Interactive Browsing via Diversified Visual Summarization for Image Search Results

Jingdong Wang Liyan Jia Xian-Sheng Hua

This manuscript is submitted to ACM/Springer Multimedia Systems Journal
for the Special Issue on "Interactive Multimedia Computing", 2010.

Abstract

Presenting and browsing image search results play key roles in helping users to find desired images from search results. Most existing commercial image search engines present them, dependently on a ranked list. However, such a scheme suffers from at least two drawbacks: inconvenience for consumers to get an overview of the whole result, and expensive computation cost to find desired images from the list. In this paper, we introduce a novel search results summarization approach and exploit this approach to further propose an interactive browsing scheme. The main contribution of this paper includes: 1) A dynamic absorbing random walk to find diversified representatives to summarize image search results; 2) A local scaled visual similarity evaluation scheme between two images through inspecting the relation between each image and other images; And 3) an interactive browsing scheme, based on a tree structure of organizing the images obtained from the summarization approach, to enable users to intuitively and conveniently browse the image search results. Quantitative experimental results and user study demonstrate the effectiveness of the proposed summarization and browsing approaches.

Key words: Visual summarization, visual diversity, interactive browsing, image search results

1 Introduction

Most existing commercial image search engines, such as Google image search, Microsoft Bing image search, and Yahoo! image search, use the associate texts (metadata) of images as the indices of images. Text-based search techniques are thus directly borrowed for image search. Image search results are usually displayed in a ranked list that is similar to the presentation of text search result. This presentation reflects the similarity of the images' metadata to the textual query.

However, such a way to display image search results is inconvenient and inefficient for the user to browse the returned images. On the one hand, there is no intuitive overview of image search results due to lacking a summarization. For example, if the user would like to get the overall content of the returned images, she has to click all the pages (Google image search, Yahoo! image search) or drag down the scroll (Microsoft Bing image search) to look through all the images. Even though the user has viewed all the images, it is still not easy for an ordinary user to immediately get known of the image content for a large number of images (usually about 1000 images). On the other hand, under the list organization of images without taking into consideration the visual contents, it is not convenient for users to search desired images. To find an image with rough visual content in mind, a user has to click all the pages or drag down the scroll to check the whole results.

In this paper, we address the problem of presenting image search results through proposing a novel visual summarization technique, called dynamic absorbing random walk. This summarization technique aims to find several example images as the summary so that 1) Each image represents a different topic (diversity); 2) Each image can well represent a group of images that belong to the corresponding topic (centralization); 3) The images are visually appealing (visual quality); And 4) the summarized images are highly related to the query (relevance). The technical novelty of this approach lies in three aspects: 1) A local scaled visual similarity evaluation scheme is adopted to grasp the local scale of each group of images; 2) A dynamic weight tuning scheme is designed for the absorbing random walk to guarantee a good diversity; And 3) the integrated framework has the ability to exploit the preference of images being representatives, e.g., the relevance, visual quality and so on.

Given the reliable and satisfactory summary of image search results, we further organize the image search results as a hierarchical structure, by recursively finding the summary for each group, and propose an effective interactive presentation scheme, including interactive browsing and browsing

path, to enable users to browse the image search results conveniently. The proposed presentation scheme provides a global overview of image search results, and also allows users to go into the local details to find desired images by a few simple clicks.

The remainder of this paper is organized as follows. Sec. 2 reviews the related work. Sec. 3 presents the proposed visual summarization approach. Sec. 4 introduces the interactive browsing scheme. Sec. 5 reports the experimental results to demonstrate the proposed summarization and browsing schemes. Finally, Sec. 6 concludes this paper.

2 Related work

One of the key research issues in presenting image search results is the organization of images. Several presentation schemes, organizing images based on visual similarties, are proposed in [10]. They focus on delivering a global overview of the selected representative images, but do not offer an interactive scheme to enable consumers to easily find images of interest. Moreover, the images are displayed in a homogeneous size, which makes users not easily capture the image content of the whole image collection at the first glance. A presentation scheme is introduced in [3] to use a visual summarization scheme, but the presented results are not satisfactory because of the limitation of its visual summarization, e.g., no hierarchical organization and without considering diversity, relevance and visual quality. The IGroup browsing scheme in [8] is only based on surrounding text information for grouping, but does not investigate the visual content for summary or offer a hierarchical summarization scheme to enable users to conveniently find images of interest.

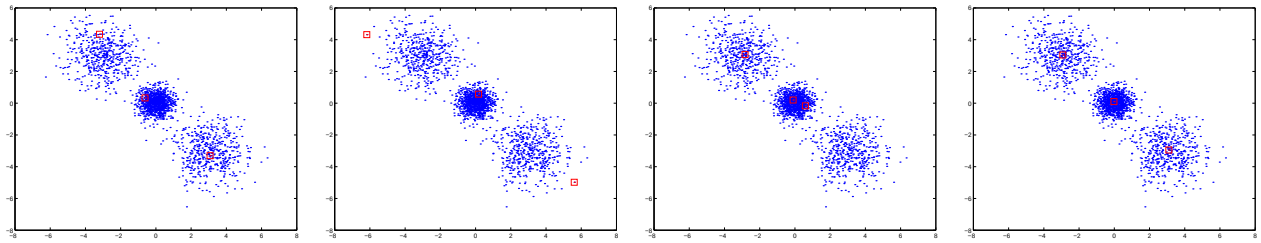
The organization of image search results is closely related to the image clustering techniques. Most existing image clustering techniques directly extract visual features or associated texts to represent each image, and then apply unsupervised clustering methods to group the images. There are many works on image clustering [1, 6, 7, 8, 13]. Some prior works have addressed the summarization task of general image collections. Most of them pay much attention (e.g., [14]) to find a summarization from the associated tags or the annotations alone. Such techniques do not yield satisfactory visual results due to the lack of considerations on the visual aspects. In the following, we mainly focus on reviewing the works on summarizing image search results.

The straightforward solution is to apply existing clustering methods on image features. For exam-

ple, a fast algorithm [7] is proposed based on affinity propagation over the visual feature to directly find the exemplars to represent the image collection. A hierarchical clustering approach [1] is presented to group Web image search results, which focuses on using visual, textual, and link analysis to cluster the search results of ambiguous targets. This method has an apparent disadvantage that the relevance and quality factors are not involved into the grouping process. A textual analysis based approach [8] is presented to find query-related semantic clusters. However, it relies too much on the text, which may lead visually inconsistent results, and another key drawback is that the preference, e.g., relevance and visual quality, is not taken into consideration. A shared nearest neighbors (SNN) approach is used to cluster images [13], which can treat both visual and textual features and take into consideration the original ranking. SNN is a density based clustering method, but it is not easy to guarantee the diversity. Technically, it relies on a global parameter to estimate the density, but with little capability to handle clusters with different compactness. The drawbacks of the above approaches can be summarized into two aspects: they only consider the centralization, i.e., the selected images can well represent the corresponding local group, but cannot guarantee the diversity, or the preference cannot be easily taken into consideration except [13].

