Reprinted from
Journal of Imaging Science and Technology® 60(6): 060401-1–060401-12, 2016.
https://doi.org/10.2352/ISSN.2470-1173.2017.1.VDA-389

AssisTag: Seamless Integration of Content-based and Keyword-based Image Exploration for Category Search

Kazuyo Mizuno, Daisuke Sakamoto, and Takeo Igarashi
The University of Tokyo, Tokyo, Japan
E-mail: kazuyokojima@gmail.com

Abstract. *Category search is a searching activity where the user has an example image and searches for other images of the same category. This activity often requires appropriate keywords of target categories making it difficult to search images without prior knowledge of appropriate keywords. Text annotations attached to images are a valuable resource for helping users to find appropriate keywords for the target categories. We propose an image exploration system in this article for category image search without the prior knowledge of category keywords. Our system integrates content-based and keyword-based image exploration and seamlessly switches exploration types according to user interests. The system enables users to learn target categories both in image and keyword representation through exploration activities. Our user study demonstrated the effectiveness of image exploration using our system, especially for the search of images with unfamiliar category compared to the single-modality image search. © 2016 Society for Imaging Science and Technology.*

INTRODUCTION

An image search on the Web has become popular in recent years. People explore images daily according to their interests, such as in on-line apparel shopping and Web site design. In the case of category search,[1] users know several example images and seek additional images that are included in the same category as example images. Unfortunately, current image search systems require users to know the appropriate keywords of the category to perform category search. This requirement makes it difficult to explore images effectively without appropriate query keywords that represent target categories.

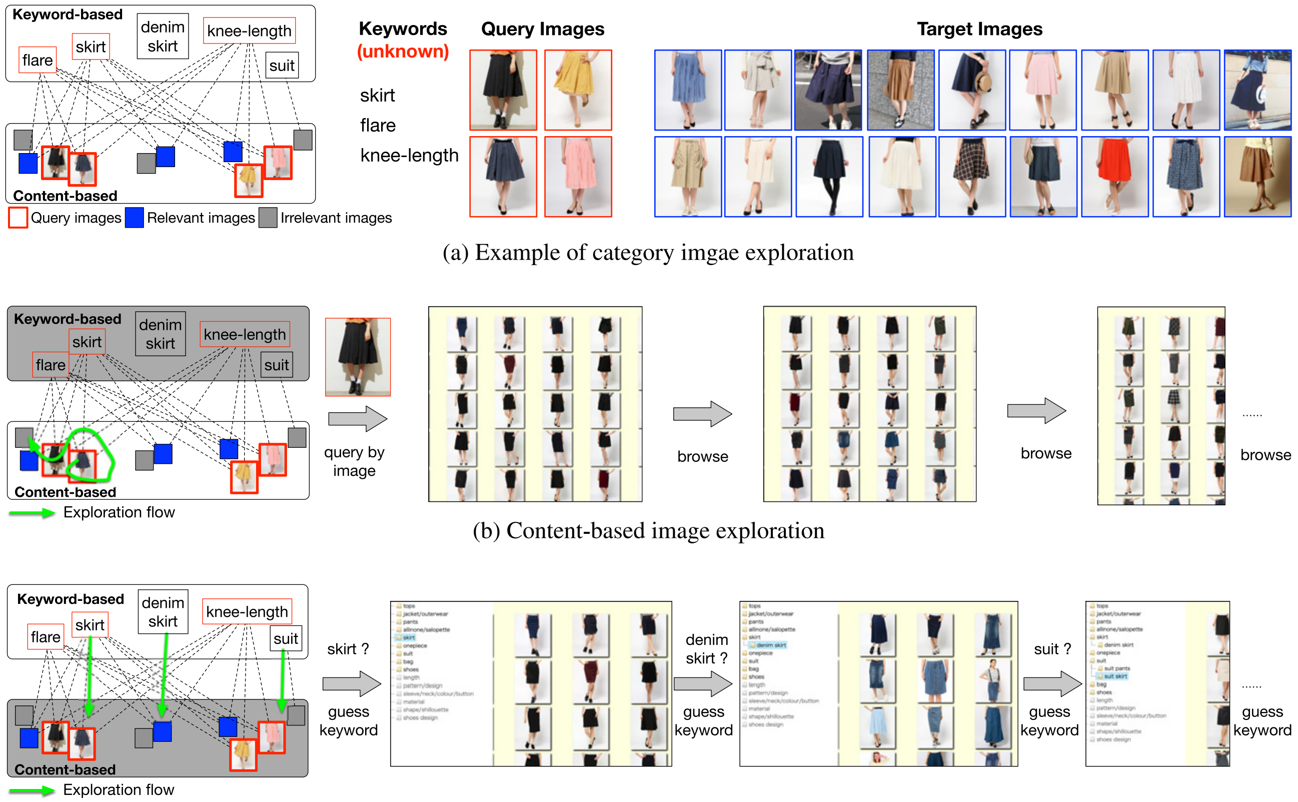
Figure 1(a) shows an example of such image exploration in the real world case of shopping search where users know the visual aspects of the target category, but they do not know the appropriate keyword for their target category. Suppose that a user wants to obtain a collection of images representing "knee-length flared skirt," but he/she does not know these specific keywords. This case is difficult to handle using traditional single-modality search. The user in the content-based search shown in Fig. 1(b) can obtain images visually similar to a query, but this fails to collect images representing objects within the same category but with different visual appearance. The user in the keyword-based search shown in Fig. 1(c) can efficiently collect target images with a different visual appearance, but he/she cannot use it without knowing the appropriate keywords beforehand.

The main objective of this research was to develop a novel image exploration system that achieves category search even if users do not know appropriate keywords for the target category. To do so, we seamlessly bridged content-based and keyword-based image exploration in detail by intuitive keyword suggestions from text annotations associated with images users are interested in. Our system assists users in understanding the target category using both visual and text representations during their exploration activities. Figure 2 is an example of an image exploration workflow with our proposed system in the same case as that in Fig. 1(a). The user starts an exploration by using a related keyword, such as "skirt." The keyword may not exactly match the target category, but the system might display some images showing the target objects. The user then switches to content-based exploration and runs the exploration using the images as search keys, obtaining more images indicating the target objects. The system interactively presents keywords associated with the user selected images such as "flare" and "knee-length," so that the user can eventually find a keyword that represents the target category "knee-length flare skirt" from the presented keywords. If the user wants to focus on the keywords, he/she can change the exploration type to keyword-based image exploration and obtain more images of the target category with different visual appearance. The user can also switch back to content-based image exploration when he/she wants to explore visually similar images.

The technical contribution of our system consists of a keyword suggestion algorithm and a user study to evaluate the proposed image exploration workflow. We automatically define implicit categories from keywords of an input dataset and controlled the relevance of keywords on each image. Our user study demonstrated that keyword suggestion based on our defined implicit categories could successfully support users' image exploration with improved accuracy.

In Related Work, we introduce the previous work in multi-modal image search and exploration-based user interfaces. System Overview presents an overview and a data assumption for our proposed system. Data Structure describes the process of off-line data construction in our proposed system. User Interface describes the interface of our proposed system. User Study presents the results and

Received June 27, 2016; accepted for publication Aug. 29, 2016; published online Dec. 5, 2016. Associate Editor: Yeong-Ho Ha.



