Effective Browsing of Web Image Search Results

Hao Liu[1], Xing Xie[2], Xiaoou Tang[1], Zhi-Wei Li[2], Wei-Ying Ma[2]

Microsoft Research Asia[2]
5F, Sigma building, No. 49, Zhichun Road
Beijing, 100080, P.R.China
{xingx, i-zli, wyma}@microsoft.com

Department of Information Engineering
the Chinese University of Hong Kong[1]
Shatin, Hong Kong
hliu@cuhk.edu.hk, xtang@ie.cuhk.edu.hk

ABSTRACT

The rapid development of web image search engines has enabled users to search hundred million of images available on the Web. However, due to the unsatisfactory performance of current search technologies, people still need to spend much time in navigating through the large number of result pages to find images of their interest. In this paper, we analyze the main characteristics of web image search results browsing from psycho-physiological and behavioral psychological views and propose to employ a similarity-based organization to present the search results. A user study is carried out to compare our approach with a ranking-based list interface and a cluster-based interface. Experimental results show that visual similarity can help users to explore image search results more naturally and efficiently.

Categories and Subject Descriptors

I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Algorithms, Human Factors

Keywords

Web image retrieval, image browsing, user interface, attention model, multidimensional scaling

1. INTRODUCTION

With the rapid improvements in both hardware and software technologies, large collections of images have been made available on the Web. To help users find images on the Web, many commercial search engines have developed technologies that allow users to search hundred million of Web images based on keywords.

Unlike those text-based retrieval systems, images need to be annotated in advance, either automatically or manually, in order to be indexed by keywords. In most commercial libraries, to ensure the service quality, the photos are manually annotated by editors with titles, keywords or short abstracts. However, for images on the Web, due to the large volume, techniques for automatic annotation or description extraction from web pages are needed.

This work was conducted at Microsoft Research Asia.

The problem of web image retrieval has already been studied for years and a number of approaches have been proposed. According to the information used in the retrieval algorithms, the existing solutions can be roughly divided into three categories:

- **Text-based Approach**: This type of approaches first extracts text information from the web pages containing the image as annotations. Then traditional text retrieval algorithms are applied to search the images. Currently, most of the popular image search engines have adopted this approach, such as Google [1] and Altavista [2].
- **Content-based Approach**: Image analysis techniques are applied to extract visual features from the images, like color, texture, orientation, and shape. These features are then used to find the most similar images to the query image. Typical content based systems have been introduced in [3] and [4].
- **Link-based Approach**: Recently, inspired by the huge success of Google's PageRank algorithm [5], researchers proposed to use the web link structure to improve the image search results. For example, in [6] the authors demonstrated that link information can be used to improve image search.

Compared with the research on algorithms for ranking and indexing Web images, the research on the presentation of search results has received less attention. A simple list-based interface is still the most frequently used interface in commercial systems, and the users still need to perform a great deal of manual page navigations to find the images they are looking for.

Although the problem of browsing and navigation of large image collections has been studied under different contexts, such as personal albums or professional libraries, none of them has taken web images into consideration. The characteristics of underlying data sets can heavily affect the effectiveness of presentation approaches. For example, home photos would be best browsed in a chronicle order [7][8], by location [9], or by person [10]. As for professional photo libraries, they contain more annotations and usually have been classified into different categories. In [11], the experimental results indicate that a random presentation sometimes is better than a similarity-based presentation. This is mainly because that their images come from a professional database and are all relevant to the query. Therefore, interesting images usually have high contrast to their neighbors and appear to stand out in a random organization.

For web images, time or location information would be either unavailable or inaccurate. Though the text in a web page can be a good source of annotation, the quality is unreliable. This makes the traditional list-based interface inadequate since images in the first page are not necessary better than those in the following pages in terms of their relevance to the query.

In this paper, we intend to study a variety of search results presentations for web image search engines. The novel contributions of our work include:

- We analyze the main characteristics of image search results browsing from psycho-physiological and behavioral psychological views and propose a novel visual similarity-based method for search results presentation.
- A user study has been carried out to compare our approach with a ranking-based list presentation and a cluster-based presentation method. Experimental results show that our proposed approach helps users search images more naturally and efficiently.

The remainder of this paper is organized as follows. In Section 2, we analyze several approaches to present the image collection and points out the key techniques required to achieve better browsing experience. In Section 3, we introduce a task-driven attention model for browsing web image search results. A scheme for similarity-based search results presentation and its main features are presented in Section 4. A user study is carried out in Section 5. Finally, Section 6 concludes the paper.

2. INFORMATION ACCESSING IN IMAGE VISUALIZATION

It is well known that the density of photoreceptors in the retina changes significantly from fovea to periphery, which results in high resolution vision at the fovea and progressively lower resolution vision toward the periphery of the vision field. Not all but only a small part of incoming visual information can reach short-term human memory for further processing, i.e., *the Attention as Filter Metaphor* [12][13]. When browsing an image collection on computer screen, such space-variant sampling necessitates frequent gaze shift and fixation stages to determine whether each image is matched with user's preference.