There exists some investigation on organizing image search results diversely according to textual feature. For example, the method in [23] is proposed to detect and resolve the ambiguity of a query based on the textual features, which makes the ambiguity nature reflect in the diversity of the result. The scheme on how the textual diversity of image search results can be achieved through the choice of the right retrieval model is presented in [17]. [14] acknowledges the need of visual diversity in search results for image retrieval and proposes a reranking method based on topic richness analysis to enrich topic coverage in retrieval results. However, these text based techniques are not well suitable to get visually diversified summaries of image search results.

Recently, the visual diversification problem is investigated in [16]. It proposed a dynamic weighting approach to combine multiple visual features and presented three clustering methods to get diversified summarization. But the centralization and diversity are not well guaranteed and even the outlier image may be selected into the summary. Technically, the dynamic weighting for visual similarity is homogeneous for each data point, which has little ability to discriminate clusters with different compactness. Moreover, this work is lack of the solution of presenting the summary to users or utilizing the summary to help users browse the image search results. A toy example in Figs. 1(a), 1(b) and 1(c)



(a) (b) (c) (d)

Figure 1: (a), (b) and (c) show the three representative points (in red boxes), selected by folding, maxmin, and reciprocal election, respectively. (d) shows the result from our approach.

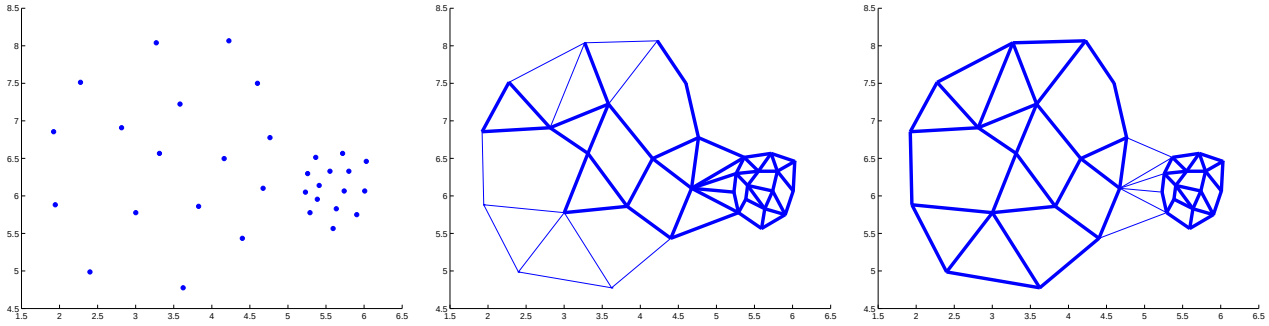
illustrates the results of those three methods, from which we can see that their obtained summaries are not satisfactory. A visual summarization approach is introduced in [3] by combining the visual similarities from different visual features, but does not address other important factors, including diversity, relevance and visual quality.

3 Visual summarization

There are several key problems of summarizing image search results, including the determination of the visual similarity measure between images, the preference selecting the images into the summary, and the good algorithm balancing both centralization (or representativeness) and diversity among the images in the summary.

Finding good visual similarity measures is a fundamental problem for image clustering. The powerful metric learning techniques are not suitable for the image clustering problem since it is an unsupervised learning problem. A dynamic feature weighting approach [16] is adopted to fuse multiple features for visual similarity evaluation, but it cannot handle the case that different groups of images have different compactness, i.e., different local scales (see Fig. 2(a) for a simple example). A locally linear reconstruction algorithm is used to learn the similarity for semi-supervised learning [18]. But the learning process is a little costly. Instead, we propose a local scaled visual similarity measure, which is helpful to cluster image search results containing image groups with each group corresponding to a different scale.

We observe that a visually appealing summary is also important to affect the satisfaction degree. The initial ranking of the search results provides an important clue on the relevances of the images,



(a) (b) (c)

Figure 2: Illustration of the local scaling scheme for similarity evaluation. (a) shows two groups of points with different densities. (b) and (c) show the similarities calculated without local scaling and with local scaling, respectively. The thickness degree of the edge indicates the degree of similarity. We can see that local scaling is useful to express the scales of each local group and get better similarity evaluation for groups of points with different local scales.

which should also be taken into consideration for generating a summary. To make use of this information, we propose to exploit them in our summary generation algorithm as the preference, so that visually high-quality images and the images initially ranked at the top have larger probability to be selected into the summary.

After defining a good similarity measure between images and calculating the preference for each image, we present an integrated approach, dynamic absorbing random walk, modeling all the issues together, to find a visually appealing, high-relevant, diverse visual summarization for the image search results.

3.1 Local scaled visual similarity

In this subsection, we propose a local scaling approach motivated by the approach in [24] to evaluate the visual similarity. Existing methods often valuate the similarity using the Gaussian function,

$$s(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}), \tag{1}$$

where σ is the scaling parameter to adjust the similarity degree. We observed that there may not be a single value for σ that works well for all the data when the input data includes clusters with different local statistics. Therefore, we propose a local scaling approach. For each data point, instead of using a

global scaling parameter, we estimate its local scaling parameter according to its distances to the other data points. Specifically, we calculate the square of distances of a data point to the others, and then compute the square root of the harmonic mean as its local scale parameter. We adopt the harmonic mean because it tends strongly toward the least elements of the distances and also tends (compared to the arithmetic mean) to mitigate the impact of large outliers and aggravate the impact of small ones.

Mathematically, the local scaling parameter is calculated as follows

$$\sigma_i^2 = \frac{n-1}{\sum_{k \neq i} \frac{1}{\max(\epsilon, d^2(\mathbf{x}_i, \mathbf{x}_k))}}, \tag{2}$$

where $\epsilon = 0.00001$ is a constant to avoid too small divisor. Then the similarity with local scaling between two data points is defined as follows

$$s(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i \sigma_j}). \tag{3}$$

In our experiment, we use several types of features, including the color, shape and texture of images, and we calculate the similarities over these features, respectively, and multiply them together to get the overall similarity on those features. Then we can get an overall affinity matrix $\mathbf{A} = [a_{ij}]_{n \times n}$, with each entry $a_{ij} = s(\mathbf{x}_i, \mathbf{x}_j)$, and n being the image number.

An example in Fig. 2 shows that the local scaled similarity is superior to the similarity using a single global scaling parameter, since it is able to capture the different scales of the clusters.

3.2 Preference

The images in the summary serve as an overview of the image search results. Hence to be good representatives, these images should be semantically relevant to the query, and also look visually satisfactory. We introduce two preferences: relevance preference and quality preference. Basically, it is expected that relevant images and high-quality images have more probabilities to be selected into the summary.