(a) Example of category imgae exploration

(b) Content-based image exploration

(c) Keyword-based image exploration

Figure 1. Example of category image search and exploration flow with conventional search interface (Source of images: http://zozo.jp).

discussion on a user study. We conclude the article in Conclusion and Future Work.

RELATED WORK

We looked back on previous work on visualization techniques of multi-modal image search and user interfaces for interactive image search.

Visualization of Image and Text Collections

The multi-modal image search is an image retrieval approach that treats multiple format information. We focused on visualization techniques that handle image local features and annotated keywords of images. We categorized previous works into two types, i.e., integrated and separated visualizations.

Integrated Visualization

Several outcomes obtained from previous work have integrated additional text information into the visualization results of an image collection. Such approaches allow users to simultaneously understand the characteristics of image and text collections. PhotoMesa[2] laid out groups of images in a manner of 2D space filling by using treemap and bubblemap. Janecek et al.[3] proposed an interactive "Focus + Context" visualization technique for exploring a text annotated image collection. They integrated the semantic information of images into the visualization of an image collection where the user could explore the image collection by learning relevant semantic relationships. Visual Islands[4] visualized both visual similarities and cluster information in grid layout images by packing images into the same concept. iMap[5] provided a stable layout of images using visual and text feature space. This system enabled the balance between content and text information to be controlled. iGraph[6] visualized images and text collections by constructing a compound graph that achieved effective visual navigation and comprehension of a massive dataset.

These integrated visualization techniques successfully visualized content and text information simultaneously. However, the layout of images for image exploration is less intuitive for users than that of only visual aspects of images.

Separated Visualization

Other previous work has separately visualized image and text information and also visualized the correlation between image and text.

The Semantic Image Browser (SIB)[7] is a technique for semantic image analysis that conveys annotation results to the user in addition to content-based image browsing. This technique allows the user to understand relationships between images and annotated texts in addition to the similarities among images. Janjusevic et al.[8,9] proposed a concept map for visualizing user query space using a Venn

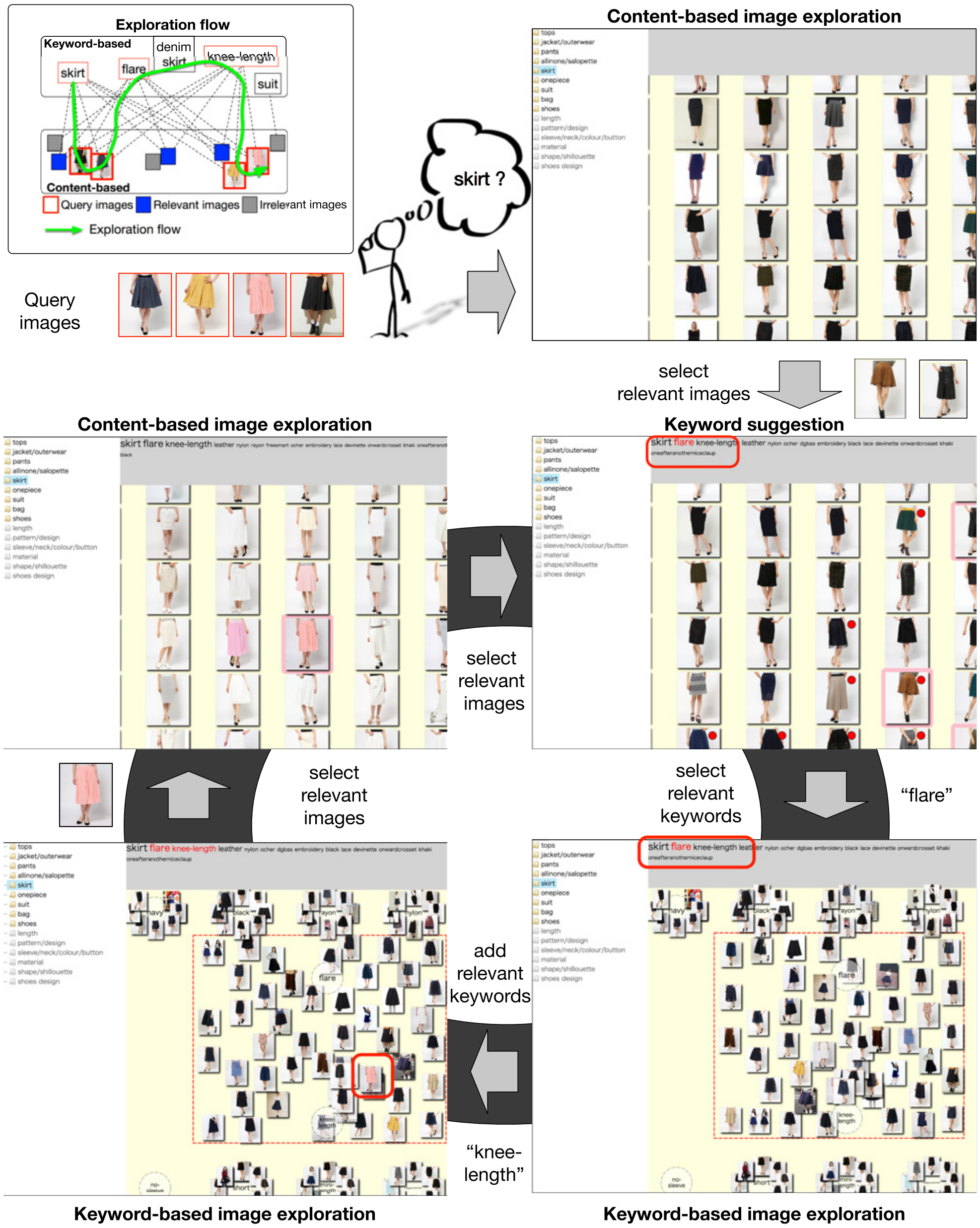


Figure 2. Image exploration flow of our proposed system (Source of images: http://zozo.jp).

diagram and Fisheye distortion. Yang et al.[10] proposed an image exploration system for a large-scale image dataset by constructing a global visual concept network and image summarization for each concept. Their visualization technique assisted users in understanding the coherence between concept-pairs and the visual properties within the concept. Truong et al.[11] proposed a system called concept-aware social image search (CASIS). They constructed a keyword relation graph to assist in finding the user search image. Their visualization approaches were suitable for understanding the overview of image/text feature space, although these approaches have difficulties with detailed image exploration in large-scale image datasets.

JustClick[12] provided a multi-modal image retrieval framework using topic network and content-based similarity visualization. The user first explores his/her interest keywords using a topic network and he/she then moves to focus on a specific topic and explores images by using content similarity. CIDER[13] provides a hierarchical structure of conceptual keywords from user input query text. The user can choose a specific concept keyword and browse images by their content similarity. JustClick and CIDER represent annotated text as an upper layer and image local features as a lower layer. Thus, these approaches have difficulty with starting category search without the appropriate keywords of the target category.

