IEICE TRANS. INF. & SYST., VOL.E98–D, NO.8 AUGUST 2015

PAPER

Using Designed Structure of Visual Content to Understand Content-Browsing Behavior

Erina ISHIKAWA[†a)], ***Student Member***, **Hiroaki KAWASHIMA**[†], ***and*** **Takashi MATSUYAMA**[†], ***Members***

SUMMARY Studies on gaze analysis have revealed some of the relationships between viewers' gaze and their internal states (e.g., interests and intentions). However, understanding content browsing behavior in uncontrolled environments is still challenging because human gaze can be very complex; it is affected not only by viewers' states but also by the spatio-semantic structures of visual content. This study proposes a novel gaze analysis framework which introduces the content creators' point of view to understand the meaning of browsing behavior. Visual content such as web pages, digital articles and catalogs are comprised of structures intentionally designed by content creators, which we refer to as *designed structure*. This paper focuses on two design factors of designed structure: spatial structure of content elements (content layout), and their relationships such as "being in the same group". The framework was evaluated with an experiment involving 12 participants, wherein the participant's state was estimated from their gaze behavior. The results from the experiment show that the use of design structure improved estimation accuracies of user states compared to other baseline methods.

key words: *eye tracking, catalog browsing, user states, content design*

1. Introduction

Understanding a user's gaze behavior while browsing visual content such as digital articles, web pages and catalogs is of great interest for various applications. Interface design, usability, and user state estimation can all be improved by understanding gaze behavior. Gaze analysis is a long-standing topic in various fields including visual psychology [1] and human computer interaction [2]. However, understanding content browsing behavior in uncontrolled environments is still challenging because human gaze can be very complex; it is affected not only by viewers' situations but also by the spatio-semantic structures of visual content. To interpret the meanings of eye movements, i.e., to associate gaze patterns with human internal states such as motivations or interests, the following question should be considered: which characteristics of content structures are to be used for interpreting eye movements and how?

Gaze transition patterns, i.e., sequential patterns of gaze-targets in the visual content, have been considered important clues in understanding browsing behavior [3]–[6]. For example, in a previous work [3], refixation patterns of the form X-Y-X... are used to identify pair comparison for analyzing multi-alternative choice processes in catalog browsing. Refixation patterns are also considered as a factor that indicates viewers' uncertainty about their answers in translation tasks [4]. Gaze transition patterns are used to manage human-computer dialogue by estimating the viewer's interests toward objects on the screen [6] and the viewer's engagement in conversation with displayed agents [5].

Although the previous methods perform well in their experimental settings, it is still difficult to create a versatile gaze analysis method that can deal with diverse situations due to the following problems. First, to achieve semantic understanding of browsing behavior, semantic information of visual content needs to be taken into account. In previous methods, each object region on the screen is annotated with semantic labels to characterize gaze transitions [4], [5]. The labels are normally task specific and heuristic, otherwise they can be too diverse due to the variety of semantic information of visual content. Therefore, appropriate labels and object region boundaries need to be defined by analysts for each situation or task. Second, human gaze is affected not only by what kind of objects the visual content contains but also by where and how they are displayed on the screen [7]–[9]. That is, the gaze model should employ content appearance structure (e.g., spatial layouts of images and text). Moreover, the semantics and appearance structures in the visual content are highly related with each other. For instance, semantically similar objects are often placed close together in space. Therefore, we need to consider how to model both of the structures and employ them jointly to understand gaze behavior.

In this study, we present a novel framework to understand browsing behavior by leveraging the content structure which reflects the content designers' point of view — for simplicity, the structure is referred to here as *designed structure*. Visual content such as web pages, digital catalogs and pamphlets usually have inherent structure designed by content creators. Content designers decide the positions and decorations of content elements (e.g., images and text) based on their perceptions and intentions (e.g., "group items by their categories") and emphasize the relationships among the elements by design components (e.g., frames). Since content designers usually organize content structure to make the information in the content more comprehensible to viewers, we hypothesize that viewers' gaze behavior is highly affected by the designed structure. At the same time, viewers may also have their own intentions while content browsing. Here we hypothesize that the effect on viewers' gaze

Manuscript received December 15, 2014.
Manuscript revised March 27, 2015.
Manuscript publicized May 8, 2015.

[†]The authors are with the Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.
a) E-mail: ishikawa@vision.kuee.kyoto-u.ac.jp

DOI: 10.1587/transinf.2014EDP7422

behavior by designed structure is dependent on the viewers' intention. For instance, a viewer might browse object regions in consecutive order when he/she does not have a strong goal/purpose, meanwhile, a viewer might ignore the order when he/she is searching for specific information.

In the proposed framework, browsing behavior is characterized by which part of designed structure attracts the focus and how behavior is influenced by designed structure (in other words, how compliant is the user?). This paper focuses on two essential factors of designed structure: low-level spatial relations of media regions (content layout), and high-level semantic relations among content elements emphasized by a content designer such as "being in the same group". The contributions of this study are as follows: (1) a gaze analysis framework which introduces *designed structure* for understanding content-browsing behavior, (2) a basic estimation method of viewers' state based on the proposed framework. Introducing designed structure enables a simple representation of visual content; therefore, we expect that it can deal with a variety of semantic and spatial structure of visual content. In this paper, we conduct an experiment to evaluate our proposed framework using two types of content design. This paper assumes a common online shopping situation, and the proposed framework is evaluated by measuring the performance of estimation of viewers' state in digital catalog browsing such as "acquiring item information" and "comparing items".

2. Related Work

Recent development of eye trackers enables us to obtain large gaze datasets with less effort. As this data has become available, more attention has been given to machine learning techniques in recent gaze studies. Gaze-motion features (e.g., durations of fixations) have been utilized as inputs to machine learning algorithms [10], [11]. However, to understand the semantic meaning of browsing behavior, both gaze-motion features and semantics of in-focus content should be considered. This section introduces previous studies on statistical learning of gaze-motion features and some gaze analysis methods that employ content information to characterize gaze behavior.

Statistical learning of gaze-motion features.