Information Foraging Theory [14] analyzes trade-offs in the value of information gained against the costs of performing activity in human-computer interaction tasks. Cognitive systems engaged in information foraging will exhibit such adaptive tendencies, when feasible, to maximize gains of valuable information per unit cost. The theory has been applied to Web navigation [15]. Information scent [16], which can be regarded as local cues used to assess and navigate towards information resources, provides an indication of the utility or relevance of a navigation path for information foraging.

We will use these theories to analyze the various approaches for web image search results browsing.

2.1 Ranking-based List Presentation

Ranking-based list presentation is a popular visualization method for current commercial image search engines. The image items are organized one by one according to the ranking of each image, and presented in a sequence of web pages.

Let us consider an ordered set of images I returned by a search engine, which can be described as:

$$I = \{I_i\} = \{(ATTR_i)\}, \quad 1 \le i \le M \tag{1}$$

where i is the image index, which indicates the ranking of this image among the returned image collection; $ATTR_i$ describes a set of attributes of the image, some of which are interesting to the user; M represents the total number of images.

Consider a subset I_K of the returned image set, which represents the images the user prefers in the search results:

$$I_K = \{I_k \mid ATTR_k \supset ATTR_P, \quad 1 \le k \le M\} \tag{2}$$

where $ATTR_P$ represents a set of attributes preferred by users, I_K is a subset of I, each member of which is a superset of $ATTR_P$. For example, if a user wants to find images of "Ming Yao," a NBA player, in the returned image collection, the images containing either he playing basketball in Houston U.S. or lunching with his family in Shanghai China are potentially interesting to the user.

One assumption of ranking-based list presentation method is that an image of the higher ranking will be more relevant to the user's preference. However, due to the large volume of relevant images on the Web and limitation of current search technologies, the image subset I_K that best matches the user's preference may be presented neither in the first part of the sequence of web pages nor in a continues way, causing a low information scent for navigation. Based on our observation, people usually start from the first page and browse each image one by one, and navigate from one page to another until they find the target images they need. Therefore, for multi-page list presentation as shown in Figure 1, the information accessing for the image subset I_K can be described as a cost function f_R:

$$f_R(I_K) = \sum_{i<l}(CI_i + CG_i) + \sum_{j<(l/N)} CP_j + CI_l \quad l = \max_{k \in K}(k) \tag{3}$$

where l is the maximum index of I_K, the I_l represents the last target image in the web pages the user is interested in, CI represents the cost people used to process the content of the images, CG is the cost for gaze shifting from one image to another, CP_j is the cost for navigating from current page to the target page j, and N represents the number of images per web page. Equation (3) assumes that the information of image can be processed in one fixation stage, that is, no more gaze shift are needed to get the information of an image. Due to the large quantity of images available, the result collection returned by a search engine is usually presented in thumbnail. So we consider it a reasonable assumption.

As we can observe from Equation (3), the cost function f_R of ranking-based list presentation depends heavily on the distribution of image subset I_K over I. When I_K distributes in the first part or continuously on I, as shown in Figure 1(a), it will be an efficient presentation method for web image search results. However, due to the unsatisfactory performance of current ranking algorithm, I_K usually spreads randomly on I, as shown in Figure 1(b). As a result, people have to scan image one after another sequentially and devote considerable efforts in gaze shifts, page navigations and processing of irrelevant images to find relevant results. Particularly, when users want to compare two images with some common attributes located at different pages, much more efforts are required because of back-and-forth page navigations.

2.2 Cluster-based List Presentation

Consider the cluster-based list presentation method as shown in Figure 2. Cluster analysis is used here to partition image

collection I into groups of image subsets such that the attributes of images in the same cluster are similar. The images of the same cluster are organized in a web page, one page for each cluster. At the same time, the images $I_R = \{I_r\}$ which stands for the profile of each cluster are extracted and presented to users as cues for page navigation. While browsing the search results, people can usually scan the representation images first, and decide which cluster the target images belongs to, and then move forward to the correspondent web page and find target images. Therefore, for cluster-based list presentation, the information accessing for the image subset I_K can be described as a cost function f_C:

$$f_C(I_K) = \sum_{r \in R}(CI_r + CG_r) + CP_t + \sum_{k \in K}(CI_k + CG_k) \qquad (4)$$

where N is the number clusters of image collection; t represents the web page which I_K belongs to. The cost function f_C composes of three parts: $\sum(CP_r+CG_r)$ is the cost used to find which cluster I_K resides; CP_t is the cost to navigate toward the web page containing I_K; $\sum(CP_k+CG_k)$ represents the cost used to process the target image subset.

According to Equation (4), cluster-based list presentation can speed up search process by navigation cues which take advantage of previewing representation images. It enables the user to navigate toward the specific subset that he is interested in. The effectiveness of this approach depends greatly on clustering performance and the quality of representative images of each cluster. Due to the continuous property of attributes space, the boundary between different images is not very clear. It is not a surprise that two images with some common attributes that user prefers will be grouped into different clusters. Under such circumstances, more efforts are required to jump around through multiple clusters. So, as shown in our experiments, it is unnatural to assign image collection into such distinct and mutually exclusive clusters.