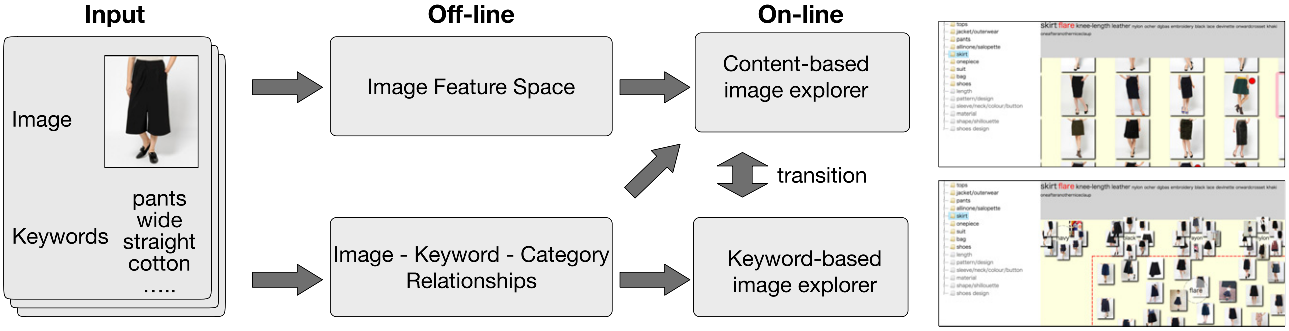

Input
Off-line
On-line
Image
Keywords
pants
wide
straight
cotton
.....
Image Feature Space
Image - Keyword - Category Relationships
Content-based image explorer
transition
Keyword-based image explorer

Figure 3. Overview of proposed method.

We explain how we separately visualize image and text information to preserve the intuitiveness of visual-similarity based image browsing in this article. We investigated how the user could be assisted to learn the target category through image exploration activities using annotated texts.

User Interface for Interactive Image Exploration

Several cases in previous work have investigated data modeling and user interface for interactive image exploration. Interactive image search is a research issue that includes how to interactively guide users to images of interest and how to update feature space and visualization according to user interactions.

Design galleries[14] is a computer-assisted methodology for determining parameters. This system displays representative images in feature space in addition to the distribution of feature space and learns user interest parameters by user image selection. Thomee et al.[15,16] proposed a content-based image exploration interface with pan and zoom operation. They connected pan and zoom operations with manipulation in image feature space. *DynamicMaps*[17] provided a content-based interactive image browser for a massive image dataset. This system interactively displayed visually similar images according to pan and zoom operations. Worring et al.[18] proposed a category search method through the interaction of graph-based visualization. Koike et al.,[19] in the perspective of a specific use case, proposed an image exploration system for on-line apparel shopping while considering women's shopping behaviors.

We investigated a way of guiding the user to the target category through image exploration activities by making use of the annotated text information of images.

SYSTEM OVERVIEW

This section describes a system overview.

Dataset Assumption

First, we describe the assumption underlying our target dataset and task. The main goal of our system is to achieve category search activities without the appropriate keywords of target categories. Our approach mainly relies on a keyword suggestion function during image exploration. Thus, our system has an assumption that each image has a sufficient number of keywords and each keyword has a sufficient number of images.

System Overview

Figure 3 shows an overview of our proposed system. Our system seamlessly integrates content-based and keyword-based image exploration and suggests appropriate keywords of the target category to the user. We compute the data structures of image local feature space and the categories of images offline. We first construct feature space based on the visual aspects of images for content-based image exploration. We also automatically define image categories from images and annotate keywords in the dataset for keyword suggestion and keyword-based image exploration.

Our user interface consists of two components: a content-based image explorer and a keyword-based image explorer. The content-based image explorer provides the image layout based on the visual aspect of images and keyword suggestion according to the user's interested images. The keyword-based image explorer focuses on the images that are included in the user's interested keywords. In addition to general operations at each explorer such as panning and zooming, we also provide a transitional effect between the two explorers.

DATA STRUCTURE

This section describes the data structure construction process. First, we construct feature space based on the visual aspect of images for content-based image exploration. Our system also automatically extracts categories from the distribution of keywords in the dataset and defines the relationship among images, topics, and keywords.

Data Model for Content-Based Image Explorer

This section describes the construction process of the data structure for the content-based image explorer. We applied the hierarchical k-neighbor data structure proposed in *DynamicMaps*,[17] which is an interactive content-based image browser especially for a massive image dataset. It begins by extracting multiple local features from each image and calculates the distance between each image. After that, the system constructs a k-neighbor graph based on the defined distance. The system also incrementally constructs an upper-level k-neighbor graph to support zoom in/out

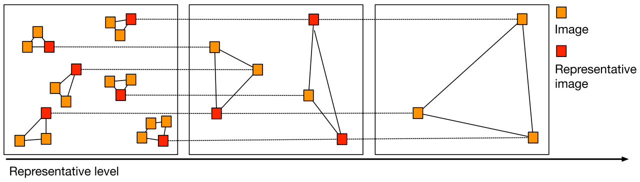


Figure 4. Hierarchical k-neighbor graph of images.

operation, as shown in Figure 4. We extract the average color, color histogram, and gist in the same way as *DynamicMaps* for image local feature extraction.

Topic Detection from Image and Keyword Collection

This section describes the implicit category detection process and relationships among categories, images, and keywords. The main purpose of implicit category detection discussed in this article is to automatically define categories from unstructured image and keyword collections. The proposed system can successfully suggest keywords of target categories by using implicit categories, according to user images and keywords of interest. In practice, we introduce two metrics; relevance between a keyword and an image and similarity between keywords.

We first automatically generate categories from the combination of images and keywords in the dataset by applying a topic modeling technique. Topic modeling is a popular data mining algorithm for discovering the hidden semantic structure in document collection. We adopted the Latent Dirichlet Allocation[20] (LDA) technique to detect topics. Each topic has multiple keywords with probabilistic values and each image has multiple topics with probabilistic values by applying LDA as shown in Figure 5. We consider these topics as implicit categories in this article. Note that keywords can appear in multiple categories so that categories are allowed to overlap.

We respectively denote an image, keyword, and category as $I_i(i = 0, 1, \dots l)$, $T_j(j = 0, 1, \dots m)$, and $C_k(k = 0, 1, \dots n)$. Furthermore, the keyword probability in the category and category probability in the image are denoted as $P_C(T_j, C_k)$ and $P_I(C_k, I_i)$ respectively. Using these probabilistic values, the relevance value between T_i and I_j can be formulated as:

$$R(I_i, T_j) = \sum_{k=0}^{n} P_C(T_j, C_k) P_I(C_k, I_i). \tag{1}$$

We also define the similarity measure between T_j and $T_{j'}$ as:

$$S_T(T_j, T_{j'}) = \sum_{k=0}^{n} X_k(T_j, T_{j'}) P_C(T_j, C_k) P_C(T_{j'}, C_k), \tag{2}$$

where $X_k(T_j, T_{j'})$ is an indicator function when the value is 1 if both T_i and $T_{j'}$ are included in C_k, and 0 otherwise.

USER INTERFACE

Our proposed system consists of two types of image explorers: content-based and keyword-based image explorers, as shown in Figure 6. Users can seamlessly switch between the two explorers according to the user's interested images and keywords.

Content-Based Image Explorer

The content-based image explorer provides an image exploration panel, a keyword suggestion panel, and a keyword filtering panel. Figure 7 shows enlarged snapshots of the user operation workflow using each panel. The user starts with limited images with an initial expected keyword using a keyword filtering function (Figs. 7(a) and (b)) and interactively pans and zooms the image collection according to visual similarities. When the user selects an image of interest, the system suggests several keywords related to the selected image. The suggested keywords are interactively updated by user image selection. Note that these suggested keywords are entrance points to keyword-based image exploration.