Statistical approaches using gaze-motion features are considered a robust and task-independent method for gaze analysis, and they are utilized in various situations. For example, Sugano et al. classify gaze-motion features using a Random Forest algorithm to estimate viewers' preference toward pictures [11]. Pasupa et al. combined features of gaze-motion and image features are used for improving the accuracy of search results in information retrieval [12]. In [10], gaze-motion features are learned by using Support Vector Machines to discriminate non-intentional and intentional eye movements. Previous studies often use multiple features related to fixations (relatively stable gazed positions), or saccades (rapid eye movements between fixations), specifically, fixation positions/durations/numbers and saccade lengths/directions/durations. Although the use of gaze-motion features is powerful with sufficiently large data and a specified situation/task, gaze-motion features alone cannot deal with the semantic meaning of browsing behavior. For example, re-fixation patterns of the form X-Y-X can be considered "comparing items" or "uncertainty" depending on the semantic relations of two regions being looked at as mentioned in Sect. 1. Therefore, not only gaze-motion features but the semantics of visual content should be considered to achieve semantic analysis of browsing behavior.

Using content semantics for gaze analysis.

In previous studies, semantic labels which were annotated to areas-of-interest (AOIs) on-screen were used to understand the content semantics of gaze behavior [5], [8]. For example, in [5], they use semantic information of visual content to estimate the viewer's engagement in conversation with displayed agents. Semantic labels such as "the agent's head" and "the object that the agent is explaining" are annotated with regions on the screen, and N-gram analysis is applied to the focused semantic label sequence. This approach is considered to be a simple and effective way to introduce semantic information of visual content. However, it can only deal with the specified domains/situations since it becomes difficult to determine appropriate labels for object regions as visual content becomes complex and diverse. For example, as the number of the objects contents in the visual content increases, the number of semantic labels will become larger.

Effects of content design on gaze.

Information about content design has been considered an important factor that controls human gaze. The previous studies on web page design found that viewers' gaze behavior is highly related to different types of content layouts [9], [13], [14]. Although previous studies obtained important findings from observing gaze data in different conditions, their main motivation is to investigate the effects of page design on viewers' task performance or on the likelihood of a user looking at a particular region of the screen. There are still few studies that introduce design information to understand the meanings of browsing behavior.

The primary contribution of this study is a framework that incorporates content design information to understand browsing behavior. Instead of using semantic properties of content objects, this study focuses on relationships among the objects to prevent the above mentioned problem with semantic complexity of visual content. Modeling gaze behavior with content design can contribute to a new generation of visual content studies such as automated content creation.

3. Modeling Visual Content for Interpretation of Gaze Transitions

This section first presents the description of the visual con-

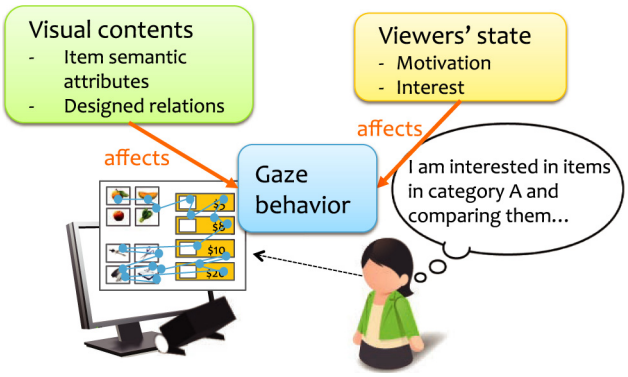

Fig. 1 A catalog browsing situation.

tent browsing situation that is studied in this paper. We assume the entities of the browsing situation as shown in Fig. 1. Visual catalogs contain structure that is intentionally designed by content creators. Simultaneously, viewers may have their own intentions while content browsing, which naturally affects the viewers' gaze behavior. The novelty of this study is to introduce the designers' perspective as one of the entities constituting the catalog browsing situation. In this section, the *designed structure* is introduced in more detail.

3.1 Catalog Browsing Situation

Suppose a viewer is browsing a digital catalog to select a gift for his/her friend. The digital catalog contains a description of several items via various media such as images and text. The viewers' 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 meanings of gaze transition behavior in catalog browsing such as "inspecting item details" or "comparing several items". In this section, three entities constituting the catalog browsing situation: item semantic attributes, designed structure, and viewers' states, are described in detail. Every entity plays an important role in the analysis of catalog-browsing behavior. Moreover, some entities are related with each other, and their relations are also explained.

3.1.1 Item Semantic Attributes

Items in a digital catalog have various semantic attributes such as brand and categories. Semantic attributes are shared by some items in the catalog, and shared attributes can be considered to be important in the analysis of gaze transition behavior. For example, if a viewer is comparing items, it is possible that the common attributes of the compared items are of interest. However, as the number of the items increases, the semantic attributes of items can be more diverse. Thus, when the gaze data is limited, it becomes difficult to obtain meaningful interpretation of browsing behavior using item semantic attributes.

3.1.2 Designed Structure

Designers compose visual content using various media (e.g., images and text) and decorations (e.g., icons and frames) based on pre-existing design criteria such as "place semantically related items close together". The composition of the resulting content structure is what we call *designed structure*. In this study, we leverage group relations among items as the most basic design criteria for the purpose of understanding browsing behaviour. That is, the content structure is represented by a set of items intentionally grouped by designers. The designed structure is considered at two levels: intention-level (IL) and appearance-level (AL). The former expresses high-level semantic relations among items that reflect designers' intention, and the latter expresses low-level appearance information of visual content.

(1) Intention-Level Designed Structure.

Designers usually have prior intentions for visual content, such as "this attribute of items should be emphasized". Designers compose the content structure by emphasizing semantic relationships among items, such as "these items are in the same group". We call the semantic relationships emphasized by designers *design relationships*, and use them to characterize gaze transitions. As the most common design relationships, this paper deals with the followings:

- **Parallel relation.** Indicates similarity of items, e.g., items in the same category are linked with this relation.
- **Contrast relation.** Indicates the difference of items.
- **Ordinal relation.** Indicates the order of items, e.g., items sorted by their ratings are linked with this relation.