Relevance preference Given image search results, we cannot get the exact relevance degree with respect to the query because existing search engines only use the text information to return images. However, the order in the search results in a large sense reflects the relevance degree. Hence, it is reasonable to use the original order as the relevance preference. For one image in position i, we compute its relevance score as a Gaussian distribution $r_i = \exp(-\frac{i^2}{2\sigma_r^2})$, where $\sigma_r = 200$ is

a parameter to determine how many top ordered images are given higher preference. The whole relevance preference over n images is denoted as a vector $\mathbf{r} = [r_1 \cdots r_n]^T$.

Quality preference We adopt the machine learning technique to obtain visual quality evaluation function. We select a set of Web images, ask the volunteers to label the image quality. The labeled images are divided according to their quality scores into three levels: the best, middle and the worst. We use the images belonging to the best as the positive samples, and the images belonging to the worst as the negative samples. The images in the middle are viewed as ambiguous samples, and discarded in the training process. In this paper, we extract 8-dimensional features [11] from the training images, including color entropy, brightness, blur, block, dynamic range, intensity contrast, image width, and image height. These features are then used to train a support vector machine (SVM) classifier. In our experiment, we use a soft SVM classifier using the probability estimation technique [2] such that the range is $[0, 1]$ with 1 high quality and 0 low quality. Then for each image, we denote its quality score by q_i. The whole quality score is denoted as a quality vector $\mathbf{q} = [q_1 \cdots q_n]^T$. Then, the relevance and quality preferences are combined together to get a whole preference $\mathbf{p} = \alpha\mathbf{r} + (1-\alpha)\mathbf{q}$ with $\alpha = 0.5$ in our experiment. To make $\mathbf{p}$ be a distribution, we normalize it so that $\sum_i p_i = 1$.

3.3 Dynamic absorbing random walk

A good summary of image collections should have the following several properties: 1) The image is a representative of a local group in the image collection, i.e., it is very similar to many other items (centralization); 2) Those images should cover as many distinct groups as possible (diversity); And 3) it incorporates some preference as prior knowledge to representative selection (preference). It is pointed out that absorbing random walk can improve the diversity of the ranking and can also be applied to text summarization [26]. Related graph based methods are also used for video annotation [21, 22, 19, 20]. We propose a new algorithm based on the absorbing random walk, by introducing a dynamic weight tuning scheme, to improve the diversity performance for visual summarization. In the following, we first review this method, analyze its drawbacks, and modify it for visual summarization.

3.3.1 Absorbing random walk

The basic idea of exploiting absorbing random walk to encourage the diversity of a ranking list is as the following. A random walk is defined on a graph over the items. The item with the largest

stationary probability is selected and pushed into the ranking list. Once an item has been selected, it is set to be an absorbing state. The absorbing states drag down the stationary probabilities of the items close to them, thus encourage the diversity. Mathematically, this process is described as follows. A transition matrix $\tilde{\mathbf{T}} = [\tilde{t}_{ij}]_{n\times n}$ is defined by normalizing the rows of $\mathbf{A}$: $\tilde{t}_{ij} = \frac{a_{ij}}{\sum_{k=1}^{n} a_{ik}}$, so that $\tilde{t}_{ij}$ is the probability that walker moves to j from i. Then a teleporting random walk $\mathbf{T}$ is obtained by interpolating each row with the preference $\mathbf{p}$.

$$\mathbf{T} = \lambda\tilde{\mathbf{T}} + (1-\lambda)\mathbf{e}\mathbf{p}^T, \tag{4}$$

where $\mathbf{e}$ is an all-1 vector, and $\mathbf{e}\mathbf{p}^T$ is the outer product. When $0 < \lambda < 1$ and $\mathbf{p}$ does not have zero elements, this teleporting random walk $\mathbf{T}$ is irreducible, aperiodic, all states are positive recurrent and thus ergodic. Therefore, $\mathbf{T}$ has a unique stationary distribution $\boldsymbol{\pi} = \mathbf{T}^T\boldsymbol{\pi}$.

Suppose a group of items $\mathcal{G} = \{g_i\}$ have been selected, we then turn them as absorbing states by setting $t_{gg} = 1$ and $t_{gi} = 0, \forall i \neq g$. We can arrange items so that the selected items are listed before the remaining items. The transition matrix $\mathbf{T}$ is thus rewritten as

$$\mathbf{T}_{\mathcal{G}} = \begin{pmatrix} \mathbf{I}_{\mathcal{G}} & \mathbf{0} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix}. \tag{5}$$

The state with the largest expected number of visits is then selected into $\mathcal{G}$ in current iteration. The average expected visit number is calculated as

$$\mathbf{v} = \frac{\mathbf{N}^T\mathbf{e}}{n - |\mathcal{G}|}, \tag{6}$$

where $|\mathcal{G}|$ is the size of $\mathcal{G}$, and $\mathbf{N}$ is a so-called fundamental matrix

$$\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}. \tag{7}$$

3.3.2 Dynamic absorbing random walk

This absorbing random walk approach can improve the diversity in ranking, but the improvement is significantly limited when the data points distribute in such a way that the number of images in different groups are not balanced. It is not likely that all the groups are with similar image numbers in a real world scenario. Fig. 3(a), (b) and (c) shows the stationary distributions (of the toy example in Fig. 1) of the first three iterations of absorbing random walk. We can see that the first three items

selected by absorbing random walk are all from the middle group. This is because the middle group contains too many points which are close to each other, so that the influence of a few absorbing states is too weak to lower the stationary probabilities of other points in this group. To get a diverse summary, we propose a dynamic absorbing random walk, which, in addition to producing absorbing states, dynamically updates the transition matrix according to the current selected items, which we call a dynamic weight tuning scheme.

