymbol{g}=(g_1,\ldots,g_N)$ to describe semantic aspects of gaze motions.

3.3 User states while catalog browsing

User states while catalog browsing can be complex and diverse, such as "being interested in item A and B, and comparing them". Existing studies have mainly dealt with behaviors under controlled conditions, and assumed user states as a discrete single state. This study aims to address such complex user states by assuming them as a mixture of states, i.e., the example above can be considered as a mixture of "being interested in item A", "being interested in item B" and "comparing several items". Since it is obviously difficult to define these base states manually, the model will discover them from observed gaze datasets in a bottom-up manner. Specifically, I will introduce a probabilistic topic model [4], which is inspired by data mining studies, in order to represent user states as a mixture of latent aspects learned from gaze data (see Fig. 2 (C)).

Let us denote a set of K states as $\mathcal{Z}=\{Z,\ldots,Z_K\}$. User states are represented as a K-dimensional parameter vector $\boldsymbol{\theta}\in[0,1]^K$, where the k-th element of $\boldsymbol{\theta}$, θ_k, is the probability that the user takes state Z_k that is, $P(Z_k)=\theta_k$ and $\Sigma_k\theta_k=1$. The proposed model particularly focuses on the frequency of the triplets, $\boldsymbol{g}$, and formalize the generative process of G by ignoring durations of intervals as follows:

$$P(G)=\prod_{n=1}^{N}\left\{\sum_{Z_k}P(Z_k|\boldsymbol{\theta})P(w_n|Z_k)\right\}^{g_n}. \qquad (1)$$

A set of conditional probabilities of gaze pattern w with state z, $\{P(w|z)\}$, can be learned from gaze data by using the Expectation Maximization (EM) algorithm as shown in [4]. The model will be capable of estimating the user state $\boldsymbol{\theta}$ given E with learned parameters.

4. PRELIMINARY EXPERIMENTS

This section describes the experimental methodology, the preliminary results from [5] and future experimental plans. I will evaluate our model in a step-wise manner by preparing several experimental conditions. In [5], I focus on product categories as semantic properties of items, and assume user states as pre-defined discrete states.

Experimental setup.

Eight subjects took part in the experiments. 10 digital

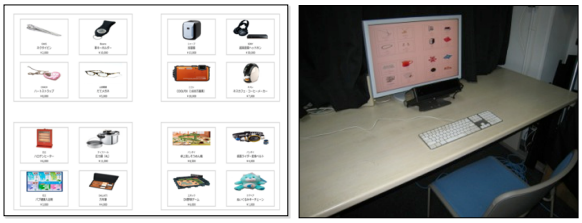

Figure 3: Experimental setup. Left: an example of displayed contents. Right: apparatuses.

catalogs were prepared to be displayed on a screen. Each content contained the information (captions and images) of 16 products, which can be grouped into one of four categories: accessories, home electronics, house-hold goods, and toys. Contents were displayed on a screen[1], and a subject was asked to sit in front of the screen (see Fig. 3). Gaze data were acquired as 2-d points on the screen by using an eye tracker[2] installed below the screen.

In the previous experiment, I aimed to discriminate user's three states, i.e., *input*, *decision*, and *free-viewing* states. Therefore, the following series of tasks was given to each subject in the expectation that the subject's internal state transit through the three states.

Task 1: input (30sec) *Browse the catalog displayed on the screen, and confirm what items are there.*

Task 2: decision (no limit) *Select a gift from the catalog for your friend.*

Task 3: free-viewing (60sec) *Watch the catalog freely.*

Preliminary results and discussions.

In this experiments, I assumed user states as one of three states. A linear classifier was employed to estimate the user states by using the frequency distributions as a feature vector, and then estimation accuracies were obtained via leave-one-out cross validation. The following two methods (B1) and (B2) were used to serve as baselines. The baseline (B1) applies a bi-gram analysis to the sequences of the products being looked at, which corresponds to the traditional temporal analysis introduced in Section 2 (B). The other baseline (B2) first extracts saccade speed information with multi-scales, and then use it as a feature. Note that (B2) does not employ any semantic information from displayed contents.

The results are shown in Tab. 1. The comparison among the proposed method and the other baselines demonstrates the effectiveness of using the semantic information of contents and analyzing the semantic relations being focused on with multiple scales. In addition, I examined estimation accuracies when classifying every two-state pair to investigate separabilities between the states. The accuracies were 66.2% (input vs decision), 73.6% (decision vs free-viewing) and 77.8% (free-viewing vs input). The results show that the gaze behavior in input states tend to be similar to ones in decision states rather than ones in free-viewing states. This can be interpreted as meaning that subjects are motivated to acquire content information more actively in decision states than in free viewing states after making a choice.

[1]MITSUBISHI RDT262WH (550.1x343.8mm)

[2]Tobii X60 (freedom of head movement: 400x220x300mm, sampling rate: 60Hz, accuracy: 0.5degrees).

Table 1: Estimation accuracies [%]. (B1): bi-gram analysis and (B2): saccade analysis.

Proposed method	(B1)	(B2)
59.7	42.9	48.8

Through experiments, I found that considering categories improves estimation accuracies of user states: acquiring information, comparing and free-viewing. However, there could be other semantic relations that are important to understand eye movements, e.g., order relations of items such as ranking. I am extending the representation of designed structures by involving various semantic relations. Moreover, I will build huge gaze datasets for statistical learning of the model and release them to the public, which can be an additional contribution of my study.

5. CONCLUSIONS

The goal of my study is to build a probabilistic model of gaze behavior in uncontrolled environments. The model employs designed structures of displayed contents to interpret the meanings of gaze behavior. Moreover, by representing the user sate as a mixture of states, the model is expected to achieve a more detailed understanding of catalog browsing behaviors. For future work, I am extending content model as well as designing experiments to collect further data of gaze for statistical learning of the proposed model.

6. ACKNOWLEDGMENTS

This work is supported by Grant-in-Aid for Scientific Research under the contract of 25·5396.

7. 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] S. Eivazi and R. Bednarik. Predicting Problem-Solving Behavior and Performance Levels from Visual Attention Data. In *IUI*, pages 9–16, 2011.

[3] S. Eivazi, R. Bednarik, M. Tukiainen, M. von und zu Fraunberg, V. Leinonen, and J. E. Jääskeläinen. Gaze Behaviour of Expert and Novice Microneurosurgeons Differs During Observations of Tumor Removal Recordings. In *ETRA*, pages 377–380, 2012.

[4] T. Hofmann. Probabilistic Latent Semantic Analysis. In *UAI*, pages 289–296, 1999.

[5] E. Ishikawa, R. Yonetani, H. Kawashima, T. Hirayama, and T. Matsuyama. Semantic Interpretation of Eye Movements using Designed Structures of Displayed Contents. In *Gaze-In*, 2012.

[6] R. J. K. Jacob and K. S. Karn. *The mind's eye: cognitive and applied aspects of eye movement research*. 2003.

[7] Y. Nakano and R. Ishii. Estimating User's Engagement from Eye-gaze Behaviors in Human-Agent Conversations. In *IUI*, pages 139–148, 2010.

[8] J. Russo and L. Rosen. An eye fixation analysis of multialternative choice. *Memomry & Cognition*, 3(3):267–276, 1975.

[9] A. P. Witkin. Scale Space Filtering. pages 1019–1022, 1983.

[10] A. Yarbus. Eye movements and vision. *Plenum*, 1967.

[11] R. Yonetani, H. Kawashima, and T. Matsuyama. Multi-mode Saliency Dynamics Model for Analyzing Gaze and Attention. In *ETRA*, pages 115–122, 2012.