These relations can be all translated from the item groups defined by content designers (details in Sect. 3.2). Although more relationships can be considered such as *part-of* relations and *abstract-of* relations, this study only consider the above three relations for simplicity.

In this paper, we assume the designers' intention is given information. This is reasonable to assume since the analysts of browsing behavior can access the information of design relationships among items (e.g., the case that the analyst prepares visual content). However, it is not always true that the analysts can access the information. We discuss a possible method to acquire the intention-level design structure from existing visual content as future work in the following discussion section (Sect. 6).

(2) Appearance-Level Designed Structure.

The information of items is described by various media such as images and text in visual content. The media regions are arranged based on a certain layout. According to the findings from previous gaze studies, the region positions are an important factor that affects viewers' gaze behavior [7]. Therefore, in this study, we focus on spatial layouts as content appearance information. To interpret gaze transitions

among regions on the screen, this study employs spatial relations among regions such as "far from each other".

We assume that the intentions of content designers and the appearance of content are highly related with each other. That is, content designers decide which types of media to be used and compose the structures of the media regions, layouts, and formats to represent their intents (e.g., items being the same group would be placed close to each other). In other words, designed structure is a realization of content designers' intentions. It should also follow that design conventions and rules to make the content information comprehensible to potential viewers. The details of the further discussions on design conventions and rules to realize designed structure as actual visual content are given in Sect. 6.

3.1.3 Viewers' States and Gaze Behavior

When viewers browse visual content, they have internal states such as "examining item A" and "comparing item A and item B". The goal of this study, namely understanding content-browsing behavior, is to associate viewers' gaze behavior with their internal states. Through the following experiment, we investigate which variables of gaze-motion features and structure of visual content are useful to represent content-browsing behavior.

3.2 The Description of Content Structures

Suppose a catalog contains information of a set of items $\mathcal{I}_{All} = \{1, \ldots, N\}$. Each item has a set of P attributes $\mathcal{P}_{All} = \{1, \ldots, P\}$, where p-th attribute can take a value of A_p possible attribute values $\mathcal{A}^{(p)} = \{1, \ldots, A_p\}$. Some attributes can have ordinal relations with each other such as *rating*. In that case, assume that the index of the attribute values are corresponding to the ordinal relations, that is, if $a > a'$ $(a, a' \in \{1, \ldots, A_p\})$, the a-th attribute value is larger than the a'-th attribute value. Here, let us introduce a function $f_p : \mathcal{I}_{All} \rightarrow \mathcal{A}^{(p)}$, where $f_p(i)$ indicates the attribute value of the p-th attribute that the i-th item has.

When content designers create a digital catalog, they decide which aspects of content should be emphasized, and allocate items based on the certain criteria (intention-level designed structure). For example, if the designer emphasize a specific attribute (e.g., *category*), all items in the same category are regarded as "in the same group". To represent this process, we introduce a set of the emphasized attributes as $\mathcal{P}_F \subseteq \mathcal{P}_{All}$. Moreover, the emphasized attributes are categorized into two types: *grouping-attribute* and *sorting-attribute*. The former indicates attributes that are used to group items, and the latter indicates the attributes that are used to sort items. We denote them as $\mathcal{P}_G \subseteq \mathcal{P}_F$ and $\mathcal{P}_S \subseteq \mathcal{P}_F$ respectively.

Here, let us denote intention-level designed relations as $L_{IL} = \{L_{IL}^{(1)}, \ldots, L_{IL}^{(D)}\}$. As mentioned above, this study considers the following three types of designed relations: *parallel, contrast* and *ordinal*. The relations between items are determined using the sets of attributes, $\mathcal{P}_G, \mathcal{P}_S$, as follows.

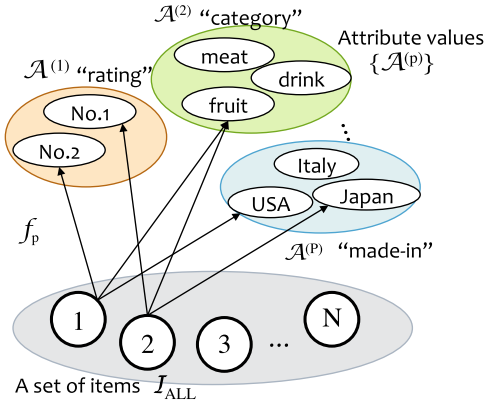

Fig. 2 Descriptions of content structures.

parallel Two different items i and j have this relation when the items share one or more grouping-attributes, that is, $f_p(i) = f_p(j), \exists p \in \mathcal{P}_G$.

contrast Two different items i and j have this relation when the items do not share any grouping-attributes, that is, $f_p(i) \neq f_p(j), \forall p \in \mathcal{P}_G$.

ordinal Two different items i and j have this relation when the items share one or more grouping-attributes, that is, $f_p(i) = f_p(j), \exists p \in \mathcal{P}_G$, and are consecutive in an sorting-attribute $p' \in \mathcal{P}_S$.

Meanwhile, items in a digital catalog are described by various media which occupy spatial regions on the screen, and the media regions are arranged based on a certain layout (appearance-level designed structure). In this paper, we consider item regions composed by a set of media regions to be a basic region for analysis. Let us denote a set of item regions constituting a visual catalog as $\mathcal{R} = \{R_1, \ldots, R_N\}$ $(R_n \subset \Omega, \Omega$: display space), where the information of n-th item is described by a region R_n. As mentioned in Sect. 3.1.2 (2), this study employs spatial relations among items to understand gaze transitions affected by the content layouts. As the layout-oriented gaze transitions, we assume, for example, "looking at items from left to right" or "looking at items along a row". Here, the layout of item regions is represented as several spatial relations $L_{AL} = \left\{L_{AL}^{(1)}, \ldots, L_{AL}^{(D')}\right\}$ between every pair of item regions.

3.3 Examples

We here give a few examples of designed structure.