We applied the *DynamicMaps*[17] technique to the image exploration panel which can interactively explore visually similar images. The user can control the similarities and diversities of displayed images in this panel by using pan and zoom operations shown in Figs. 7(b) and (c). The user can move positions in the k-neighbor graph of image local feature space using a pan operation and change the representative level of the graph by a zoom operation.

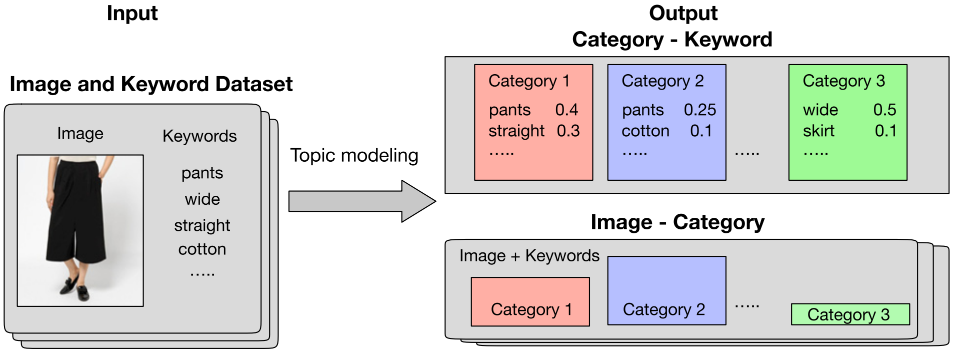
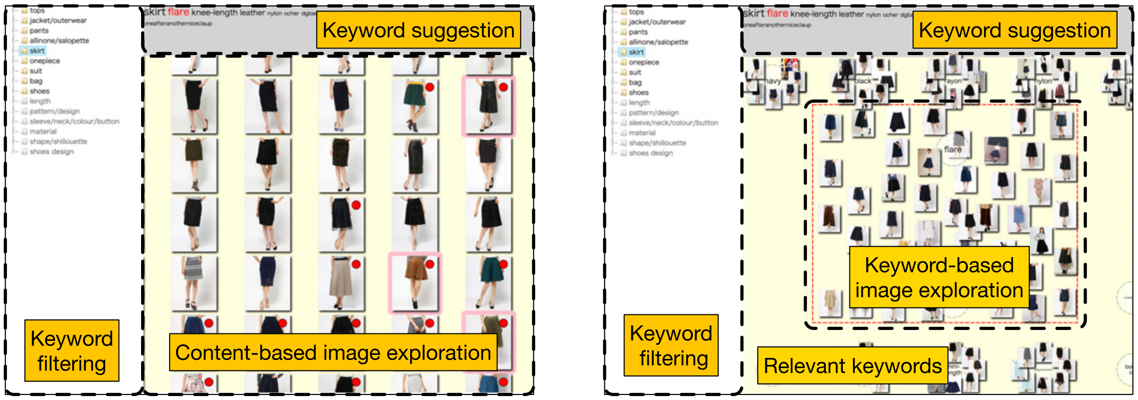


Figure 5. Topic modeling.



(a) Content-based image explorer (b) Keyword-based image explorer

Figure 6. User interface. (Source of images: http://zozo.jp).

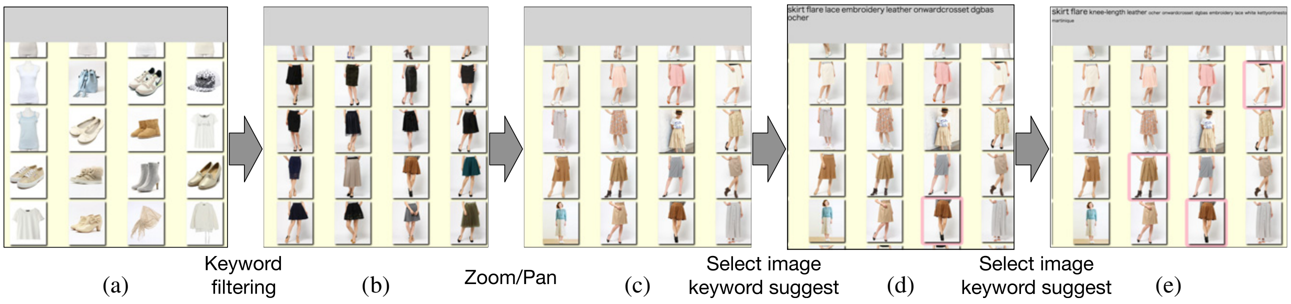

skirt flare lace embroidery leather onwardcrosset dgbas ocher
skirt flare knee-length leather
(a)
Keyword filtering
(b)
Zoom/Pan
(c)
Select image keyword suggest
(d)
Select image keyword suggest
(e)

Figure 7. Example of user exploration workflow. (Source of images: http://zozo.jp).

The keyword suggestion panel provides relevant keywords according to user image of interest to help the user learning the keywords of the target category. Relevant keywords are ordered according to the keyword relevance of selected images, $I_{i'}\,(i' = 0, 1, \ldots l')$. We formulated the relevance value between annotated keyword T_j and selected images $I_{i'}$ as:

$$W_T(T_j) = N_I(T_j) + \sum_{i'=0}^{l'} R(I_{i'}, T_j), \tag{3}$$

where $R(I_{i'}, T_j)$ is the relevance value formulated in Eq. (1) and $N_I(T_j)$ is the number of annotated images in user selected images. Note that we exclude relevant keywords where $W_T(T_j) < 1$. We visualized the keyword relevance value as the character font size.

Keyword-based Image Explorer

The keyword-based image explorer consists of two components: focused keywords and a relevant keywords panel. The user explores images included in focused keywords and he/she also customizes focused keywords using relevant keywords.

The focused keyword panel displays user selected keywords and their relevant images. These keywords are selected by the user at the content-based image explorer. We first choose displayed images and then layout keywords and images while preserving their relationships. We define the relevance value between image I_i and focused keywords $T_{j'}\,(j' = 0, 1, \ldots m')$ as:

$$W_I(I_i) = N_T(I_i) + \sum_{j'=0}^{m'} R(I_i, T_{j'}), \tag{4}$$

where $R(I_i, T_{j'})$ is the relevance value formulated in Eq. (1) and $N_T(I_i)$ is the number of user selected keywords in I_i. Note that, we exclude images from displayed images where $W_I(I_i) < 1$.