2.3 Similarity-based Visualization

Consider the similarity-based visualization method as shown in Figure 3. Some similarity measurement criteria can be employed to narrow down the spatial distribution of images which contain specific attributes, so as to increase the information scent. It provides an overview of image collection, which enables users, while browsing the image collection at first sight, to locate some specific subsets quickly. Therefore, for similarity-based visualization method, the information accessing for the image subset I_K can be described as a cost function f_S:

$$f_S(I_K) = CG_{k_s} + \sum_{I_k \in I_K}(CI_k + CG_k) \quad I_{k_s} \in I_K \qquad (5)$$

where k_s is an arbitrary member of K, which is referred as the first image which attracts user's attention, CG_{ks} presents first gaze shift cost from the original gaze point to I_{ks}.

According to Equation (5), the cost function f_S of similarity-based visualization depends on the distribution of image subset I_K over two-dimensional presentation penal. High information scent can speed up search process and amplify cognition.

Although such similarity-based visualization can help users find the right image quickly especially for web images search results, some hurdles still need to be crossed in order to make people really enjoy it. One problem of similarity-based visualization is to find an effective tool to measure the similarity between two images. Another problem is information loss caused by image down-sampling when we try to put more images on one screen. So a better thumbnail generation technology other than directly down-sampling is needed to use the space more efficiently. In the next Section, a task driven attention model will be discussed to solve above problem.

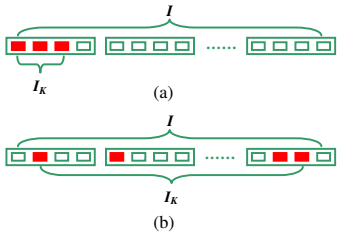


Figure 1. Ranking-based list presentation.

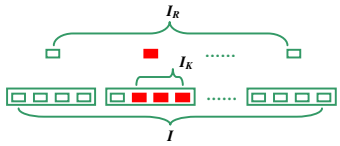


Figure 2. Cluster-based list presentation.

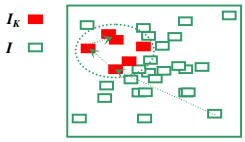


Figure 3. Similarity-based visualization.

3. TASK DRIVEN ATTENTION MODEL

It has been systematically shown by using eye-movements (EMs) experiment that specific tasks can significantly alter the viewing strategy [12]. In most of the time, the user has a specific task when conducting web image search, therefore, the different image regions are of different importance to him. For example, while a user is trying to find "Porsche" on the Internet, suppose an image containing a Porsche running along a sea-side road is returned by search engine. Under such circumstance, the ocean will be regarded as the background, and it will be less attractive for users than the Porsche itself. If the same image is returned while the user is trying to find "ocean", the ocean will be more attractive than Porsche in this case.

An extensible attention model for image adaptation has been proposed recently [17]. Instead of treating an image as a whole, it manipulates each region-of-interest in the image separately, thus allows delivery of the most important region to the client with higher priority. Since general attention includes only limited types of attention objects such as saliency, face and text, we extend the existing model to accommodate use's preference, that is, include those regions that are related to the user's tasks.

Because search engines typically only present thumbnail of images to users, the Minimal Perceptible Scale (*MPS*) attribute [17] is not considered in our design.

Definition 1: The visual attention model for an image is defined as a set of attention objects:

$$\{AO_i\} = \{(ROI_i, AV_i)\}, \quad 1 \le i \le N \tag{6}$$

where

AO_i,	the i^{th} attention object within the image
ROI_i,	Region-Of-Interest of AO_i
AV_i,	attention value of AO_i
N,	total number of attention objects in the image

We assign two attributes to each attention object, which are *Region-Of-Interest (ROI)*, *attention value (AV)*. The notion of '*Region-Of-Interest (ROI)*' is borrowed from JPEG 2000, which is referred as a spatial region within an image that corresponds to an attention object. *Attention value (AV)* is a quantified value indicates the weight of each attention object in contribution to the information contained in the original image.

As shown in Figure 4, we use three steps to build the attention model of each returned image. First of all, we try to map user's task to a set of low-level visual features using the approach in [19]. In the second step, we extract the task related region objects of each image and add these regions to the attention model of the image. Finally, we detect generic attention objects such as saliency and face objects as described in [17][18]. Based on this model, cropping-based thumbnail generation is then employed to present the most important regions of the image to users.

Since some attention objects are extracted based on user's query, we call such image model task-driven attention model. Many time-consuming processes such as image segmentation, saliency and face object detection, can be executed offline, so time cost for building task-driven attention model is acceptable.

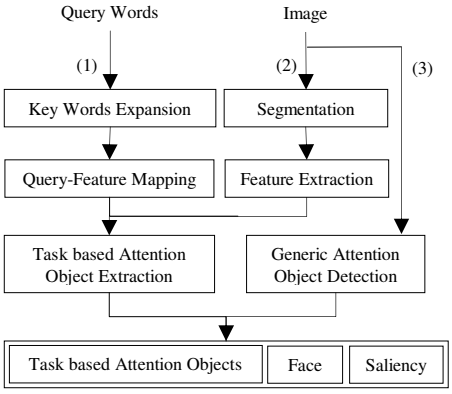


Figure 4. Task Driven Attention Model.