Example 1: category-based layout

The layout is shown in Fig. 3 left. The catalog includes information about 16 different items ($N = 16$), and each item has two attributes $P_{All} = \{category, price\}$. Here, the designers' focused attribute is $P_F = \{category\}$. The items are grouped according to their categories. The items in the same categories are linked by *parallel* relations and other items in different groups are linked by *contrast* relations. That is,

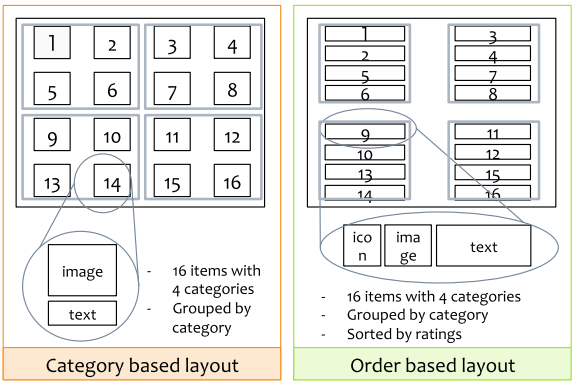

Fig. 3 Examples of catalog layouts.

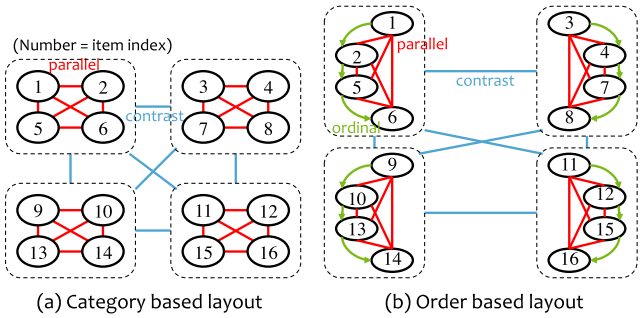

Fig. 4 The intention-level designed structure of the example catalogs. The relation between dotted frames indicate that every item in the frame has the relation with every item in the other frame.

design relations in this layout are $L_{IL} = \{parallel, contrast\}$.

Example 2: order based layout

The layout is shown in Fig. 3 right. The catalog includes information about 16 different items ($N = 16$), and each item has two attributes $P_{All} = \{category, price, rating\}$. Here, the designers' focused attributes are $P_F = \{category, rating\}$. The items are grouped according to their categories and sorted by their ratings in each group. Their ratings are described by an icon of a number from 1 to 4. Besides two design relations in the above example 1, items with successive ratings are linked with *ordinal* relations, i.e., $L_{IL} = \{parallel, contrast, ordinal\}$.

4. Interpretation of Gaze Transitions

This section first describes how to interpret gaze transitions based on the designed structure defined in the previous section. The estimation method of viewers' state based on proposed framework is also described. As a comparative method, we consider using the semantic attributes of items in digital catalogs to interpret gaze transitions.

4.1 Labeling Gaze Transitions Using Designed Structures

Assume the viewers' gaze data are obtained as a sequence of gaze points on the screen, $E = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_T)$ $(\boldsymbol{e}_t \in \Omega)$. Gazed regions can be easily identified by associating each gaze point with an item region. Let us define a function $R : \Omega \to \mathcal{R}$, where $R(\boldsymbol{e}) = R_n$ for $\boldsymbol{e} \in R_n$. As the first step in interpreting gaze transition patterns, a sequence of timings of gaze transitions is obtained as $T = (t_1, \ldots, t_J)$ by finding the timing $t_j \in [2, T]$ that satisfy $R(\boldsymbol{e}_{t_j-1}) \neq R(\boldsymbol{e}_{t_j})$. A sequence of regions being looked at is acquired as $r = (r_0, \ldots, r_J)\left(r_0 = R(\boldsymbol{e}_1), r_j = R(\boldsymbol{e}_{t_j})\right)$. Here, the temporal interval U_j that r_j is being looked at is described as $U_j = [t_j, t_{j+1} - 1]$.

First, for *intention-level design features*, the gaze transitions are associated with design relation labels derived from intention-level designed structure. For each gaze transition on timing t, the two item regions looked at, i, j, are found. According to the rules in Sect. 3.2, the design relation between two items can be determined. Each gaze transition at t_j is represented by a D-dimensional vector $\boldsymbol{x}_j^{IL} \in \{0, 1\}^D$, where the d-th element of $\boldsymbol{x}_j^{IL}$ is 1 if the two items has the d-th design relation label L_d. Finally, a sequence of annotated gaze transitions is obtained as $\boldsymbol{x}^{IL} = \left(\boldsymbol{x}_1^{IL}, \ldots, \boldsymbol{x}_J^{IL}\right)$.

Second, for *appearance-level design features*, each gaze transition is labeled with their corresponding spatial relations among item regions. This study considers spatial directions from a given item region to its four neighbors, i.e., the set of spatial relations is as follows: $L_{AL} = \{$*left-of, right-of, below, above, far*$\}$. Gaze transitions between item regions that are not within four neighbour distance are labeled as *far*. The appearance-level design feature at t_j is represented by one-of-K representation, i.e., the feature is denoted as a $|L_{AL}|$-dimensional vector $\boldsymbol{x}_{jk}^{AL} \in \{0, 1\}^{|L_{AL}|}$, where one of the elements corresponding to the spatial relation as the timing t_j is 1 and all other elements are 0.

Eye trackers sometimes contain noise or miss viewers' gaze points because of blinks. Therefore, the sequence of gaze points is first smoothed by applying a median filter[†]. Moreover, the sequence of regions looked at is modified by discarding intervals shorter than a threshold. In the following experiment, 100ms is used as the threshold. If successive intervals with the same item ID are interrupted by a blank, the intervals are combined to a longer interval.

4.2 Comparative Content Features

To evaluate the effectiveness of using designed structure to understand browsing behavior, we prepare features that use the item-semantic attributes as comparative features. Let us denote a set of attribute values of the item looked at the timing t_j as $\mathcal{S}^{(n)} = \{f_1(n), \ldots, f_P(n)\}$ $(n \in [1, N])$. Here, the attribute values, $\mathcal{S}^{(n)}$, are also described as one-of-K representation, i.e., the p-th attribute is denoted as a Q_p-dimensional vector $\boldsymbol{x}_{jp}^{S} \in \{0, 1\}^{Q_p}$, where one of the elements corresponding to $f_p^{(n)}$ is 1 and all other elements are 0. The vector representing all semantic attributes is obtained by combin-

[†]In this paper, the window size of the median filter is 5 sampling points at 60 Hz (corresponding to about 83msec).

ing the vectors of attributes as $\boldsymbol{x}_j^S = (\boldsymbol{x}_{j1}^{S\mathrm{T}}, \ldots, \boldsymbol{x}_{jP}^{S\mathrm{T}})^{\mathrm{T}}$. Finally, we obtain a sequence of semantic attributes looked at as $\boldsymbol{x}^S = (\boldsymbol{x}_1^S, \ldots, \boldsymbol{x}_J^S)$.