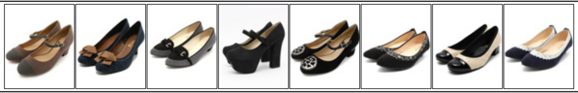
After displayed images are chosen, we arrange the layout of focused keywords and relevant images. Figure 8 shows the process of layout generation. We first place keywords in the uniformed manner shown in Fig. 8(a). We then place relevant images according to the relevance value ratio of the focused keywords (Fig. 8(b)). The position of relevant image I_i is formulated as:

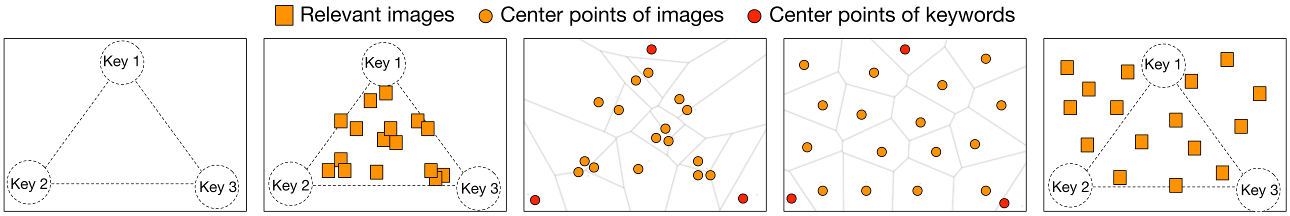
$$\boldsymbol{p}_i = \sum_{j'=0}^{n'} R_n(I_i, T_{j'})\boldsymbol{t}_{j'} \tag{5}$$

$$\sum_{j'=0}^{n'} R_n(I_i, T_{j'}) = 1, \tag{6}$$

where $\boldsymbol{t}_{j'}$ is the position of focused keyword T_j and $R_n(I_i, T_{j'})$ is the normalized relevance value between I_i and T_j calculated with Eq. (1). We generate a Voronoi diagram,

Table I. Examples of categories with their relevant keywords and images. (Source of images: http://zozo.jp).

Category	Keyword	Image
1	$\mathrm{tite}(0.294), \mathrm{skirt}(0.256), \mathrm{knee\text{-}length}(0.222), \mathrm{cotton}(0.112), \mathrm{rayon}(0.049), \mathrm{nylon}(0.039)$	
2	$\mathrm{pants}(0.307), \mathrm{tapered}(0.247), \mathrm{cropped}(0.222), \mathrm{cotton}(0.147), \mathrm{standard}(0.067)$	
3	$\mathrm{clutchbag}(0.350), \mathrm{bag}(0.283), \mathrm{silver}(0.146), \mathrm{black}(0.081), \mathrm{nylon}(0.064), \mathrm{mejane}(0.033)$	
4	$\mathrm{short}(0.393), \mathrm{tee\text{-}length}(0.262), \mathrm{pants}(0.191), \mathrm{cotton}(0.109), \mathrm{rayon}(0.029), \mathrm{nylon}(0.012)$	
5	$\mathrm{pumps}(0.314), \mathrm{shoes}(0.234), \mathrm{round\text{-}toe}(0.129), \mathrm{pointed\text{-}toe}(0.091), \mathrm{artificial\text{-}leather}(0.074)$	



(a) Keyword layout (b) Image initial layout (c) Voronoi diagram (d) Centroid Voronoi tessellation (e) Image final layout

Figure 8. Image layout process.

avoiding image overlaps, from the positions of images and keywords, as shown in Fig. 8(c). We then apply centroid Voronoi tessellation shown in Fig. 8(d) and obtain the final layout (Fig. 8(e)).

The user controls the similarity and diversity of displayed images in the focused keyword area with zoom in/out operation. He/she particularly changes the interval of the chosen index in ordered relevant images according to zoom operation.

The relevant keyword panel displays relevant keywords from user selected keywords. Relevant keywords are chosen according to the relevance topics of selected keywords and their probability value. The relevance value between arbitrary keyword T_j and focused keywords $T_{j'}$ $(j' = 0, \dots m')$ is defined as:

$$R_T(T_j) = \sum_{j'=0}^{m'} S_T(T_j, T_{j'}). \tag{7}$$

The representative images of keywords according to the keyword relevance value of images is also calculated with Eq. (1).

The user can add and delete keywords of interest by dragging keyword objects between the relevant keyword panel and the focused keyword panel. Focused keywords and their relevant images are interactively updated by this user operation. The user can switch back to the content-based image explorer by double-clicking an image of interest in the keyword-based image explorer.

Note that, the user can switch the type of explorer by double-clicking images or keyword objects. We keep common images and keywords in the two explorers and smoothly animate them to keep a mental map in the user's mind, as shown in Figure 9.

USER STUDY

We conducted an explorative user study to evaluate the usability and effectiveness of our system, especially for category search. We set the system conditions and tasks as follows.

Dataset

We employed a situation in apparel product exploration which involves realistic tasks in our daily lives with limited participants' knowledge. We collected 300,000 images and 1656 keywords of apparel products from an on-line apparel E-Commerce (EC) site.[21] Our dataset stored thumbnail image files of all products and keywords. We collected product categories, colors, materials, and bland names as keywords. We set a threshold value for the number of images on each keyword from 0.01% to 30% based on the total number of images in the dataset to exclude keywords that were too rare or too common.

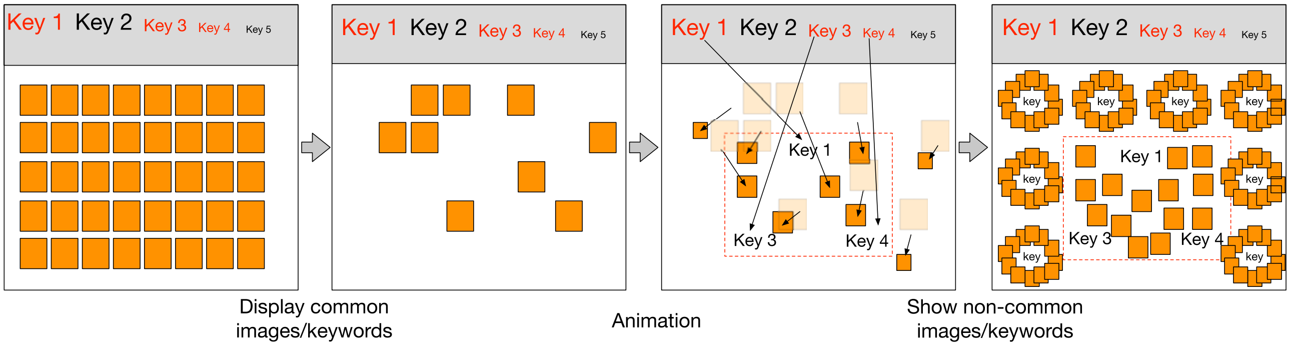

Key 1
Key 2
Key 3
Key 4
Key 5
key
Display common images/keywords
Animation
Show non-common images/keywords

Figure 9. Transition process.

We generated implicit categories from the dataset by applying the topic modeling technique as described in Data Structure. We set the number of topics experimentally while observing the KL-divergence between topics. 200 categories were generated in this study. Table I lists examples of categories and their relevant keywords with probabilistic values. Note that we iteratively increased the number of keywords in one category until all keywords were included in at least one category.

Setup for Study

Equipment

Our prototype system was implemented on a desktop PC with Quad-Core Intel Xeon CPUs (3.2 GHz and 8 MB cache) and 8 GB of RAM; the source code was written in C++ and Python for pre-computation, Javascript for user interface, respectively.

Participant

We recruited eighteen participants (nine males and nine females). We used a recruitment company to find most of the participants. Ten participants were recruited by this service. Another eight participants were recruited by snowball sampling, which involved asking study participants to refer to other people who might also want to participate. Overall, the mean age of the participants was 31.94 (SD = 6.96), and their background, occupations, and expertise were not controlled.