The basic idea is as follows. We dynamically adjust the transition probability between two items in the remaining items according to their similarities to the selected item. Intuitively, if two items are very similar to the selected image, we reduce the similarity between them and in turn reduce their transition probability, which will consequently reduce the probability they are selected as the representatives. Formally, we tune the transition as the following

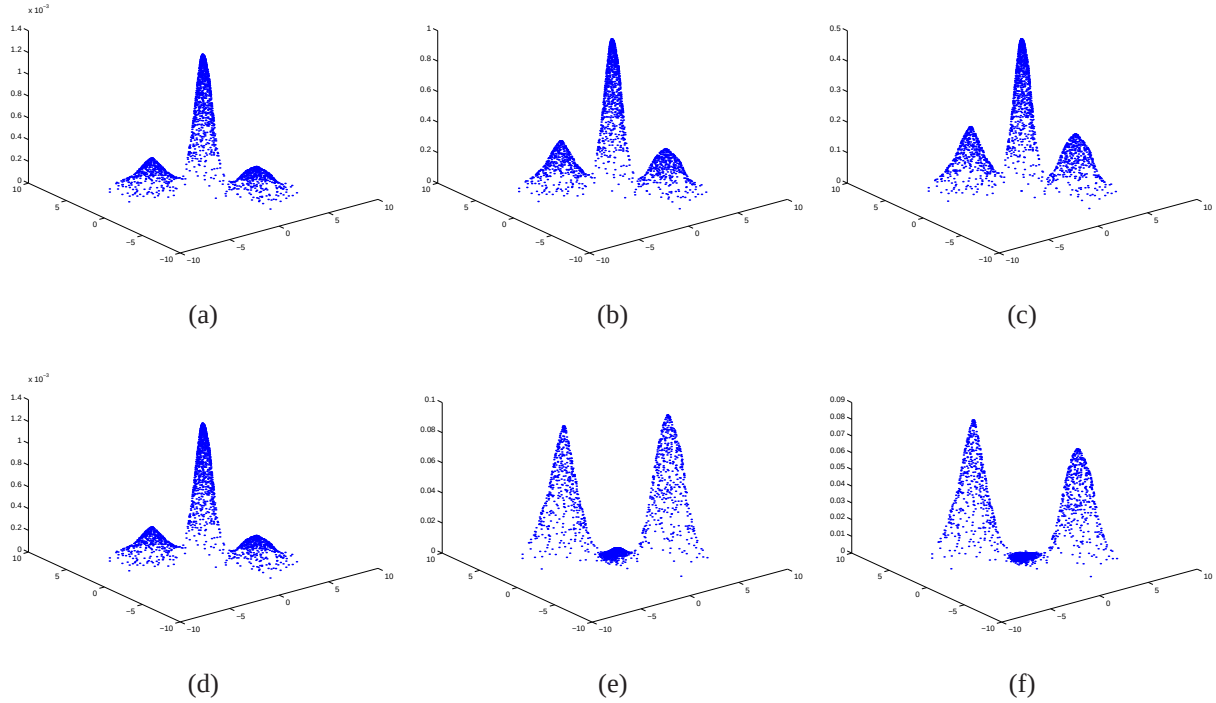
$$\tilde{t}_{jk}^{|\mathcal{G}|} = \begin{cases} \frac{t_{jk}^{|\mathcal{G}-1|}}{\exp(\rho \times t_{ji}^{0} \times t_{ki}^{0})}, & i \neq j, i \neq k \\ t_{jk}^{|\mathcal{G}-1|}, & \text{otherwise} \end{cases} \tag{8}$$

where i is the index of the item selected in the $|\mathcal{G}-1|$-th iteration, t_{ji}^{0} and t_{ki}^{0} are the initial transition probabilities from items j and k to item i, and $\rho = 2$ is a parameter to control the degree of the adjustment of the transition probabilities. We normalize each row in $\tilde{\mathbf{T}}^{|\mathcal{G}|}$ to get a transition matrix $\mathbf{T}^{|\mathcal{G}|}$. Then the sub transition matrix over the remaining images is denoted as $\mathbf{Q}^{|\mathcal{G}|}$. Similar to the absorbing random walk, we can calculate the expected number of visits of the remaining images, and select the item with the maximum expected number of visits. The toy example shown in Fig. 3 (d), (e) and (f) shows the effectiveness of our dynamic absorbing random walk. The whole algorithm is summarized as Alg.1.

Algorithm 1 Computing a diversified visual summary via dynamic absorbing random walk

1. Compute the stationary distribution $\boldsymbol{\pi}^1$ of the random walk $\mathbf{T}$, so that $\boldsymbol{\pi}^1 = \mathbf{T}^T \boldsymbol{\pi}^1$.
2. Push i to visual summary $\mathcal{S}$, with $i = \arg\max_{j=1}^{n} \pi_j^1$.
3. Set image i to be an absorbing state.
4. Update the transition matrix $\mathbf{T}^{t-1}$ to $\mathbf{T}^t$ according to Eqn. (8).
5. Calculate the average visiting number according to Eqn. (6).
6. Push i into $\mathcal{S}$ with $i = \arg\max_j v_j$.
7. Repeat step 3 to step 6 till the number of images in $\mathcal{S}$ reaches the preset number.

Complexity analysis The computation cost of the proposed summarization algorithm mainly comes



(a) (b) (c)

(d) (e) (f)

Figure 3: (a), (b) and (c) illustrate the original absorbing random walk and show the stationary distributions over the three groups of 2D points when no representative points are selected, after the first representative point is selected, and after the second representative point is selected. (d), (e) and (f) demonstrate the effectiveness of our dynamic absorbing random walk and show the stationary distributions over the three groups of 2D points when no representative points are selected, after the first representative point is selected, and after the second representative point is selected. From the figures, we can see that our approach is able to find the representative points from all the three local groups due to our dynamic weight tuning scheme while the original absorbing random walk has no such ability.

from two aspects. One aspect is the maintenance of the affinity matrix $\mathbf{A}$, and the other is the algorithm shown in Alg.1. The affinity matrix $\mathbf{A}$ is built off-line. The complexity of its computation is $O(n^2)$, with n being the number of images considered for each query. For our implementation we will only require a sparse matrix with $O(n)$ non-zero entries. For the second aspect, each iteration costs $O(n^2)$ as it conducts a matrix inversion operation of a sparse matrix in Eqn. (6). Hence, the whole computation cost is $O(kn^2)$ with k being the number of images in the summary $\mathcal{S}$. For an offline process, such computation cost is acceptable. We can also follow the implementation in `http://image-swirl.googlelabs.com/`, and build summaries only for a limited set of popular queries.

4 Interactive browsing

We have presented the approach to get a visual summary for image search results. To utilize the summarization for browsing image search results, we will categorize the remaining images by selecting each remaining image into the category that corresponds to the most similar image in the summary. To make users intuitively browse each category of images, we further do the visual summarization on each category. This process can be recursively performed until the number of images in each category is smaller than some pregiven number. This process is called hierarchical summarization. In our hierarchical summarization, each category will be divided into four sub-categories, i.e., the size of $\mathcal{S}$ in Alg.1 is set as four. With this process, an image collection is organized by a forest. The time cost is $O(n^2 \log n)$ and still acceptable.

The forest representation brings out the advantage of well organizing the image collection. But it is still insufficient, because a good visualization scheme is also essential to allow users intuitively to browse the search results. The visualization system should allow users to easily get a whole overview of the image search results, i.e., a visual summary of the image collection, and conveniently present the details according to the user target.