4.3 Analysis of Interpreted Gaze Transitions

As a result of the interpretation, viewers' gaze behavior is represented as multiple sequences of vectors, $\boldsymbol{x}^{IL}$, $\boldsymbol{x}^{AL}$ and $\boldsymbol{x}^S$. In statistical gaze analysis approaches (see Sect. 2 for details), occurrence frequencies of gaze-motion features are often used. The frequency distributions of gaze-motion features are calculated with a certain time period or on each area-of-interest. In this study, the frequency distributions of features are calculated with each interpreted gaze transition sequence ($\boldsymbol{x}^{IL}$, $\boldsymbol{x}^{AL}$ and $\boldsymbol{x}^S$) as

$$\boldsymbol{X}^{IL} = \sum_j \boldsymbol{x}_j^{IL}/J \tag{1}$$

$$\boldsymbol{X}^{AL} = \sum_j \boldsymbol{x}_j^{AL}/J \tag{2}$$

$$\boldsymbol{X}^{S} = \sum_j \boldsymbol{x}_j^{S}/J. \tag{3}$$

4.4 Gaze-Motion Features

The purpose of our experiment is to investigate which content feature is more effective when combined with gaze-motion features. This study uses gaze-motion features from [11] including fixation features (positions, durations, and time) and saccade features (directions, length, duration, and time). Although the gaze-motion features for each right and left half of the screen are defined separately in [11], we do not distinguish them in this paper. Moreover, note that the features related to the time information are ignored in our task estimation (Sect. 5). This is because the task estimation would be obvious if the time information is used since the tasks in the experiment are sequentially given to the participants. In addition to the gaze-motion features, the durations of regions being looked at are also examined. As described in the previous section, the duration of each gaze region r_j is denoted as $|U_j| = t_{j+1} - t_j$. We use the mean value and variance of the durations $\{|U_j|\}$. The gaze-motion features are described as a vector $\boldsymbol{X}^{Gaz}$. All the features used in this paper are listed in Table 1.

5. Experiment

In the experiment, we aim to verify the effectiveness of the proposed framework for interpreting content browsing behavior. Since it is difficult to obtain the ground truth of the meaning of browsing behavior from observed gaze data, in this study, we assume that different types of browsing behavior occur for different states of viewers. A set of tasks are given to participants to induce a variety of states, then the proposed framework is measured by the performance of estimating viewers' tasks from their gaze. This section presents the experimental methodology and the results.

Table 1 Features used in the experiment. The left column shows the name of features. The second and third column from the left shows feature ID with category based layout (CL ID) and feature ID with order based layout (OL ID), respectively. The right column shows the particular labels/relations/values used in the experiment.

	CL ID	OL ID	Features
Item-semantic Feature (price range)	1	1	20,000- yen
	2	2	10,000-15,000 yen
	3	3	5,000-7,000 yen
	4	4	1,000-3000 yen
(Category)	5	5	Accessories
	6	6	Home electronics
	7	7	House-hold goods
	8	8	Toys
(Rating)	N/A	9	First
	N/A	10	Second
	N/A	11	Third
	N/A	12	Fourth
Appearance-level Design Feature (positions)	9	13	Far
	10	14	Left-of
	11	15	Right-of
	12	16	Below
	13	17	Above
Intention-level Design Feature	14	18	Contrast
	15	19	Parallel
	N/A	20	Order
Gaze-motion Feature (duration)	16	21	Mean
	17	22	Variance
(Fixation Position)	18	23	Mean (Horizontal)
	19	24	Mean (Vertical)
	20	25	Variance (Horizontal)
	21	26	Variance (Vertical)
	22	27	Covariance
(Fixation Duration)	23	28	Mean
	24	29	Variance
(Saccade Direction)	25	30	Mean (Horizontal)
	26	31	Mean (Vertical)
	27	32	Variance (Horizontal)
	28	33	Variance (Vertical)
	29	34	Covariance
(Saccade Length)	30	35	Mean
	31	36	Variance
(Saccade Duration)	32	37	Mean
	33	38	Variance

5.1 Experimental Settings

12 participants took part in the experiment. Each participant was asked to sit in front of a screen showing a digital catalog (see Fig. 5). Gaze data of the participants were acquired as 2-d points on the screen by using an eye tracker[†] installed below the screen.

Digital catalogs.

To investigate the general versatility of the proposed framework, we used digital catalogs with two different types of layouts: category based and order based (see Fig. 3).The details of each layout are described in Sect. 3.3. With each layout, four digital catalogs are prepared. Each digital catalog contained the description (images and text) of 16 items,

[†]Tobii X120 (freedom of head movement: 400x220x300mm, sampling rate: 60Hz, accuracy: 0.5degrees)

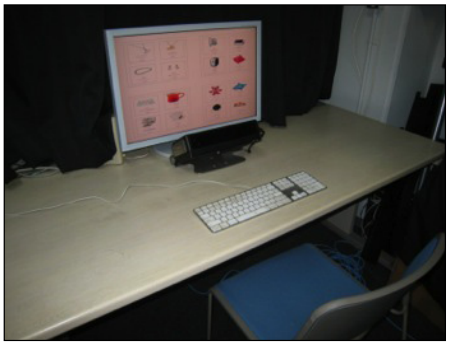

Fig. 5 The experimental environment.

and the items can be grouped into one of four categories: accessories, home electronics, house-hold goods, and toys.

Tasks.