Before we started the study, we administered a pre-study questionnaire and found that the participants' experience of using image exploration was different; one group had experiences with using image search, but the other had not. In addition to image search, we asked the participants about the frequency with which they used on-line shopping. Finally, we divided the participants into two groups: *Experts* and *Non-experts* in image exploration. Participants in the Expert group were selected if: (1) they answered they frequently used the image search (20+ times a week) or (2) they frequently used on-line shopping sites (buy something on a shopping site once a month). Overall, nine participants were selected for the Expert group.

Conditions

We prepared three types of image search interfaces for the user study. These three types were used as conditions.

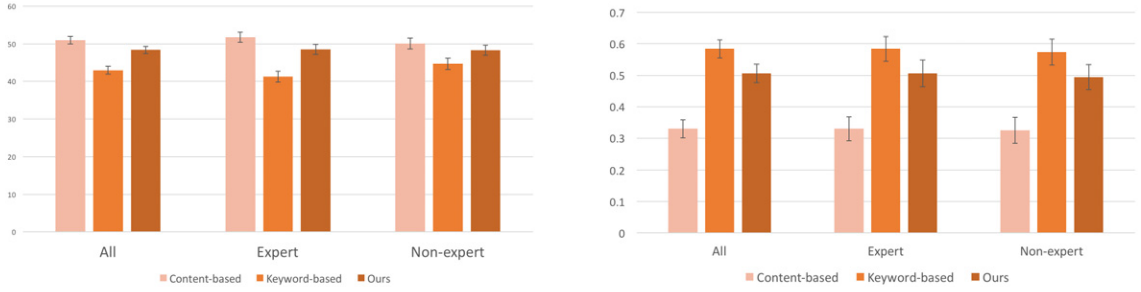
- Content-based search only: An image search system that only provides a browser based on content similarity. In particular, we used the *DynamicMaps*[17] interface.
- Keyword-based search only: Users can only use keyword filtering for browsing images.
- Integrate content-based and keyword-based search (our system): Users can browse images by content similarity and keywords and switch with each other.

Task

Tasks were image exploration tasks, in which participants were asked to search an image on the interface. We first provided participants with several reference images of the same categories without a category name (it was hidden from the participants). At this point in time, we asked participants about the category name of reference images. After that, participants searched and collected images that were in the same category as the reference image using the interface. Finally, they selected one image as an answer from their collected images, and reported its category name. We limited the search time to one minute. Participants tried ten different tasks per one Condition (interface). We prepared thirty different image exploration tasks (reference images and category name) so that no participant experienced the same task within their thirty trials. Note that, we chose a category for task in which participants were not familiar with the field of the category (e.g., male participants sought women's fashions).

Procedure

First, participants were asked to answer the pre-study questionnaire, which included the background information. Next, they were asked to use the three different interfaces (content-based search only, keyword-based search only, and our integrated search interface) ten times each. The order of the task and condition (interface) were counterbalanced among the participants. Prior to the study, the participants completed two training sets per condition (interface). We also conducted semi-structured interviews after the study.



(a) Task completion time

(b) Correct answer ratio (image)

Figure 10. Task completion time and correct answer ratio.

We measured the task completion time, correct answer ratio, and System Usability Scale[22] for quantitative evaluation. The correct answer ratio of images was calculated from the number of common annotated keywords between reference images and the answered image. We also measured the correct answer ratio of category names at before and after exploration to evaluate the learning effectiveness. The correct answer ratio of keywords was calculated from the number of common keywords between the keywords of reference images and user input keywords. We also conducted an interview for the qualitative evaluation.

Quantitative Result

Task Completion Time

The average task completion times for the three interfaces in each groups are plotted in Figure 10. These average times were analyzed using a mixed two-way analysis of variance (ANOVA) (within-subject plan (type of interface) and between-subject plan (expertise) as independent variables and task completion time as the dependent variable). The results revealed no significant differences in interaction effects [F(2,30) = 1.87, n.s.] and the main effect of the expertise (expert or non-expert) [F(1,15) = 0.32, n.s.], but there were significant differences in the main effect of the interface [F(2,30) = 17.20, $p < 0.01$]. Multiple comparison using a Holm test on the simple main effect of the interface indicated that the task completion time for the keyword-based search only interface was significantly faster than that for the other two (MSe = 17.4086, 5% level).

We conducted another one-way ANOVA (within-subject plan (type of interface) as an independent variable and task completion time as the dependent variable) for each expertise (expert and non-expert). The results under the Expert condition revealed that there was a significant difference in the interfaces [F(2,16) = 16.26, $p < 0.01$]. Multiple comparison using a Holm test on the interfaces indicated that the keyword-based search only interface was significantly faster than the other two interfaces (MSe = 16.3512, 5% level). Looking at the non-expert condition, there was a significant difference in the interfaces [F(2,14) = 4.15, $p < 0.05$]. Multiple comparison using a Holm test on the interfaces indicated that there were no significant differences between interfaces (MS3 = 18.6170, 5% level).

Correct Answer Ratio (Images)

The average correct answer ratios (image) for the three interfaces in each groups are given in Fig. 10. These average ratios were analyzed using a mixed two-way analysis of variance (ANOVA) (within-subject plan (type of interface) and between-subject plan (expertise) as independent variables and correct answer ratio (image) as the dependent variable). The results demonstrated no significant differences in interaction effects [F(2,30) = 0.3, n.s.] and the main effect of the expertise (expert or non-expert) [F(1,15) = 0.32, n.s.], but there were significant differences in the main effect of the interface [F(2,30) = 5.29, $p < 0.05$]. Multiple comparison using a Holm test on the simple main effect of the interface indicated that the correct answer ratio (image) for the keyword-based search only interface was significantly faster than the content-based search only interface (MSe = 0.0393, 5% level).

We conducted another one-way ANOVA (within-subject plan (type of interface) as a variable and the correct answer ratio (image) as the dependent variable) for both types of expertises (expert and non-expert). The results under the expert condition demonstrated that there was a significant difference in the interfaces [F(2,16) = 5.43, $p < 0.05$]. Multiple comparison using a Holm test on the interfaces indicated that the keyword-based search only interface had a significantly higher score than the content-based search only interface (MSe = 0.0289, 5% level). Looking at the non-expert condition, there ware no significant differences between the interfaces [F(2,14) = 1.33, n.s.].