The naive approach to browsing the forest is a nested list view. At the beginning, only the root images of the trees in the forest are displayed, which provides a quick overview of the image collections. When the user clicks an image, then the images that are the children of this image are automatically displayed. A user can recursively click an image to explore the corresponding subtree

to find the images of interest. The disadvantage is that the view may display too many images so that the view looks too messy, which influences the overall experience when the user explores the image near the leaf node. Therefore, we will present a graph based interactive scheme to enable users to easily browse such a forest, which benefits from both the forest summarization structure and the similarity of all the images. Our browsing scheme is a unified way to combine similarity-based global visualization and local detail visualization, which enables users to know the overall relations of all the images, and at the same time to get the detail view of target images.

To obtain the global visualization, we propose to embed the image space to a 2-dimensional space, such that the visually similar images are embedded neighborly and visually dissimilar images are placed far away. We adopt the nonlinear dimensionality reduction algorithm, isometric feature mapping (ISOMAP) [15], to embed high-dimensional images into a 2-dimensional space because ISOMAP can find a 2D embedding to preserve the geodesic distance in the high dimensional space, which is calculated according to the local Euclidean distance in the original space.

Specifically, the ISOMAP algorithm is described as the following. First, a weighted K-nearest neighbor graph is constructed, with the weight on each edge corresponding to its Euclidean distance. Second, a geodesic distance matrix $\mathbf{D}_g$ is defined over the images as the sum of edge weights along the shortest path between two nodes, computed using the Dijkstra's algorithm. Third, an inner product matrix is computed by $\mathbf{K} = -\frac{1}{2}\mathbf{H}\mathbf{D}_g^{.2}\mathbf{H}$, where $\mathbf{D}_g^{.2}$ means the element-wise square of the geodesic distance matrix $\mathbf{D}_g$, and $\mathbf{H}$ is a centering matrix with $h_{ij} = \delta_{i=j} - \frac{1}{n}$. Finally, the two eigenvectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ corresponding to the two maximum eigenvalues λ_1 and λ_2 of $\mathbf{K}$ are used to form the 2D embedding. The 2D coordinate of image i is calculated as $\mathbf{y}_i = [\sqrt{\lambda_1^i}\alpha_1, \sqrt{\lambda_2^i}\alpha_2]^T$. Fig. 4 (a) shows such 2D projections.

Since we will combine this global 2D embedding with the local visualization, we record this 2D embedding in our hierarchical tree structure using the relative coordinates rather than the original absolute coordinates. Specifically, for each image, we compute a relative coordinate $\bar{\mathbf{y}}$ by assuming its parent image to be at the origin. Since the root images of the trees in the forest have no parent, we calculate its relative coordinate by subtracting their mean coordinate, $\bar{\mathbf{y}}_i = \mathbf{y}_i - \frac{1}{k}\sum_{j=1}^{k}\mathbf{y}_j$, where k is the number of root images.

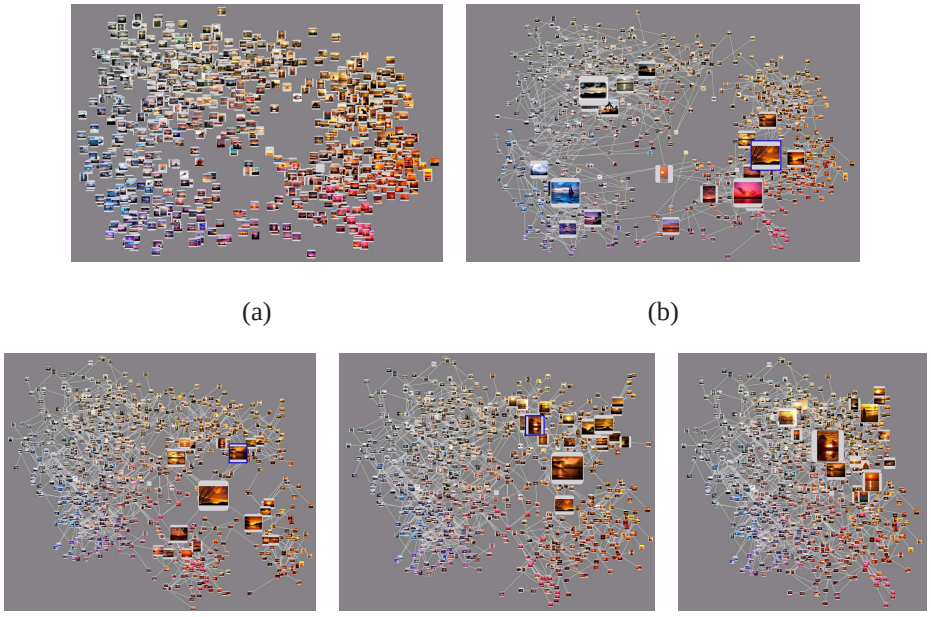
To show the local detailed images, we introduce an inhomogeneous image scale scheme. The basic idea is that the detailed image is displayed in a larger level and the associated relative coordi-

nates are also in a larger scale while the non-detailed image is displayed in a smaller level and its associated relative coordinates are also in a smaller scale. Technically, this scheme consists of two issues: detailed image determination and the scale. We solve the first issue according to the user's interactivity. At the beginning, a dummy root node is introduced to unify the forest as a tree, and this dummy node is viewed as the detailed image. In the interactivity process, the image clicked by the user, called active image, is viewed as the detailed image. For the second issue, the basic idea is to inspect the minimum path length between each image and the detailed image in the tree structure to scale and position the images. Denote the path length of one image I_i from the detailed image by l_i, and we also introduce an indicator variable b_i to show whether the image is a successor or an ancestor of the active image.

Specifically, the displaying level for image i is calculated as $z_i = 0.1 \times (10 - 2 \times l_i)$ when $l_i < 3$ if the image is a successor of the current active image, $z_i = 0.1 \times (10 - 2 \times l_i)$ when $l_i < 2$ if the image is an ancestor of the current active image, and $z_i = 0.15$ for other images. This displaying adjustment will highlight the active image and the images around it in our visual forest organization. To compute the relative coordinates of image i, we compute a scale as $s_i = -a \exp(-l_i) + b$ so that the scale is guaranteed to be between $b - a$ and b. We set $a = 15$ and $b = 20$ for the detailed images and $a = 5$ and $b = 5$ for the non-detailed images, which helps highlight the detailed images. Then we compute the coordinates of each image as $\tilde{\mathbf{y}}_i = \tilde{\mathbf{y}}_{p_i} + s_i \times \bar{\mathbf{y}}_i$. This relative distance adjustment will result in that the detailed images are displayed in a boarder area. Finally, we transform all the coordinates so that it fits the canvas view size. We show an interactive browsing process in Fig. 4.

Browsing path We present an assistant tool, browsing path, to enable the user to easily get track of his browsing path and backtrack other nodes in the forest. We introduce a dummy node as a root node to combine the trees in the forest into a single tree to show the browsing path. The input of the browsing path is the active image, i.e., the image which is clicked. In the canvas display three types of images: the active image, its children images, and the path from the image to the root.

