Semantic Interpretation of Eye Movements Using Designed Structures of Displayed Contents (Authors Version)

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

Ryo Yonetani
Kyoto University
yonetani@vision.kuee.kyoto-u.ac.jp

Hiroaki Kawashima
Kyoto University
kawashima@i.kyoto-u.ac.jp

Takatsugu Hirayama
Nagoya University
hirayama@is.nagoya-u.ac.jp

Takashi Matsuyama
Kyoto University
tm@i.kyoto-u.ac.jp

ABSTRACT

This paper presents a novel framework to interpret eye movements using semantic relations and spatial layouts of displayed contents, i.e., the designed structure. We represent eye movements in a multi-scale, interval-based manner and associate them with various semantic relations derived from the designed structure. In preliminary experiments, we apply the proposed framework to the eye movements when browsing catalog contents, and confirm the effectiveness of the framework via user-state estimation.

Categories and Subject Descriptors

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

General Terms

Algorithm, Experimentation

Keywords

eye movements, semantic network, user states

1. INTRODUCTION

Gaze-based intelligent user interfaces (IUIs) employ user eye movements not only as a cursor to indicate his/her explicit intentions but as a crucial cue reflecting implicit states, such as "comparing several items" and "trying to decide items to buy" in catalog browsing. To achieve the latter states, it is important to analyze the semantics underlying both eye movements and displayed contents. Our goal is to analyze those semantics in detail enough to understand users' implicit states while they are browsing catalog contents.

Analyzing a relationship between user states and his/her eye movements is a longstanding topic in visual psychology with the pioneering work performed by Yarbus [8], and it is now taken over by the estimation techniques of user states from eye movements such as [1, 3, 9]. In particular, several gaze-based IUIs have tried to analyze eye movements to understand user internal states in order to provide meaningful information [2, 6]. On the other hand, a high-level (i.e., semantic) analysis of eye movements have recently had a lot of attention in the field of eye-tracking study. For this analysis, semantic properties of objects displayed in contents are often considered [4, 5]. However, when understanding the user states such as comparing items and selecting one from them, not only semantic properties but their relations, e.g., "some of the items are in the same category" and "items are sorted in some semantic order", are required to be considered.

In this paper, we propose a novel framework that interprets eye movements as a focus of the semantic relations. The main contribution of this work is to actively utilize various semantic relations underlying the displayed contents for the interpretation, which differs from the previous work annotating eye movements by the objects or their semantic property being looked at. To handle the contents including various semantic relations, we use a combined representation of semantic networks and content spatial layouts, which we refer to as a *designed structure*.

For the designed structures, objects are assumed to be placed close to each other if they share some semantic relations. In other words, the spatial layouts of the objects are assumed to correspond to the semantic relations. Under these assumptions, eye movements from one object to another, which are observed as a gaze-point sequence on a screen, can be associated with the semantic relations among objects. Moreover, the user states mentioned above can be estimated by statistically learning the semantic relations being focused on.

The following section introduces the concept of our proposed framework. In Section 3, we report preliminary experiments evaluating the effectiveness of the proposed framework via

Gaze-In'12, October 26, 2012, Santa Monica, California, USA.

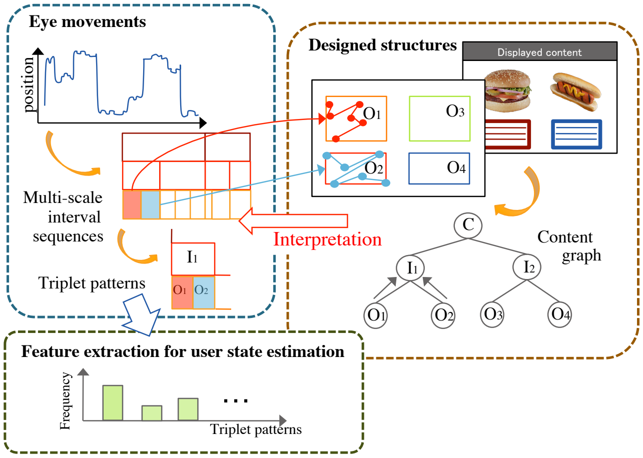

Figure 1: Overview of our proposed framework.

estimation of user states while browsing catalog contents.