Correct Answer Ratio (Keywords)

The average correct answer ratios (Keywords) for the three interfaces in each group are plotted in Figure 11. The pre- and Post-correct answer ratios are separately illustrated in the figure. The purpose of this analysis was to understand what effect the interface had on the learning of keywords while participants were using the interface. These average ratios were analyzed using a mixed three-way analysis of variance (ANOVA) (within-subject plan (type of interface and pre/post results) and between-subject plan (expertise) as independent variables and the correct answer ratio (keywords) as the dependent variable). The results revealed

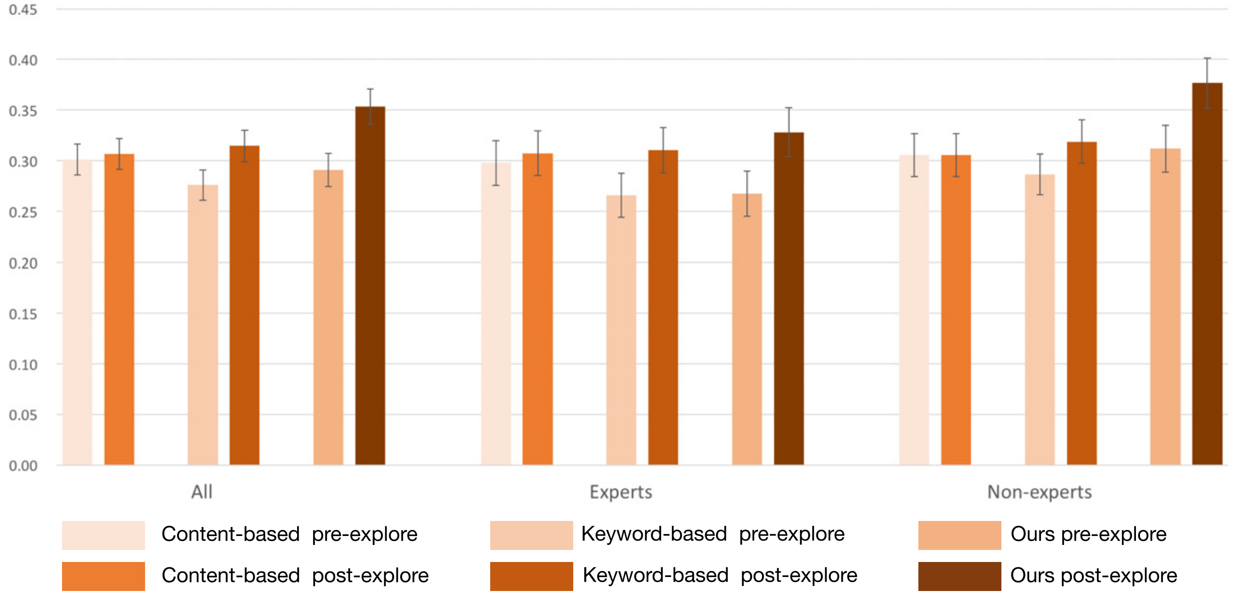


Figure 11. Correct answer ratio (keywords).

there was a significant interaction effect in the type of interface and pre/post results [F(2,30) = 6.16, $p < 0.01$] and the main effect of the pre/post results [F(1,15) = 25.24, $p < 0.01$], but there were significant differences in the main effect of the interface [F(2,30) = 0.07, n.s.] and expertise [F(1,15) = 2.21, $p < 0.10$]. The interaction effect was further analyzed and the results indicated that there were significant differences in pre/post results and the keyword-based search only interface [F(1,15) = 14.18, $p < 0.01$] and the pre/post results and integrated interface (proposed interface) [F(1,15) = 12.7, $p < 0.01$]. There was also a marginal significant difference in the pre/post results and content-based search only interface [F(1,15) = 3.3, $p < 0.10$].

We concluded from the results in Figs. 10 and 11 that our system significantly improved the accuracy of the target category name in the users' minds.

System Usability Scale

The average score on the System Usability Scale (SUS) for the three interfaces in all expertise groups are plotted in Figure 12. These average scores were analyzed using a mixed two-way analysis of variance (ANOVA) (within-subject plan (type of interface) and between-subject plan (expertise) as independent variables and the score on SUS as the dependent variable). The results indicated no significant differences in interaction effects [F(2,32) = 0.84, n.s.] and the main effect of the expertise [F(1,16) = 0.87, n.s.], but there were significant differences in the main effect of the interface [F(2,32) = 3.5, $p < 0.05$]. Multiple comparison using a Holm test on the simple main effect of the interface revealed that the correct answer ratio (image) for the keyword-based search only interface was significantly higher than that for the content-based search only interface (MSe = 223.4520, 5% level).

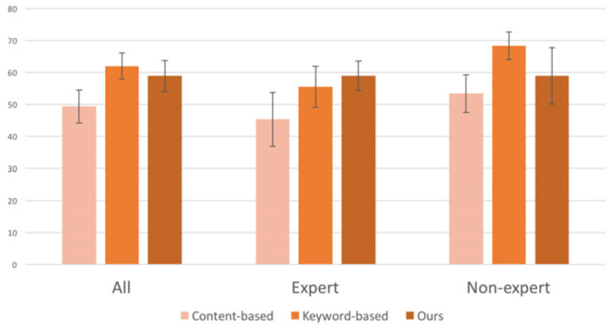


Figure 12. Results from system usability scale.

We conducted another one-way ANOVA (within-subject plan (type of interface) as an independent variable and the SUS score as the dependent variable) for each type of expertises (expert and non-expert). As a result, there were no significant differences between interfaces in any expertise groups [Expert:F(2,16) = 2.5, n.s.; Non-expert:F(2,16) = 1.95, n.s.].

Qualitative Evaluation

We conducted semi-structured interviews after participants had finished their tasks. This section explains how we summarized the qualitative evaluation obtained from the interviews and task activity observations.

Keyword Discovery During Image Exploration Activities

Almost all participants discovered additional keywords for the target category using our proposed system. Participants discovered unknown keywords or known keywords that they could not remember before starting image exploration. For example, P_6 said, "When I was searching images of pants, I found more detailed keywords for target category like 'damage and skinny'." P_9 said, "When I was searching images of shoes, I found an unknown keyword 'side-gore.' I focused on this keyword by changing to keyword-based image exploration and understood the visual aspects of this keyword."

Keyword Suggestion

Almost all participants agreed on the usefulness of this function, which was used for discovering additional keywords of the target category and refine exploration using suggested keywords. For example, P_8 said, "When I sought denim-pants images, I eventually found the keyword 'indigo-blue' on the keyword suggestion panel. I could refine search using this keyword."

Integration of Content-based and Keyword-based Image Exploration

Participants in the expert group preferred this approach. In contrast, non-expert participants expressed an opinion on this approach that they required additional work for image exploration and preferred the keyword-based search only system. For example, P_9 said, "This system is useful for understanding the visual aspects of unknown keywords."

Observations of task activity also indicated that there was a difference between the exploration activities of expert participants and non-expert participants. Participants in the expert group tended to check many images and explore them in more detail by using the entire explore function. In contrast, participants in the non-expert group did not make much effort to refine exploration and tended to use familiar search functions such as category filtering.

DISCUSSION

Summary of User Study

Overall, participants successfully used our proposed interface. We confirmed the following points from the results of the two experiments with regard to the task completion time, the keyword-based search only interface was significantly faster than the other two interfaces. With regards to learning effectiveness, on the other hand, our proposed interface and the keyword-based search only interface were significantly more effective than that for the content-based only interface, and specifically more effective than the previous work on *DynamicMaps*. In short, keyword-based search only interface demonstrated best performance in the quantitative results.

Looking at the quantitative results, participants reported that our proposed interface facilitated their learning of unknown category names or words. Regarding the results from correct answer ratios (keywords), the proposed interface demonstrated better improvements compared to pre- and post-exploration answers. The results obtained from the SUS score indicated expert users' answers on our proposed interface were slightly better than those on the keyword-based search only interface. All in all, the keyword-based only interface quantitatively demonstrated significantly better performance, and our proposed interface attracted positive feedback from participants in the qualitative results. In conclusion, our system improved the accuracy of the target category in the users' minds by suggesting appropriate keywords in users' image of interest in a timely fashion.