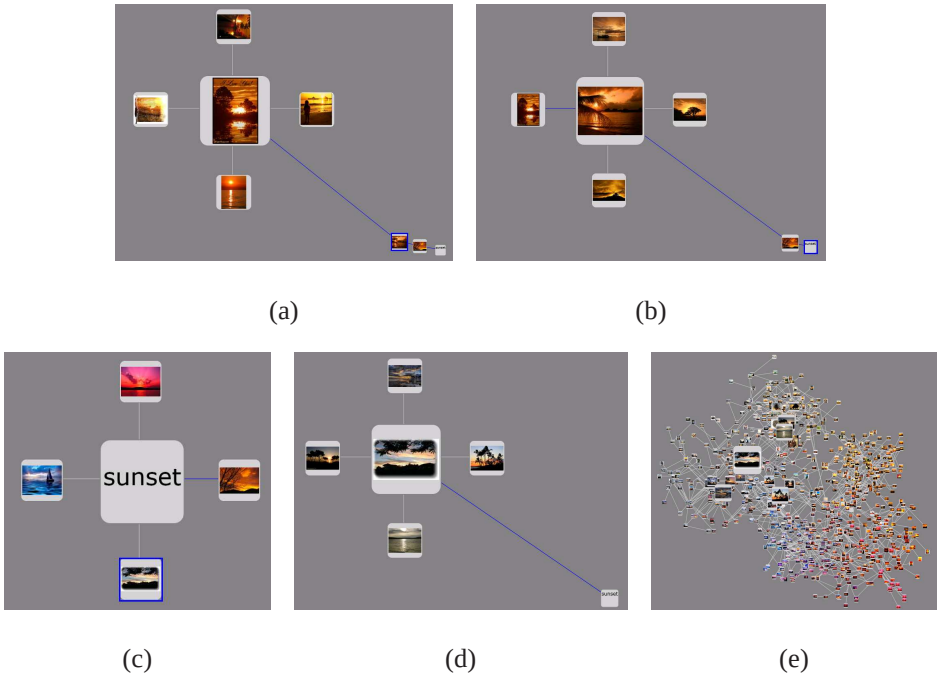
The browsing path interface also presents a simple browsing scheme by clicking an image as the active image. There are three actions according to the clicked image type. When clicking the image child, the following actions take place. Its parent is pushed into the path, and the path displaying is updated. This image is displayed in the center, and its child images are displayed. When clicking the image in the path, all the images including this image and its children are popped out from the path,



(a) (b)

(c) (d) (e)

Figure 4: Illustration of the interactive browsing scheme. (a) shows the 2D projections directly from ISOMAP. (b) - (e) show the interactive browsing process of our scheme. (b) shows the initial presentation of our scheme, (c) - (e) show the presentation scale change after a user clicks one image which is indicated by a blue bounding box.



(a) (b) (c) (d) (e)

Figure 5: Illustration of browsing path. (a) shows the browsing path corresponding to the presentation shown in Fig. 4 (e), (b) - (d) show the results of interaction on the browsing path interface, and (e) shows the presentation by switching (d) to the browsing interface.

and it is displayed in the center, and its children are displayed around this image. When clicking the active image that is being expanded, no action is performed.

During the browsing process, users can switch the browsing interface to the browsing path interface which indicates this browsing history. In the browsing path interface, the user can also switch back to the interactive browsing interface. The interactive browsing interface and the browsing path are synchronized by using the active image. With performing actions in the browsing path, consumers can easily back-browse the image search results. An example shown in Fig. 5 shows the switch results between the interactive browsing interface and the browsing path interface.

5 Experiment

To evaluate the proposed approach, we crawled top 1000 images in the returned images from Google image search for a specific text query. There are totally 59 queries, which are selected from popular queries such that those cover different types of queries, such as object, scene, portrait and etc. In the

following, we present experimental results to justify the proposed visual clustering technique and the subsequent interactive browsing schemes.

5.1 Visual summarization evaluation

Ground truth establishment To evaluate the performance quantitatively, we establish a ground truth by asking assessors to manually group the images. We distribute 59 queries to 10 users, and ask them to categorize the images based on the visual properties. The specific grouping process is as follows.

1. Overall picture: We first present the top 50 image in one page, and allow the assessor to inspect these images quickly. This helps the assessors get an overall picture of these images and get a rough idea for visual categories in these images.

2. Group forming: The assessor selects several distinct images to represent each group, and then drags the remaining images and assigns it to some representative image, and finally gets the manual grouping. During the process, the assessor may select one image to form a new group if she finds this image does not belong to any existing group.

3. Summary construction: After the groups are formed, we ask the assessor to select one image to represent each group by at the same time considering the selected representative image in other groups to guarantee the diversity.

Evaluation criteria Given the ground truth of clustering, we follow [16] to use two common used criteria, the Fowlkes-Mallows index [4] and the variation of information criterion [12], to evaluate the performance. The detailed description of these two criteria can be found in [16].

Compared methods To evaluate the performance, we compare it with several existing representative methods.

1. Affinity propagation [5] (AP). It is to find a good subset of exemplars for a whole set of data points, by considering all data points as candidate exemplars such that they can represent the image collection very well. Then a scalar message propagation scheme over the data points is derived to efficiently find the exemplars. But this method does not take into consideration diversity explicitly.

2. Affinity ranking [25] (AR). It is a reranking approach for Web search results, by optimizing two metrics: diversity – which indicates the variance of topics in a group of documents; and information richness – which measures the coverage of a single document to its topic. The two metrics are calculated from a directed link graph, named affinity graph, which models the structure of a group of documents based on the asymmetric content similarities between each pair of documents.

3. Folding [16]. It appreciates the original ranking of the search results as returned by the textual retrieval method. Images higher in the ranking list have larger probabilities being selected as cluster representatives. In a linear pass the representatives are selected, the clusters are then formed around them. But the centralization of the representatives in this method is not guaranteed.

4. Maxmin [16]. Similar to the folding approach, it also performs representative selection prior to cluster formation, but discards the original ranking and finds representatives that are visually different from each other. Similar to folding, the centralization can not be guaranteed.

5. Reciprocal election [16]. It allows all the images to cast votes for other images that they are best represented. Strong voters are then assigned to their corresponding representatives, and taken off the list of candidates. This process is repeated as long as there exist unclustered images. In this method, the diversity is not well guaranteed.

6. Absorbing random walk [26] (ARW). A ranking algorithm (GRASSHOPPER) that is similar to PageRank but encourages diversity in top ranked items, by turning already ranked items into absorbing states to penalize remaining similar items. This method is reported to outperform the K-means approach. Our approach is based on it, and improves it by defining a local scaled visual similarity measure and introducing a dynamic weight tuning scheme.