3.1 Query-Feature Mapping

To infer the user's tasks for search results representation, the query keyword sent by the user is the most direct and useful information source.

We employ a compact image thesaurus [19] to help to map the query keyword to visual feature. It includes two components, two-level tree structure texture labels and a list of visual features related to each semantic leaf node of the tree. The relation between visual feature and leaf node is learned and constructed by taking advantages of the abundant image textual annotations available from web pages. The visual feature used here is a linear combination of three level color moments and 36-bin color correlogram.

Firstly we extract hypernym as well as synonyms of the key words by WordNet [20] and expand the query keywords to form a task descriptor vector $QV=\{Q_i\}$. It can be regarded as a set of texture labels representing the attributes of interested images. Secondly, we try to map each item Q_i in QV to a leaf node in the image thesaurus and extract low level features related to the node. When Q_i is mapped to a leaf node, its corresponding visual feature will be assigned to QFV_i; when Q_i is mapped to a non-leaf node, the corresponding visual features of all its children nodes will be assigned to $QFV_i=\{QFV_i^k\}$, where k is referred to the index of children node features; when no match node is available, we will stop the modeling process. After query-feature mapping process, we get a visual feature vector $QFV=\{QFV_i\}$ which describes the QV.

3.2 Task-based Attention Object Extraction

Before extracting task-driven attention objects from image, we segment each image into homogeneous regions using the JSEG [21] algorithm. Each region can be expressed as an object.

Definition 2: The segmented image is defined as a collection of region objects:

$$\{RGN_j\} = \{(RECT_j, RFV_j)\}, \quad 1 \le j \le N \tag{7}$$

where

$RECT_j$,	Region rectangle of RGN_i
RFV_j,	Feature vector of RGN_i
N,	Total number of regions in the image

Two attributes are assigned to each region object. $RECT_j$ is the spatial attribute of the region. Feature vector of RGN_i contains the visual feature of the region.

Then we will use Euclidean distances to measure the similarity between each image region RGN_i and query feature vector QFV_i:

$$D(Q_i, RGN_j) = \| QFV_i^k - RFV_j \| \tag{8}$$

When D_{ij} is smaller than a predefined value, the region j will be labeled as a task related attention object and added to the attention model of the image.

The importance of a task relevant object is usually reflected by its region size and position:

$$AV_{task} = Area_{task} \times W_{task}^{pos} \tag{9}$$

where $Area_{task}$ denotes the area of the task relevant object and W_{task}^{pos} refers to the weight of position in [17].

3.3 Attention-based Thumbnail Generation

As discussed in Section 2, in order to avoid information loss caused by directly image down-sampling, we generate a thumbnail for each image based on task-driven attention model. Our approach, which is similar to the techniques employed in [17][22], automatically crops less informative regions and keep the most informative part of the image, which is called attention region. Cropping can increase the signal to noise ratio of the images presented to user, that is, the ratio of important regions containing target attributes to the unimportant regions containing

distracter attributes. Thus, the spatial resources are used in a more efficient way.

Figure 5 gives an example of the thumbnail generation based on attention model. Originally, three generic attention objects are detected in this image, including two face objects and one saliency objects. When a user submits query keyword "LeBron James", no leaf node is mapped to the query and an optimal thumbnail is generated as shown the right top in this example. When the submitted keywords are "LeBron James AND basketball", a leaf node "basketball" is mapped to the query and a task related attention object is detected. The corresponding thumbnail is then generated as shown the right bottom in the example.

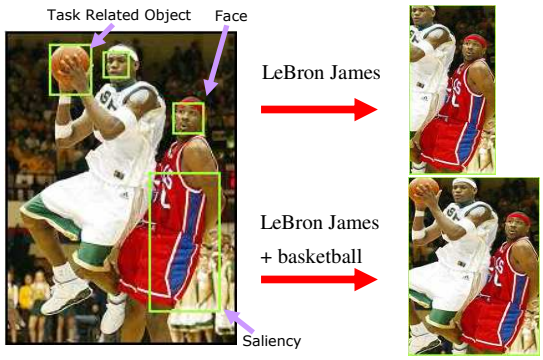


Figure 5. Attention-based thumbnail cropping.

3.4 Attention-based Similarity Measurement

For web image search results, image similarity can be generated based on many properties of the image, such as web link, surrounding text or semantic features. We adopt two sources of information to evaluate the image similarity information: one is the content feature of the whole image and the other comes from attention region of the image, so as to incorporate the semantic concept of image into the similarity measurement.

The similarity of image item and attention region can be measured as Euclidean distance:

$$IM_{ij} = \| FV_i - FV_j \|; \quad (10)$$

$$RM_{ij} = \| FV_{AR_i} - FV_{AR_j} \|; \quad (11)$$

where IM is the image similarity matrix, which is referred as the similarity between image I_i and I_j; RM is the region similarity matrix, which is referred as the similarity between the attention region of image I_i and I_j.

The final similarity measurement between two images is the combination of above two similarity matrix:

$$VM_{ij} = \alpha RM_{ij} + (1-\alpha) IM_{ij} \quad (12)$$

Where α is the weight to achieve a balance between overall similarity, which represent the content of the entire image, and attention region similarity, which represents the attributes of the image preferred by users.