The task required our participant to select a gift for his/her friend. According to previous studies in the marketing research field, the buying decision process can be divided into the following five stages [15]: (1) problem recognition, (2) information search, (3) pre-purchase alternative evaluation, (4) purchase, and (5) post purchase evaluation. Taking the stages into consideration, in the experiment, we use 3 tasks which correspond to stage (2), stage (3) to (4), and the phase just after (4), respectively. Since stage (1) and (5) are not directly related to browsing behavior, we ignore these two stages in the experiment. Specifically, the following tasks were given to each participant, which we call *input, decision* and *free-viewing*, respectively.

1. (30 sec) *Browse a digital catalog to confirm what products exist.*
2. (no limit) *Select a gift from the catalog for a designated person considering his/her profile.*
3. (60 sec)*Browse the catalog freely.*

If the participants were to select a gift for their real friends/acquaintances, it is possible that the personality of the recipient of the gift or the relationship between the viewer and the recipient can affect the viewers' behavior. Since such information is hard to be acquired through the experiment, we designate the recipient by showing the profile of a certain person. The provided profile includes the information of the fictitious relationship with the viewer (i.e., the participant), the hobby, and the portrait.

5.2 Task Estimation

As explained above, the proposed framework is evaluated by measuring the performance of task estimation. The basic idea of the task estimation in this study is to classify the frequency distributions calculated in Sect. 4.3 in a supervised learning manner. In the experiment, we investigate which content feature is more effective when combined with gaze-motion features. For gaze-motion features, in addition to fixation features and saccade features from [11], the mean value and variance of duration lengths between gaze transitions are used. For the details on the gaze-motion features in the experiment, see Sect. 4.4 and Table 1.

Table 2 The estimation accuracies (precisions) and F measures of the three tasks: input (IN), decision (DE) and free-viewing (FV). The 6 rows from the top show the results with the category based layout, and the next 6 rows show the results with the order based layout. The first row shows the results obtained using only gaze-motion features (Gaze). The following four rows show the accuracy obtained combining gaze-motion features with item-semantic features (Sem), appearance-level (AL) design features, intention-level (IL) design features and two content-design-oriented features (AL and IL design features), respectively. The 6th row shows the accuracy obtained combining gaze-motion features with all content features: item-semantic features, AL design features and IL design features.

	Accuracies				F measure		
	Avg.	IN	DE	FV	IN	DE	FV
	Layout Type: Category based						
Gaze	0.583	0.667	0.542	0.542	0.674	0.500	0.584
Gaze+Sem	0.535	0.563	0.521	0.521	0.607	0.459	0.556
Gaze+AL	0.597	0.750	0.521	0.521	0.758	0.485	0.556
Gaze+IL	0.604	0.709	0.563	0.542	0.731	0.529	0.559
Gaze+AL+IL	**0.611**	0.750	0.500	0.583	0.750	0.495	0.589
Gaze+All	0.604	0.729	0.521	0.563	0.737	0.495	0.587
	Layout Type: Order based						
Gaze	0.514	0.583	0.458	0.500	0.615	0.411	0.533
Gaze+Sem	0.556	0.625	0.438	0.604	0.652	0.433	0.586
Gaze+AL	0.576	0.729	0.417	0.583	0.729	0.417	0.583
Gaze+IL	0.583	0.688	0.396	0.667	0.673	0.409	0.660
Gaze+AL + IL	**0.597**	0.667	0.458	0.667	0.696	0.449	0.653
Gaze+All	0.569	0.667	0.375	0.667	0.667	0.391	0.640

For the classification algorithm, we used Random Forest [16]. Random Forest is an ensemble learning method using a set of decision trees. At the training phase, the Random Forest algorithm builds each decision tree using a subset of training data via random sampling with replacement. After training, an unseen feature vector can be classified based on a majority vote from the learned trees. Random Forest is known as a robust and fast learning method and it allows us to assess the importance of feature variables. After the training phase, a part of the training data would be left without being used. Using the out-of-bag (OOB) data as test data, an unbiased classification accuracy can be obtained. The OOB importance of the n-th variable can be calculated by measuring the difference between the original classification accuracy and the accuracy obtained by randomly permuting the n-th variable values in the OOB data.

Finally, we obtained 144 sequences of gaze data (12 participants $\times$ 4 catalogs $\times$ 3 tasks) with each layout type of catalog. The features and their ID numbers are listed in Table 1. For item-semantic features, we consider categories and price ranges of items for the category based layout. For order based layout, we consider ratings of items in addition to price ranges and categories. For more detailed information on the features we used, see Sect. 4.

Estimation accuracies (precisions) were obtained by leave-one-subject-out cross-validation, i.e., gaze data of one subject was used as test data and the rest of the gaze data were used as training. The results are shown in Table 2. With both category based and order based layouts,

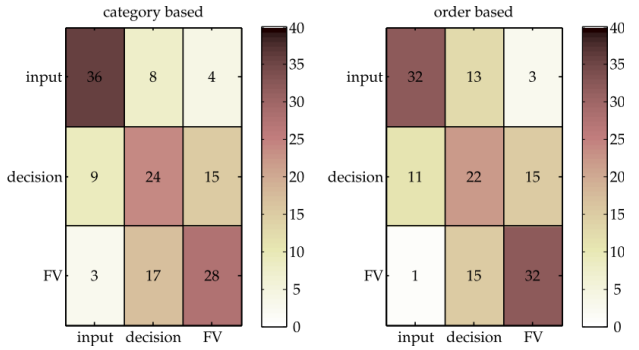

Fig. 6 The confusion matrices of the task estimation. The horizontal axis indicates estimated tasks and the vertical axis indicactes actual tasks. Each number in the matrices indicates the number of gaze data.