2. PROPOSED FRAMEWORK

Suppose that catalog contents, such as advertisements, are displayed on a screen and a user is browsing it. The eye movements of the user are observed as a sequence of gaze points on the screen by using an eye tracker. The contents have several objects such as images, captions, etc. The contents also have a specific layout designed by a content creator, which reflects some semantic relations.

Our approach to understand eye movements and user states (e.g., acquiring information, comparing, free-viewing) is to utilize the designed structures for associating the eye movements with the semantic relations. For this, the switches of gaze targets with multiple scales are first extracted from eye movements. We associate them with the semantic relations by referring to semantic properties of the gaze targets using the designed structures.

2.1 Designed Structure

Semantic relations of objects play an important role to interpret eye movements. For instance, suppose that a user first stares at object O_1, switches the gaze target to object O_2, and gazes back to object O_1 (i.e., the sequence of gaze targets becomes $O_1O_2O_1$). Here the semantic relations between the objects O_1 and O_2 imply the meaning of the gaze pattern $O_1O_2O_1$. If both objects O_1 and O_2 describe the same item I, the meaning of the gaze pattern can be interpreted as an exploration of I. Meanwhile, if O_1 and O_2 describe different items, I_1 and I_2, respectively, it is a comparison of I_1 and I_2.

To handle such semantic relations, we introduce the designed structures, which describe semantic relations of objects together with their spatial layouts (Fig. 1 right). The semantic aspects in the designed structures are here modeled by a directed graph (*content graph*), where the nodes and edges of the graph describe various entities and their semantic relations, respectively.

Consider the catalog contents consisting of commercial products which can be classified into several product categories (Fig. 2). The product and its category are, for instance, *Apple* and *Fruit*. In this case, the *Fruit* is a property of the *Apple*, and the edge from the *Apple* node to the *Fruit* node describes an *is-a* semantic relation. In addition, the commercial product in the catalog contents can be described by a form of images, captions, etc. For this, the product is a property of the descriptions, and there is a *describe* edge from the descriptions to the product.

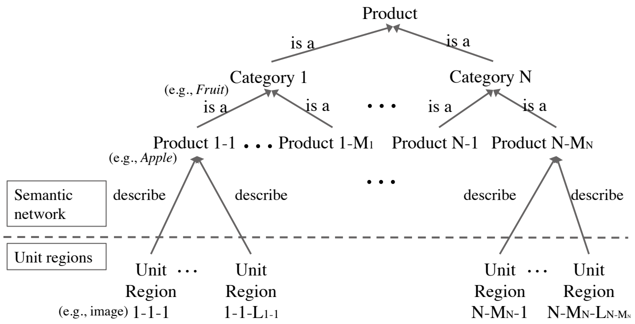

Figure 2: The content graph of a catalog.

Besides the semantic aspects modeled by the content graph, the designed structures have spatial layout information of a content as well. Specifically, descriptions such as images and captions occupy some spatial regions (unit regions). We utilize these regions to associate gaze points with the semantic properties in the following section.

2.2 Semantic Interpretation of Eye Movements

Human eye movements contain saccades, which indicate attentional shifts from object to object. To achieve semantic interpretation of eye movements, intervals between saccades are required to be annotated by semantic properties, which are derived from the designed structures. On the other hand, there are various lengths of saccade strokes, and they describe viewing strategies of users. For example, when a user tries to obtain the detailed information of a particular object, the gaze points inside the target region can be observed. Meanwhile, if the user is acquiring the outline of contents, gaze positions jump across the wide area of a screen.

To take such aspects into account, we introduce multi-scale saccade detection. Specifically, we obtain a sequence of eye-motion speeds from gaze points and apply a scale-space filter [7] to the sequence to detect saccades in multiple scales.

After the saccade detection, the hierarchical structure of interval sequences is obtained (top-left of Fig. 1). The first step to interpret eye movements is to identify the unit regions being looked at for each of the intervals 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 higher-semantic-property nodes 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.