Limitation

The proposed interface is a proof-of-concept implementation and the setup for the user study was not comprehensive so that we recognized that there were several limitations.

Comparison with Keyword-based Exploration

As indicated in the results for learning effectiveness, our system successfully helped users to understand the target category during exploration activities. However, our system did not significantly improve the accuracy of answered images. This implies that our system should more intuitively guide users from keywords in their minds to images included in these keywords.

Looking at the results from the other side, our study design compared (1) a well-known, accustomed interface (keyword-based search only interface) and (2) a first-time, not very familiar interface (our proposed interface). Our interface demonstrated similar performance with the keyword-based search only interface, though which we considered that we could demonstrate the significant potential of the interface for the future improvements or practical usage.

Dataset

Our user study only used an apparel product dataset. To exclude the influence of individual participants' prior knowledge about the dataset, we should apply another type of dataset such as natural images in a photo-sharing service (e.g., iStockphoto (http://www.istockphoto.com/)).

CONCLUSION AND FUTURE WORK

This article proposed an integrated image exploration system for category search without users having prior knowledge of appropriate keywords about the target category. We integrated content-based and keyword-based image exploration and bridged each exploration by using a keyword suggestion framework. Our proposed system successfully assisted users to understand the search target category by suggested keywords and keyword-based image exploration. Our user study using an apparel product dataset demonstrated that our system provided learning effectiveness in the search target category during image exploration activities. As future work, we would like to improve user interaction workflow for more intuitively guiding users from keywords in their minds to images included in these keywords. We also intend to integrate the results from automatic image annotations or descriptions into our keyword database in future work, which is one interesting direction for our research topic.

ACKNOWLEDGMENT

This work was supported by Grant-in-Aid for JSPS Fellows (15J00405).

REFERENCES

[1] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1349–1380 (2000).

[2] B. B. Bederson, "PhotoMesa: a zoomable image browser using quantum treemaps and bubblemaps," *Proc. of the 14th Annual ACM Symposium on User Interface Software and Technology (UIST '01)* (ACM, New York, NY, 2001), pp. 71–80.

[3] P. Janecek and P. Pu, "Searching with semantics: an interactive visualization technique for exploring an annotated image collection," *On The Move to Meaningful Internet Systems 2003: OTM 2003 Workshops: OTM Confederated Int'l. Workshops* (Springer, Berlin, Heidelberg, 2003), pp. 185–196.

[4] E. Zavesky, S.-F. Chang, and C.-C. Yang, "Visual islands: intuitive browsing of visual search results," *Proc. of the 2008 Int'l. Conf. on Content-based Image and Video Retrieval (CIVR '08)* (ACM, New York, NY, 2008), pp. 617–626.

[5] C. Wang, J. P. Reese, H. Zhang, J. Tao, and R. J. Nemiroff, "iMap: a stable layout for navigating large image collections with embedded search," Proc. SPIE **8654**, 86540K–86540K-14 (2013).

[6] Y. Gu, C. Wang, J. Ma, R. J. Nemiroff, and D. L. Kao, "iGraph: a graph-based technique for visual analytics of image and text collections," Proc. SPIE **9397**, 939708 (2015).

[7] J. Yang, J. Fan, D. Hubball, Y. Gao, H. Luo, W. Ribarsky, and M. Ward, "Semantic image browser: bridging information visualization with automated intelligent image analysis," *Proc. IEEE Symposium On Visual Analytics And Technology (VAST '06)* (IEEE, Piscataway, NJ, 2006), pp. 191–198.

[8] T. Janjusevic and E. Izquierdo, "Visualising the query space of the image collection," *Proc. 13th Int'l. Conf. Information Visualisation (IV '09)* (IEEE, Piscataway, NJ, 2009), pp. 86–91.

[9] T. Janjusevic, Q. Zhang, K. Chandramouli, and E. Izquierdo, "Concept based interactive retrieval for social environment," *Proc. 2010 ACM Workshop on Social, Adaptive and Personalized Multimedia Interaction and Access (SAPMIA '10)* (ACM Press, New York, NY, 2010), p. 15.

[10] C. Yang, X. Feng, J. Peng, and J. Fan, "Efficient large-scale image data set exploration: visual concept network and image summarization," *Proc. 17th Int'l. Conf. on Advances in Multimedia Modeling—Volume Part II (MMM '11)* (Springer-Verlag, Berlin, Heidelberg, 2011), pp. 111–121.

[11] B. Q. Truong, A. Sun, and S. S. Bhowmick, "CASIS: a system for concept-aware social image search," *Proc. 21st Int'l. Conf. Companion on World Wide Web (WWW '12 Companion)* (IW3C2, Geneva, Switzerland, 2012), pp. 425–428.

[12] J. Fan, D. A. Keim, Y. Gao, H. Luo, and Z. Li, "JustClick: personalized image recommendation via exploratory search from large-scale flickr images," IEEE Trans. Circuits Syst. Video Technol. **19**, 273–288 (2009).

[13] E. Hoque, O. Hoeber, and M. Gong, "CIDER: Concept-based image diversification, exploration, and retrieval," Inf. Process. Manage. **49**, 1122–1138 (2013).

[14] J. Marks, W. Ruml, K. Ryall, J. E. Seims, S. M. Shieber, B. Andalman, P. A. Beardsley, W. T. Freeman, S. F. F. Gibson, J. K. Hodgins, T. Kang, B. V. Mirtich, and H. Pfister, "Design galleries: a general approach to setting parameters for computer graphics and animation," *Proc. 24th Annual Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH '97)* (ACM, New York, NY, 1997), pp. 389–400.

[15] B. Thomee, M. J. Huiskes, E. Bakker, and M. S. Lew, "An exploration-based interface for interactive image retrieval," *Proc. 6th Int'l. Symposium on Image and Signal Processing and Analysis (ISPA '09)* (IEEE, Piscataway, NJ, 2009), pp. 188–193.

[16] B. Thomee, M. J. Huiskes, E. Bakker, and M. S. Lew, "Deep exploration for experiential image re-trieval," *Proc. 17th ACM Int'l. Conf. on Multimedia (MM '09)* (ACM, New York, NY, 2009), pp. 673–676.

[17] Y. Kleiman, J. Lanir, D. Danon, Y. Felberbaum, and D. Cohen-Or, "*DynamicMaps*: similarity-based browsing through a massive set of images," *Proc. 33rd Annual ACM Conf. on Human Factors in Computing Systems (CHI '15)* (ACM, New York, NY, 2015), pp. 995–1004.

[18] M. Worring, O. de Rooij, and T. van Rijn, "Browsing visual collections using graphs," *Proc. Int'l. Workshop on Multimedia Information Retrieval (MIR '07)* (ACM, New York, NY, 2007), pp. 307–312.

[19] E. Koike and T. Itoh, "An interactive exploratory search system for on-line apparel shopping," *Proc. 8th Int'l. Symposium on Visual Information Communication and Interaction (VINCI '15)* (ACM, New York, NY, 2015), pp. 103–108.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res. **3**, 993–1022 (2003).

[21] ZOZOTOWN: "http://zozo.jp" (Accessed: 18 Oct. 2016).

[22] J. Brooke, and Others, "SUS-A quick and dirty usability scale," *Usability Eval. Ind.* **189**, 4–7 (1996).