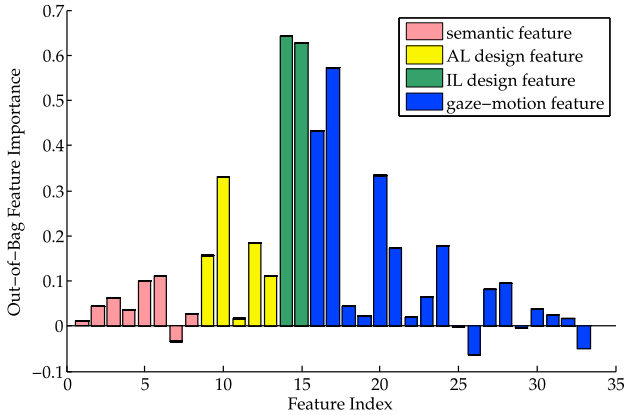

Fig. 7 Out of bag feature importance with category based layout. The feature IDs of horizontal axis correspond to feature ID listed in Table 1. The color of bars indicates the type of features.

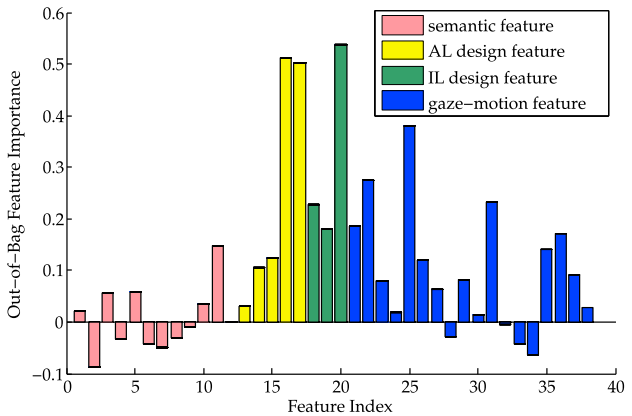

Fig. 8 Out of bag feature importance with order based layout. The feature IDs of horizontal axis correspond to feature ID listed in Table 1. The color of bars indicates the type of features.

the highest estimation accuracy is obtained with the combination of gaze-motion features and two design-oriented features (intention-level and appearance-level design features). The results show that using content design information to interpret browsing behavior is effective than compared with item-semantic features. Moreover, intention-level and appearance-level design features perform almost the same as each other, which indicates that both designers' intentions and content appearance are an important factor in the viewer task estimation.

Moreover, we examined the confusion matrices of the three-task classification to investigate the separability between the tasks (Fig. 6). With both layouts, the input task is estimated with the highest accuracy. It can also be seen by comparing the F measures in Table 2. For the input task, it is possible that participants do not have a clear intention for content-browsing because it occurs before they see a profile of the target of selecting a gift; meanwhile, for the decision task and the free-viewing task, participants browse the digital catalogs based on their own interests/intentions. The results indicate that browsing behavior is more separable in the proposed method if there is a huge gap in degrees of viewers' sense of purpose and/or intentions.

The OOB feature importance is obtained using the 144 sequences of gaze data combining gaze-motion features with all of the content features (item-semantic features, appearance-level design features and intention-level design features). The OOB feature importance with category based layouts is shown in Fig. 7, and one with order based layouts is shown in Fig. 8. In Fig. 7, with category based layouts, the results show that intention-level design feature (feature ID: 14 and 15) has the most importance among other features. Among gaze-motion features, mean value and variance of the durations of each gaze region (feature ID: 16 and 17, respectively), variance of fixation positions (feature ID: 20 and 21) and variance of fixation duration (feature ID: 24) contribute to the estimation. In Fig. 8, with the order based layout, one of variables of intention-level design feature: the frequency of gaze transitions that follow *ordinal* relations (feature ID: 20), has the most importance as well. Moreover, it also shows that horizontal spatial relations in appearance-level design feature (feature ID: 16 and 17) contribute to the estimation. These results indicate the effectiveness of the design-oriented features to represent content-browsing behavior with both content layouts compared to other features.

6. Discussions

In this section, we discuss the limitations and future work of the proposed framework.

The limitation of our approach.

Although the experimental results showed that designed structure is useful to interpret browsing behavior, there is a limitation in that it can only deal with the situation where the viewer understands designed structure by looking at visual content. That is, if the appearance of the visual content does not reflect its designed structure in a comprehensible way to viewers, designed structure has less impact on viewers' gaze behavior. For example, in the category based layout, items with *parallel* relations are surrounded with a frame (see Fig. 3). As a preliminary experiment, we investigated the effects of the representation way of designed structure using

category based layouts without frames. Using the catalog without frames, the accuracy of task estimation was 0.527 using the combination of the gaze-motion features and both of the design oriented features, which is lower compared to the results shown in Table 2. The comparison implies that the proposed method can deal with only gaze behavior with well-designed visual content.

Extending the representation of designed structure

In this paper, we focused on simple design structure with a few types of intention-level/appearance-level design relations among items. For future work, we are extending the representation of design structure to deal with a greater variety of content structures, including hierarchical structures. For this, we consider to introduce a directed graph to represent various relations among content elements. Moreover, for appearance-level design structure, we aim to introduce more rich appearance information such as saliency and characteristics of media.

Potential applications of the proposed framework

As one of possible applications of the proposed framework, we consider to build an information system that can provide suitable information to viewers by estimating viewers' states/situations. In this paper, since the main purpose of the experiment is to verify the effectiveness of the proposed framework, only a simple gift-selecting situation has been considered. To achieve the information system described above, we need to investigate how to model viewers states during content browsing. For modeling viewers' states, we consider to introduce findings from existing research on user state modeling to our framework [17].

How to realize designed structure as visual content?

For the above limitation, we are investigating how to realize the impact of designed structure on viewers' browsing behavior in more detail. If we view the limitation from the opposite side, it implies the possibility of designing visual content which can give us more clues to understand viewers' browsing behavior. We expect that investigating the effects of different realizations of designed structure on browsing behavior would be contributory to various fields such as automated content creation and user interface design. We also aim to build a generative gaze model that can simulate gaze flow with a given visual content.

How to extract design structure from visual content for gaze analysis?

As mentioned in Sect. 3, this study assumes that the designed structures are given, i.e., the analysts of browsing behavior also create visual content or they can access the information of designers' intent. However, it is not always true that the analysts can obtain the information of designed structure. For future work, we are investigating strategies to obtain designed structure from existing visual content. Since it is still a very difficult problem to extract semantic information from a single image, we focus on digital contents such as web pages, and aim to use their source code. Mark up languages, such as HTML and XML, already have descriptions for representing relations of elements composing the visual content, therefore, we consider that it is more feasible to extract designers' intentions based on them.