4. SIMILARITY BASED IMAGE SEARCH RESULTS PRESENTATION

In this section, we will introduce the design of a novel web image search interface which is based on the discussion in Section 2. Each image item is presented in an attention-based thumbnail generated by the method described in Section 3.

4.1 Similarity-based Overview

As discussed in Section 2, a similarity-based overview is helpful for browsing web image search results, particularly when the ranking performance is not good. In our approach, visual similarity described in Section 3.3 is used to generate the similarity-based overview of search results.

Given the image similarity matrix VM, we employ the Multidimensional Scaling (MDS) [23] to map each image items into two-dimensional layout space. MDS can achieve this objective by treating inter-object dissimilarities as distances in high dimensional space, and approximating them in a low dimensional output configuration. In this way, similar images are placed nearby and relations between image items are well preserved as shown in Figure 6(a).

A problem of similarity-based visualization is the overlapping layout design, which makes some part of images invisible to user. When two images reside nearby with each other in presentation panel, their relationship in Z-dimension (perpendicular to the panel) is determined by the ranking returned by search engine, that is, the image of higher ranking will take the advantage to appear in front of the one of lower ranking.

4.2 Fitting Images to a Grid View

Although a small degree of overlapping will not affect the user understanding the content of an image [24], while aggressive overlapping will prevent users from finding certain images. In addition, overlapping design produce a very high visual density, Pirolli et al [16] concluded that strong information scent expands the spotlight of attention whereas crowding of targets in the compressed region narrows it. Therefore it is important to provide a control scheme to achieve the balance between information scent and density collection presentation. To solve these problems, we develop a two-dimension grid view to fit all the images into this grid while maximizing the original similarity relationship.

An grid algorithm with space requirement of $O(m^2)$ and time requirement of $O(m^2)+O(n^2)$ was presented in [11], where m is the grid length and n is the number of points in the configuration. Although, it can preserve distances between most closely related objects, the space/time requirement is relatively high. We observe that human vision system is not very sensitive to the absolute grid relationship between each image item and, at the same time, the space/time requirement of the algorithm is more important for search engine to render the grid view in real time. So we design a simple alternative grid algorithm to achieve the balance between the grid precision and space/time requirement.

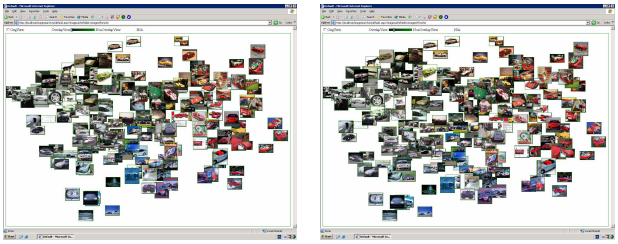
Suppose X and Y are the number of columns and rows of the image presentation panel. Let $I = \{I_i(x_{Sim}, y_{Sim}) \mid 1 \le i \le M\}$ be the returned image dataset, where M is the number of images, $(x_{Sim}$ $y_{Sim})$ is spatial position of I_i in two-dimensional visual space. Let $J = \{1, 2, \ldots, M\}$ be an index set. We order image set $I_1, I_2, \ldots, I_M$ to a sequence $I_{\varphi(1)}, I_{\varphi(2)}, \ldots, I_{\varphi(M)}$ such that $I_{\varphi(i)}(x_{Sim}) < I_{\varphi(j)}$

(x_{Sim}) for $i < j$, where φ is a permutation of the index set J. For each $\{s\}=\{\ s \mid s \in I,\ \max(s+1)\cdot Y < M\}$, denote $K_s = \{\ sY+1, sY+2, \ldots, sY+Y\}$ an index set, reorder image subset $I_s=\{I_{\varphi(sY+1)}, I_{\varphi(sY+2)}, \ldots, I_{\varphi(sY+Y)}\}$ to a sequence $I_{\psi(\varphi(sY+1))}, I_{\psi(\varphi(sY+2))}, \ldots, I_{\psi(\varphi(sY+Y))}$ such that $I_{\psi(\varphi(sY+i))}(y_{Sim}) < I_{\psi(\varphi(sY+j))}(y_{Sim})$, for $i < j$, where ψ is a permutation of the index set K_s.

$$I_i(x_{Grid}) = \lfloor \psi(\varphi(i))/Y \rfloor \tag{13}$$

$$I_i(y_{Grid}) = \psi(\varphi(i)) \bmod Y \tag{14}$$

Since x_{Grid} and y_{Grid} in Equation (13) (14) are integers, we normalize them to fit into the image panel. Noted that X and Y are interchangeable, it is a grid algorithm of optimization in one dimension and sub-optimization in another. By employing the quick sort method, the time and space requirement of the new algorithm is $O(2n\ log(n)\text{-}n\ log(m))$, where n is the number of image item and m is the number of columns or rows.



(a) Similarity-based overview (b) Overlapping ratio γ=0.25

(c) Overlapping ratio γ=0.5 (d) Fitting to a Grid view

Figure 6. From Similarity-based overview to Grid View.