Results The comparison results are shown in Tab. 1. This quantitative comparison follows the methodology [16], and is performed on the top 50 images for each query. We found that those top 50 images often are visually of high-quality and the relevance is almost guaranteed. Hence we do not take into consideration the visual quality preference and the relevance preference for this quantitative evaluation.

Table 1: Performance comparison of our approach and six other representative approaches: affinity propagation (AP), affinity ranking (AR), folding, Maxmin, reciprocal election, and absorbing random walk (ARW). FM represents the Fowlkes-Mallows index, and the performance is better if its score is larger. VI represents the variation of information, and the performance is better if its score is smaller. Our approach gets the best performances in the two criteria. The relative increase of FM score and the relative reduction of VI score are also presented in this table.

	AP	AR	Folding	Maxmin	Reciprocal election	ARW	Our approach
FM	0.2427	0.2261	0.2533	0.2579	0.2125	0.2472	**0.2628**
Relative increase of FM	8.281%	16.23%	3.750%	1.899%	23.67%	6.311%	-
VI	2.525	2.379	2.204	2.214	2.359	2.184	**2.141**
Relative reduction of VI	15.21%	10.00%	2.858%	3.297%	9.241%	1.969%	-

In the comparison, we compute the Fowlkes-Mallows index (FM) as well as the variation of information (VI). From Tab. 1, we can observe that our approach gets the best performance. The relative increase of the FM score and the relative reduction of the VI score of our approach with respect to other approaches are also depicted in the table. The relative increase of method i with respect to method j for the FM score is calculated as $\frac{FM_i - FM_j}{FM_j}$. The relative reduction of method i with respect to method j for the VI score is calculated as $\frac{VI_j - VI_i}{VI_j}$.

The superiority of our approach over other approaches mainly comes from two aspects: the local scaled visual similarity measure and the dynamic weighting tuning scheme in absorbing random walk. The former local scaling scheme is capable to mine the visual variance for the local groups and hence gets more reliable visual similarity. The latter dynamic weight tuning scheme for absorbing random walk is capable to make the cluster so diverse that it can cover distinct local groups of images.

Justification of local scaling and dynamic tuning We present the performance comparison to illustrate the affects of local scaling for visual similarity evaluation and dynamic weight tuning in dynamic absorbing random walk. We report the the Fowlkes-Mallows index and the variation of information score for four combinations of the two schemes in Tab. 2 and Tab. 3, respectively. From the two tables, we can see that the proposed two schemes, local scaling and dynamic tuning, can improve the performance a lot.

Table 2: The effects of local scaling and dynamic tuning for the Fowlkes-Mallows index. The best score is highlighted in bold fonts.

	without local scaling	with local scaling
without dynamic tuning	0.2233	0.2472
with dynamic tuning	0.2570	**0.2628**

Table 3: The effects of local scaling and dynamic tuning for variation of information. The best score is highlighted in bold fonts.

	without local scaling	with local scaling
without dynamic tuning	2.319	2.184
with dynamic tuning	2.207	**2.141**

5.2 Interactive browsing evaluation

To evaluate the presentation schemes proposed in this paper, we perform a user study experiment since it is not easy to find a quantitative criterion to measure the performance. We recruit 10 volunteers, students from university campus and our research lab, to take part in the user study. Their grades vary from freshman to graduate grade 3. Their ages range from 19 to 24. All participants are Web image search engine users.

Procedure Users are asked to mainly focus on two aspects: the ability to present a good overview of the entire image search results, and the convenience to access the target images from users' point of view. The former aspect is related to the summarization. We have presented the quantitative evaluation for each query over top 50 images. To evaluate the summarization on about 1000 images, it is not practical to allow users to label the ground truth because manual construction of ground truth for 1000 images is very difficult. So we conduct user study to evaluate the summarization over 1000 images. We first allow users to see the whole image search results for each query, then present the summaries, the initial view of our interactive browsing scheme, without indicating which method each result corresponds to, next ask them to give each of these summaries a score ranging from 1 (the worst) to 5 (the best) by jointly considering the relevances, diversity, and the representativeness (centralization of each group) of image search results. During their evaluation, the user can return to image search results to check the representativeness performance.

The latter aspect mainly focuses on the browsing performance. We first introduce the users about

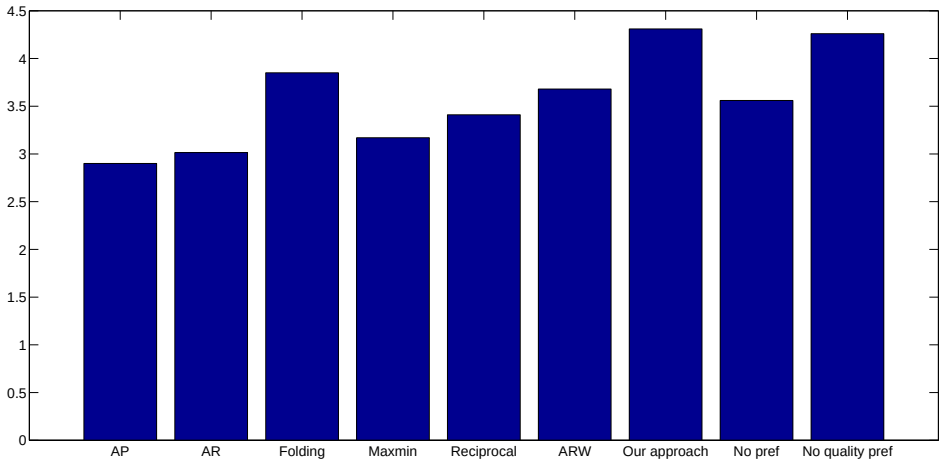
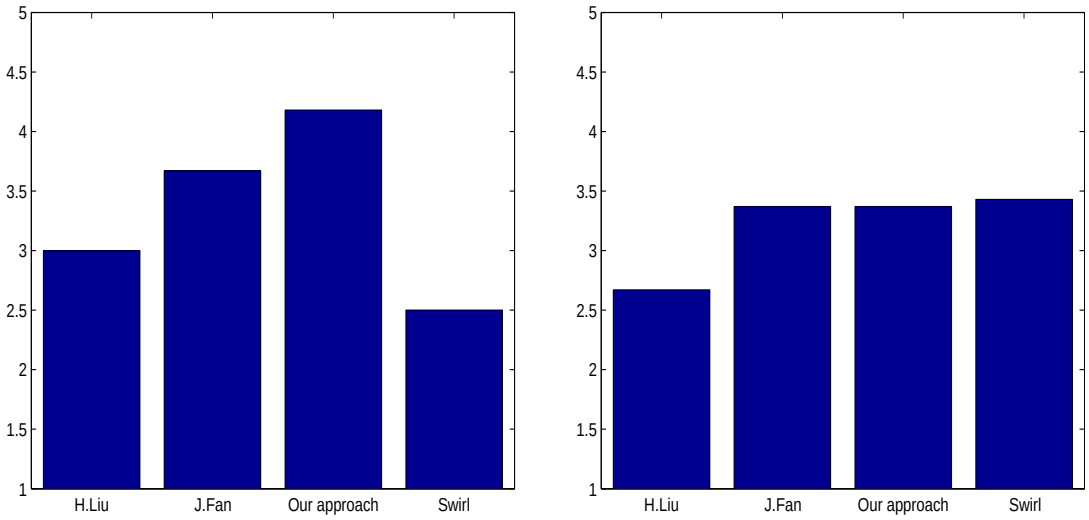