7. Conclusion

This paper presented a novel framework to interpret content-browsing behavior introducing content design information including spatial layouts and content designers' intentions. An estimation method of viewers' state is also proposed based on the proposed framework. Through the experiment, we confirmed the effectiveness of using content design for gaze analysis by measuring the performance of the viewer state estimation.

For future work, we are investigating relationships between a greater variety of design structures and gaze behavior to build a generative gaze model that can simulate gaze flow with designed structures of a given visual content. Moreover, we also want to understand how to represent design structures to maximize their effect on human gaze, which could be an important contribution to many fields such as automated content creation and user interface design.

Acknowledgments

This work was supported by Grant-in-Aid for JSPS Fellows Grant Number 25·5396 and JSPS KAKENHI Grant Number 26280075.

References

[1] A. Yarbus, Eye movements and vision, Plenum Press, 1967.
[2] 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, ed. R. Radach, J. Hyona, and H. Deubel, pp.573–605, North Holland, 2003.
[3] J.E. Russo and L.D. Rosen, "An eye fixation analysis of multialternative choice," Mem. Cogn., vol.3, no.3, pp.267–276, 1975.
[4] H. Takagi, "Recognizing users' uncertainty on the basis of eye movement patterns: A step toward an effective task assistance system," J. IPS Japan, vol.41, no.5, pp.1317–1327, 2000.
[5] Y.I. Nakano and R. Ishii, "Estimating user's engagement from eye-gaze behaviors in human-agent conversations," Proc. International Conference on Intelligent User Interfaces (IUI2010), pp.139–148, 2010.
[6] P. Qvarfordt and S. Zhai, "Conversing with the user based on eye-gaze patterns," Proc. ACM Conf. on Human Factors in Computing Systems (CHI2005), pp.221–230, 2005.
[7] J.H. Goldberg, M.J. Stimson, M. Lewenstein, N. Scott, and A.M. Wichansky, "Eye tracking in web search tasks: Design implications," Proc. Symposium on Eye Tracking Research and Applications (ETRA2002), New York, NY, USA, pp.51–58, ACM, 2002.
[8] B. Steichen, M.M.A. Wu, D. Toker, C. Conati, and G. Carenini, "Te, Te, Hi, Hi: Eye gaze sequence analysis for informing user-adaptive information visualizations," Proc. International Conference on User Modeling, Adaptation, and Personalization (UMAP2014), vol.8538, pp.183–194, Springer International Publishing, Cham, 2014.

[9] M.L. Resnick and W. Albert, "The Impact of Advertising Location and User Task on The Emergence of Banner Ad Blindness: An Eye Tracking Study," Proc. Human Factors and Ergonomics Society Annual Meeting, vol.57, no.1, pp.1037–1041, 2013.
[10] R. Bednarik, H. Vrzakova, and M. Hradis, "What do you want to do next: A novel approach for intent prediction in gaze-based interaction," Proc. Symposium on Eye Tracking Research and Applications (ETRA2012), pp.83–90, 2012.
[11] Y. Sugano, Y. Ozaki, H. Kasai, and K. Ogaki, "Image preference estimation with a data-driven approach: A comparative study between gaze and image features," Eye Movement Research, vol.7, no.3, pp.1–9, 2014.
[12] K. Pasupa, C.J. Saunders, S. Szedmak, A. Klami, S. Kaski, and S.R. Gunn, "Learning to rank images from eye movements," International Conference on Computer Vision Workshops (ICCV Workshops), pp.2009–2016, 2009.
[13] L. Chen and P. Pu, "Users' eye gaze pattern in organization-based recommender interfaces," Proc. International Conference on Intelligent User Interfaces (IUI2011), pp.311–314, ACM Press, 2011.
[14] B. Pan, H.A. Hembrooke, G.K. Gay, L.A. Granka, M.K. Feusner, and J.K. Newman, "The seterminants of web page viewing behavior: An eye-tracking study," Proc. Symposium on Eye Tracking Research and Applications (ETRA2004), pp.147–154, 2004.
[15] P. Kotler, Marketing management, Millenium Edition, Prentice-Hall, 2000.
[16] L. Breiman, "Random forests," Machine Learning, vol.45, no.1, pp.5–32, 2001.
[17] K. Shimonishi, H. Kawashima, R. Yonetani, E. Ishikawa, and T. Matsuyama, "Learning aspects of interest from gaze," Proc. the 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction (Gaze-in 2013), pp.41–44, 2013.

Erina Ishikawa received the B.E. degree in electrical and electronic engineering and the M.S. degree in informatics from Kyoto University, Japan in 2009 and 2011, respectively. She is currently a Ph.D. student at the Graduate School of Informatics, Kyoto University. Her research interests include human-machine interaction, designing user interface and human vision. She is a student member of IEICE, ACM and IEEE Computer Society.

Hiroaki Kawashima received his MS and PhD in informatics from Kyoto University, Japan in 2001 and 2007, respectively. He is currently an associate professor at the Graduate School of Informatics, Kyoto University, Japan. From 2010 to 2012, he was a JSPS postdoctoral fellow for Research Abroad, and a visiting researcher at the School of Electrical and Computer Engineering, Georgia Institute of Technology. His research interests include hybrid systems, networked control systems, pattern recognition, machine learning, and human-computer interaction. He is a member of IEICE, the Information Processing Society of Japan, Human Interface Society, and the IEEE Computer Society.

Takashi Matsuyama received his B.S. degree and D.Eng. in electrical engineering from Kyoto University, Japan, in 1974, 1976, and 1980. He is currently a professor in the Graduate School of Informatics, Kyoto University. His research interests include knowledge-based image understanding, computer vision, cooperative distributed vision, 3D video, and human-machine interaction. He has received nine best paper awards from Japanese and international academic societies including the Marr Prize at the International Conference on Computer Vision in 1995. He is a fellow of the International Association for Pattern Recognition and the Information Processing Society of Japan, and a member of the Japanese Society for Artificial Intelligence, IEICE, and the IEEE Computer Society.