4.3 Dynamic Overlapping Adjustment

In fact, the best overlapping ratio is hard to decide, since it depends on both the image collection and the user. Though different overlapping ratio can be generated automatically using similarity-based overview and Grid view, when both cannot satisfy the user's requirement due to too much information loss caused by overlapping or relationship loss caused by grid algorithm, it is reasonable to allow users to adjust overlapping ratio themselves. Therefore, we provide a slider to let users adjust the spatial position of images to modify the overlapping ratio of the presentation. In our system, the new point where image resides can be determined by the following equation:

$$P^i_{new} = \gamma P^i_{Sim} + (1-\gamma) P^i_{Grid} \tag{15}$$

where γ is the overlapping ratio, P_{Sim} and P_{Grid} is the spatial position where image resides in Similarity-based overview and Grid view. The user can adjust the overlapping ratio r by a slider bar to achieve a suitable presentation. Figure 6(a)-(d) show the prototype of overlapping adjustment from γ=0 to γ=1 with a real browsing example.

Figure 7. Fisheye view.

4.4 Fisheye View

Since users are usually interested in a small part instead of the whole image collection, users will feel more convenient if part of the image collection can be presented in a clearer way. In our approach, users can customize current view by interacting with the interested image item through mouse click, and a fisheye view is then generated to provide such detailed view function.

Fisheye view, an analogue of fisheye lens, is a valuable tool to help users see both local detail and global context information simultaneously [25]. The fisheye view in our implementation, as shown in Figure 7, uses a distorted polar coordinate system, consequently distorting only the spatial relationship of images on the presentation panel. At the same time, the focus image will be substituted with the original (non-cropped) one and shown in detail. The position of the images that are further away from the focus image will appear slightly squashed, that is the further image items are positioned away from the focus, the closer they appear in the image panel. We use following distortion function in our current implementation:

$$G(r) = \frac{(0.5+1)r}{0.5r+1} \tag{16}$$

where r is the normalized distance from peripheral image to focus image.

5. USER STUDY

In order to compare different types of schemes for browsing web image search results, a controlled user study experiment has been carried out. Experimental results are presented and discussed in this section.

5.1 Participants

There were ten participants, six male and four female. They were recruited from nearby universities. All of them were undergraduate students and majored in computer science or communication engineering, except one who majored in physics. The only criterion for selecting the subjects was that they should have frequently performed image search tasks on the web before the study. One subject was for our pilot study and his results were used to revise the interface design.

5.2 Data Sets

The same data set was used in comparing different types of interfaces. Seventeen queries were selected and sent to a commercial image search engine. The first 200 results returned for each query were saved as our test data. Therefore, the data set contains 3400 images in total. The 17 queries we used are:

Table 1. Query terms used in the experiment.

	Query terms	
Practicing	Summer Palace	Christmas card
Testing	dog	fire
	fish	flame tree
	Porsche	Great Wall
	Harry Potter	Chinese
	Hawaii	Lord of Rings
	Forbidden City	New Zealand
	ocean	PDA
	sunrise	

These queries are intended to represent the interests of a large variety of people. Among them, the first two were provided to allow subjects to practice and get familiar with the interfaces.

5.3 Three Types of Interfaces

In the experiment, we compared our approach with a ranking-based list presentation and a cluster-based list presentation.

As shown in Figure 8, we downloaded Google's search results web pages for each query and used them as the ranking-based list presentation. It divides the results into a set of pages, each page contains 20 images and images in the same page are usually arranged in a grid. People can jump among different pages to explore more results by clicking the links at the end of each web page.

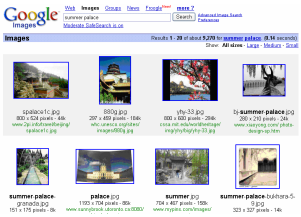


Figure 8. A ranking-based list interface.

For the cluster-based list presentation, as shown in Figure 9, 200 results in each query are grouped into five clusters by k-means algorithm based on the same features used in our approach. For each cluster, four representation images are selected, and provided to users on the left column as the cluster preview. Users can browse the search results by looking at the preview of each cluster at first, and then move forward by clicking the link of interested cluster so as to explore the correspondent images on the right column. Users can switch among five clusters until they find target images in the cluster.

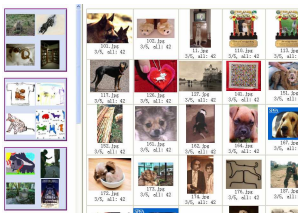


Figure 9. A cluster-based interface.

All the three interfaces are implemented using ASP.NET and java script. Users can select a query and one of the three presentation schemes to view the search results. A 17 inch LCD monitor was provided to each subject and the resolution was set at1280x1024.

5.4 Tasks

Each user was asked to complete all the fifteen testing queries in the same order. For each query, they tried to find a few images most relevant to the query terms. The interface presentation order was varied for each user: three users were given the list-based, then the cluster-based and finally our similarity-based approach; three saw the cluster-based first, our approach and the list-based last; and three saw our approach first, the list-based and the cluster-based last. Thus, each query was used with three users on each interface. This balanced ordering meant that all interfaces were used equally with all queries and any learning or query-interface biases were reduced. In order to reduce any performance influence due to familiarity and experience, the subjects were first asked to try all the three types of interfaces using the two sample queries for a sufficient amount of time.