Figure 6: The average evaluation scores of user study for visual summarization over the whole 1000 images for all the queries for the 7 methods. We can see that our approach gets the best score.

two schemes: interactive browsing and browsing path, show how to use them, and tell them the meaning of the operations. We randomly divide 59 queries into two parts: 10 for training and the remaining 49 for testing. The training part is used for users to get familiar with the usage of our system. The remaining testing part is used for users to test our system. The scores by users are also ranging from 1 (the worst) to 5 (the best).

Results The average score over all the queries, about the judgement of the capability that the summary represents the whole image search results, is shown in Fig. 6. We also report other related summarization algorithms. From this figure, we can observe that our approach performs the best in all the 7 methods. This is because our approach takes into account the diversity of the representative images and the local scaling visual similarity scheme. In addition, we also conducted experiments to check the effect of the preference to the performance. From Fig. 6, we have two observations. On the one hand, the quality preference can result in a little better performance since it only affects a few queries. On the other hand, the relevance preference is important to get satisfactory performance for 1000 search results, which is also justified in the two approaches, Folding and Maxmin.

For the second aspect, we also report the results from three other representative works with our approach. The first one is the similarity based approach in [10] (denoted as H. Liu), which yields a global view of the whole image collection without giving local detailed views. The second one is the recently proposed approach in [3] (denoted as J. Fan), which is based on topic network and representativeness-based sampling but does not consider the multiple-representative case in the image



(a) (b)

Figure 7: User study for browsing image search results. (a) for the global overview for the whole image search results, and (b) for the local detail for finding target images.

collection. The last one is recently released in Google Image Swirl [9], available from `http://image-swirl.googlelabs.com/`, which is similar to our browsing path, with more beautiful interfaces, but lack of the global overview of search results. The average results are depicted in Fig. 7. We can see that our approach gets the best performance. The investigation from the users shows that two key aspects are important for browsing image search results. The first one is an overview for a group of image search results, which enables users to quickly get the whole picture. The second one is the local detailed view of image search results, which enables users to easily and intuitively find the target images. For this aspect, Google image Swirl is a little better than ours only because the interface looks better.

6 Conclusion

In this paper, we present an interactive browse scheme for image search results to help users to conveniently look through them. This scheme is based on the proposed visual summarization approach, which has the following properties: It can well represent the image search results; The images in the summary are as diverse as possible; And the summarization is visually appealing because the relevance and visual quality are also taken into consideration. The quantitative evaluation results

and user study demonstrate the powerfulness of the proposed visual summarization approach and the interactive browsing scheme.

References

[1] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of WWW image search results using visual, textual and link information. In *ACM Multimedia*, pages 952–959, 2004.

[2] C.-C. Chang and C.-J. Lin. LIBSVM – a library for support vector machines. 2009.

[3] J. Fan, Y. Gao, H. Luo, D. A. Keim, and Z. Li. A novel approach to enable semantic and visual image summarization for exploratory image search. In *Multimedia Information Retrieval*, pages 358–365, 2008.

[4] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78:553–569, 1983.

[5] B. J. Frey and D. Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315:972–976, February 2007.

[6] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q. Cheng, and W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *ACM Multimedia*, pages 112–121, 2005.

[7] Y. Jia, J. Wang, C. Zhang, and X.-S. Hua. Finding image exemplars using fast sparse affinity propagation. In *ACM Multimedia*, pages 639–642, 2008.

[8] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W.-Y. Ma. Igroup: web image search results clustering. In *ACM Multimedia*, pages 377–384, 2006.

[9] Y. Jing, H. A. Rowley, C. Rosenberg, J. Wang, and M. Covell. Visualizing Web images via Google Image Swirl. In *NIPS Workshop on Statistical Machine Learning for Visual Analytics*, 2009.

[10] H. Liu, X. Xie, X. Tang, Z. Li, and W.-Y. Ma. Effective browsing of web image search results. In *Multimedia Information Retrieval*, pages 84–90, 2004.

[11] T. Mei, X.-S. Hua, C.-Z. Zhu, H.-Q. Zhou, and S. Li. Home video visual quality assessment with spatiotemporal factors. *IEEE Trans. Circuits Syst. Video Techn.*, 17(6):699–706, 2007.

[12] M. Meilă. Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98(5):873–895, 2007.

[13] P.-A. Moëllic, J.-E. Haugeard, and G. Pitel. Image clustering based on a shared nearest neighbors approach for tagged collections. In *CIVR*, pages 269–278, 2008.

[14] K. Song, Y. Tian, W. Gao, and T. Huang. Diversifying the image retrieval results. In *ACM Multimedia*, pages 707–710, 2006.

[15] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 5500:2319–2323, 22 December 2000.

[16] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *WWW*, pages 341–350, 2009.

[17] R. van Zwol, V. Murdock, L. G. Pueyo, and G. Ramírez. Diversifying image search with user generated content. In *Multimedia Information Retrieval*, pages 67–74, 2008.

[18] J. Wang, F. Wang, C. Zhang, H. C. Shen, and L. Quan. Linear neighborhood propagation and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(9):1600–1615, 2009.

[19] J. Wang, Y. Zhao, X. Wu, and X.-S. Hua. Transductive multi-label learning for video concept detection. In *Multimedia Information Retrieval*, pages 298–304, 2008.

[20] J. Wang, Y. Zhao, X. Wu, and X.-S. Hua. A transductive multi-label learning approach for video concept detection. *Pattern Recognition*, to appear, 2010.

[21] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song. Unified video annotation via multi-graph learning. *IEEE Trans. Circuits Syst. Video Techn.*, 19(5):733–746, 2009.

[22] M. Wang, X.-S. Hua, J. Tang, and R. Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Transactions on Multimedia*, 11(3):465–476, 2009.

[23] K. Q. Weinberger, M. Slaney, and R. van Zwol. Resolving tag ambiguity. In *ACM Multimedia*, pages 111–120, 2008.

[24] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, 2004.

[25] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. In *SIGIR*, pages 504–511, 2005.

[26] X. Zhu, A. Goldberg, J. V. Gael, and D. Andrzejewski. Improving diversity in ranking using absorbing random walks. In *The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2007.