2.3 User State Estimation

As a result of the semantic interpretation of eye movements, we obtain interval sequences with multiple scales whose elements (intervals) contain node labels of the content graph, i.e., unit regions or the higher semantic properties. The basic idea of estimating user states is that we statistically analyze the interval sequences which represent typical semantic relations observed in the multi-scale sequences.

Let us first consider the two analyses of the multi-scale interval sequences. The analysis of the sequences along time captures the temporal changes of the node labels being focused on. Meanwhile, the analysis along the scales captures the relations of the node labels in different scales. Since each node label is a unit region or a higher semantic property, the temporal analysis extracts how the properties of interest change over time, and the scale analysis extracts what properties are looked at depending on the area of interest. Previous papers such as [5] have mainly dealt with the former temporal analysis. On the other hand, the latter analysis shows the hierarchical relations of semantic properties, where the large strokes of saccades approximately correspond to the switches of the higher levels of semantic properties.

To employ both analyses simultaneously, we here extract the triplet patterns from the multi-scale interval sequences, which consist of two temporally successive intervals and the interval in the next-coarse scale which contains the two intervals (top-left of Fig. 1). Notice that the patterns describe the semantic relations being focused on. We extract all the triplets from the multi-scale interval sequences and use the frequency of the triplet patterns as a feature vector to estimate user states (bottom-left of Fig. 1).

3. EXPERIMENTS

Experimental setup. Eight subjects (two females and six males) took part in this experiment. Each subject was asked to sit in front of a screen [1]. An eye tracker [2] was installed below the screen, and eye movement data were acquired as 2-d points on the screen. The distance between the subject and the screen was around 1000mm. Each displayed content included 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.

In the experiments, we aim to discriminate user's three states, i.e., *input*, *decision*, and *free-viewing* states. In the experiments, we therefore gave each subject the following series of tasks in the expectation that the subject's internal state transit through the three states.

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

Task 2:decision (no limit) *Select a gift from the catalog for one of your close acquaintances. Press the keyboard button when you have decided what to select.*

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

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

Proposed method	(B1)	(B2)
59.7	42.9	48.8

Results and discussions. A linear classifier was employed to estimate the user states, and estimation accuracies were obtained via leave-one-out cross validation. The following two methods (B1) and (B2) were used to serve as baselines. (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 2.3. The other baseline (B2) extracts saccade speed information with multi-scales and utilize it as a feature. That is, (B2) does not employ any semantic information from displayed contents. The results are shown in Table 1. The comparison between the proposed method and the other two baselines demonstrates the effectiveness of utilizing the semantic information of contents and analyzing the semantic relations being focused on with multiple scales.

Additionally, estimation accuracies when classifying every two-state pair were examined to clarify the separability between the states. The accuracies were 66.2% (input vs decision), 73.6% (decision vs free-viewing) and 77.8% (free-viewing vs input). These results suggest that subjects tend to acquire content information more actively in both input and decision states than in free-viewing states after decision.

4. CONCLUSIONS

This paper presented a framework to interpret eye movements using the designed structure of displayed contents. By taking into account semantic relations and spatial layouts of displayed contents, it enables us to understand the meaning of observed eye movements in detail enough to estimate user states. For future work, we are extending the representation of designed structures by involving a variety of semantic relations (e.g., order relations of objects such as ranking), and we are also applying the framework to achieve the dynamic control of displayed contents.

5. 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] 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.

[5] Y. Nakano and R. Ishii. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *IUI*, pages 139–148, 2010.

[1] MITSUBISHI RDT262WH (550.1×343.8mm).

[2] Tobii X60 (freedom of head movement: 400×220×300mm, sampling rate: 60Hz, accuracy: 0.5degrees).

[6] P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In *CHI*, pages 221–230, 2005.

[7] A. Witkin. Scale-space filtering. In *IJCAI*, pages 1019–1022, 1983.

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

[9] R. Yonetani, H. Kawashima, and T. Matsuyama. Multi-mode saliency dynamics model for analyzing gaze and attention. In *ETRA*, pages 115–122, 2012.