All the user interactions and corresponding timestamps were recorded for later analysis, including page navigation, dragging of the slider and clicking of the fisheye view. A small questionnaire was given to the subjects after the searching task, in order to get their feedback on the interface design. The questions were:

Table 2. The questions used in the experiment.

ID	Question
1	When do you think web image search will be useful? Give a few scenarios.
2	Which of the three interfaces do you like the best?
3	Do you think using visual similarity to organize images can help searching?
4	Give a score for the usefulness of the three features in our approach (thumbnail cropping, slider and fisheye view).
5	Do you think there are any other potential improvements?

5.5 Experiments Results

The subjects thought that web image search would be very useful while designing homepages, making greeting cards, or writing reports. Some of them also searched images just for fun, for example, searching for photos related to recent news or some famous people. But none of them thought it could be applied to serious scenarios, like authoring books or preparing slides for important talks.

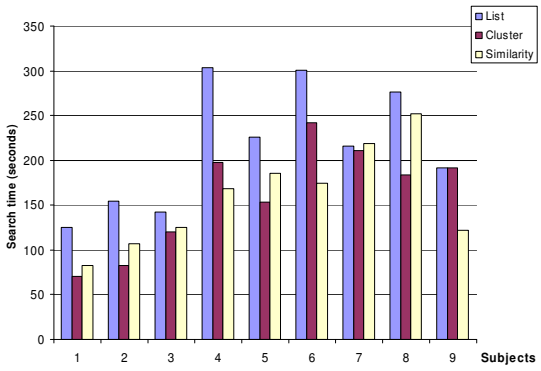


Figure 10. The average search time for each people/interface combination.

The experimental results were analyzed by one way ANOVA with search time as the dependent variable. We applied a log transform to the time to make the distribution more normal. It is found that the type of interface significantly affected the search time ($F(2,132)=9.56$, $p<0.001$). On average, users spent 215 seconds for each query using the list-based interface, 161 seconds for the cluster-based interface and 160 seconds for the similarity-based interface. As we expected, our approach outperformed the list-based interface by reducing 26% of the total search time ($F(1, 88)=15.76$, $p<0.001$). There is no significant difference between our approach and the cluster-based interface ($F(1,88)=0.001$, $p=0.97$).

It is found that the search time was heavily dependent on users ($F(8,126)=14.6$, $p<0.001$). Different people tend to devote different amounts of efforts in searching for satisfying images. Figure 10 shows the average search time for each people/interface combination. The fastest user (user 1) only spent an average of 93 seconds per query while the slowest one (user 6) took 239 seconds. In spite of the variety of searching speeds, most users spent more time on the list-based interface except one user (user 7) who used nearly the same time for all three interfaces. Moreover, query terms did not have a significant impact on the search time ($F(14,120)=0.65$, $p=0.81$). This indicates that there were no significant learning effects.

Seven out of nine liked similarity-based approach while the other two preferred the cluster-based interface. None of the subjects preferred the list-based approach. People who liked similarity-based approach thought that it was more intuitive and interesting, also convenient for comparing similar images. They said that visual similarity was in good accord with human perception, therefore, was helpful to narrow down the searching process. Whereas people who preferred to cluster-based view think that similarity-based approach required more downloading time. One subject suggested that it will be more convenient if three browsing interfaces are provided at the same time so as to let users to decide which one to be used for different requirements.

Table 3. The average scores for the three features.

Function	Thumbnail cropping	Slider	Fisheye view
Score	3.8	4.3	2.8

The average score for the three features are listed in Table 3. The slider got the highest score since it was a convenient way to reduce the overlapping ratio. Many people thought thumbnail cropping is very useful but sometimes delete important information, while fisheye view was not favored comparing to the other two functions.

We also found that current visual similarity is still insufficient especially when the low level features can not represent high level concepts clearly. Under such circumstances, users usually felt the presentation is somewhat disordered. One subject reported that sometimes images of similar concepts were distributed sparsely over the screen. Another subject preferred to classify images into photographs, cartoons, and graphics first, and then use similarity to organize the images within each category. Actually, the technologies proposed in [26] can be applied here to fulfill his needs.

In summary, unlike personal photos or professional photos, visual similarity is found to be useful for browsing web image search results. But the effectiveness of a similarity-based presentation is still limited, mainly due to the gap between semantics and low level visual features. A better interface design should take real application scenarios into consideration, such as, designing web pages or searching for a cartoon or a portrait of a particular person, where some dedicated similarity measures can be adopted.

6. CONCLUSIONS

Currently, the predominant method for image search results browsing is ranking-based list presentation. Due to the unsatisfactory performance of current ranking algorithm, it is a time-consuming process for users to find images of interest in returned image collection.

In this paper, we proposed a novel web image search results browsing strategy based on image similarity to facilitate navigation process over the image search results. A one-page similarity-based overview is constructed and a task driven attention model is applied to optimize the display of information in each thumbnail. Users can interact with the overview by dragging a slider to adjust the global overlapping ratio, and interact with interested image to generate a clear local view. Experimental evaluations indicated that our approach can improve the user's browsing experiences and speed up search process.

We are currently considering improving the image similarity measurement in our approach by incorporating more properties of the image, such as web link, surrounding text or other semantic features. We will continue to investigate these directions in our future work.

7. REFERENCES

[1] Google Image Search. http://images.google.com

[2] AltaVista Image Search. http://www.altavista.com/image

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M, Gorkani, J. Hafner, D. Lee. D. Petkovic, D. Steele, and P. Yanke, the QBIC system, IEEE Computer, Vol. 28, No. 9, pp23-32, 1995.

[4] Y. Rui, T. Huang, and S. Chang, Image Retrieval: Past, Present, and Future, J. of Vis. Com. and Image Rep. Vol. 10, No. 4, pp39-62, Apr. 1999.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, the PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Computer Science Dept., Stanford University, 1998.

[6] R. Lempel and A. Soffer, PicASHOW: Pictorial Authority Search by Hyperlinks on the Web, ACM Trans. on Information Systems, Vol. 20, No. 1, pp1-24, Jan. 2002.

[7] K. Rodden and K. Wood, How Do People Manage Their Digital Photographs, ACM CHI 2003, Ft. Lauderdale, FL, USA, Apr. 2003.

[8] J.C. Platt, M. Czerwinski, and B.A. Field, PhotoTOC: Automatic Clustering for Browsing Personal Photographs, Microsoft Technical Report, MSR-TR-2002-17, Feb. 2002.

[9] K. Toyama, R. Lofan, and A. Roseway, Geographic Location Tags on Digital Images, ACM Multimedia 2003, Berkeley, CA, USA, Nov. 2003.

[10] L. Zhang, L.B. Chen, M.J. Li, and H.J. Zhang, Automated Annotation of Human Faces in Family Albums, ACM Multimedia 2003 poster, Berkeley, CA, USA, Nov. 2003.

[11] K. Rodden, W. Basalaj, D. Sinclair, and K. Wood, Does Organisation by Similarity Assist Image Browsing, ACM CHI 2001, Seattle, WA, USA, Mar. 2001.

[12] D.A. Chernyak, and L.W. Stark, Top-Down Guided Eye Movement, IEEE Trans. on Systems, Man and Cybernetics, Vol. 31, pp. 514-522, 2001.

[13] R. Desimone, and J. Duncan, Neural Mechanisms of Selective Visual Attention, Annual Review of Neuroscience, Vol. 18, pp.193-222, 1995.

[14] P. Pirolli, and S.K. Card. Information Foraging, Psychological Review, Vol. 106, No. 4, pp643-675, 1999.

[15] E.H. Chi, P. Pirolli, K. Chen, and J. Pitkow, Using information scent to model user information needs and actions on the Web, ACM CHI 2001, Seattle, Washington, Mar. 2001.

[16] P. Pirolli, S. K.Card, and M. Van Der Wege, The Effect of Information Scent on Searching Information Visualizations of Large Tree Structures, AVI 2000, Palermo, Italy.

[17] L. Q. Chen, X. Xie, X. Fan, W. Y. Ma, H. J. Zhang, and H. Q. Zhou, A Visual Attention Model for Adapting Images on Small Displays, ACM Multimedia Systems Journal, Vol. 9, No.4, pp. 353-364, 2003.

[18] Y. F. Ma, and H. J. Zhang, Contrast-based Image Attention Analysis by Using Fuzzy Growing, ACM Multimedia 2003, Berkeley, CA, USA, Nov. 2003.

[19] X. J. Wang, W. Y. Ma, and X. Li, Data-Driven Approach for Bridging the Cognitive Gap in Image Retrieval, ICME 2004, Taipei, Taiwan, June 2004.

[20] C. Fellbaum, WordNet: An Electronical Lexical Database, MIT Press, Cambridge, Mass., 1998.

[21] Y. Deng, B.S. Manjunath, and H. Shin, Color Image Segmentation, CVPR'99, Fort Collins, CO., 1999.

[22] B. Suh, H. Ling, B.B. Bederson and D.W. Jacobs, Automatic Thumbnail Cropping and its Effectiveness, ACM Symposium on User Interface Software and Technology (UIST 2003), Vancouver, Canada, Nov. 2003.

[23] W. Basalaj, Incremental Multidimensional Scaling Method for Database Visualization. In Visual Fata Exploration and Analysis VI, Proc. SPIE, V3643, 1999.

[24] H. Intraub, C.V. Gottesman, E.V. Willey, and I.J. Zuk, Boundary Extension Briefly Glimpsed Photographs: Do Common Perceptual Processes Result in Unexpected Memory Distortions? J. of Memory and Language, Vol. 35, No. 2, pp118-134, 1996.

[25] M. Sarkar, and M. H. Brown, Graphical Fisheye Views, Communications of the ACM, 1993.

[26] V, Athitsos, M.J. Swain, and C. Frankel, Distinguishing Photographs and Graphics on the World Wide Web, IEEE Workshop on Content-Based Access of Image and Video Libraries, June 1997.

.