IntentSearch: Capturing User Intention for One-Click Internet Image Search

Xiaoou Tang, *Fellow*, *IEEE*, Ke Liu, Jingyu Cui, *Student Member*, *IEEE*,
Fang Wen, *Member*, *IEEE*, and Xiaogang Wang, *Member*, *IEEE*

Abstract—Web-scale image search engines (e.g., Google image search, Bing image search) mostly rely on surrounding text features. It is difficult for them to interpret users' search intention only by query keywords and this leads to ambiguous and noisy search results which are far from satisfactory. It is important to use visual information in order to solve the ambiguity in text-based image retrieval. In this paper, we propose a novel Internet image search approach. It only requires the user to click on one query image with minimum effort and images from a pool retrieved by text-based search are reranked based on both visual and textual content. Our key contribution is to capture the users' search intention from this one-click query image in four steps. 1) The query image is categorized into one of the predefined adaptive weight categories which reflect users' search intention at a coarse level. Inside each category, a specific weight schema is used to combine visual features adaptive to this kind of image to better rerank the text-based search result. 2) Based on the visual content of the query image selected by the user and through image clustering, query keywords are expanded to capture user intention. 3) Expanded keywords are used to enlarge the image pool to contain more relevant images. 4) Expanded keywords are also used to expand the query image to multiple positive visual examples from which new query specific visual and textual similarity metrics are learned to further improve content-based image reranking. All these steps are automatic, without extra effort from the user. This is critically important for any commercial web-based image search engine, where the user interface has to be extremely simple. Besides this key contribution, a set of visual features which are both effective and efficient in Internet image search are designed. Experimental evaluation shows that our approach significantly improves the precision of top-ranked images and also the user experience.

Index Terms—Image search, intention, image reranking, adaptive similarity, keyword expansion.

1 Introduction

Many commercial Internet scale image search engines use only keywords as queries. Users type query keywords in the hope of finding a certain type of images. The search engine returns thousands of images ranked by the keywords extracted from the surrounding text. It is well known that text-based image search suffers from the ambiguity of query keywords. The keywords provided by users tend to be short. For example, the average query length of the top 1,000 queries of Picsearch is 1.368 words, and 97 percent of them contain only one or two words [1]. They cannot describe the content of images accurately. The search results are noisy and consist of images with quite different semantic meanings. Fig. 1 shows the top ranked images from Bing image search using "apple" as query. They belong to different categories, such as "green apple," "red apple," "apple logo," and "iphone" because of the ambiguity of the word "apple." The ambiguity issue occurs for several reasons. First, the query keywords' meanings may be richer than users' expectations. For example, the meanings of the word "apple" include apple fruit, apple computer, and apple ipod. Second, the user may not have enough knowledge on the textual description of target images. For example, if users do not know "gloomy bear" as the name of a cartoon character (shown in Fig. 2a) and they have to input "bear" as query to search images of "gloomy bear." Lastly and most importantly, in many cases it is hard for users to describe the visual content of target images using keywords accurately.

In order to solve the ambiguity, additional information has to be used to capture users' search intention. One way is text-based keyword expansion, making the textual description of the query more detailed. Existing linguistically-related methods find either synonyms or other linguistic-related words from thesaurus, or find words frequently co-occurring with the query keywords. For example, Google image search provides the "Related Searches" feature to suggest likely keyword expansions. However, even with the same query keywords, the intention of users can be highly diverse and cannot be accurately captured by these expansions. As shown in Fig. 2b, "gloomy bear" is not among the keyword expansions suggested by Google "related searches."

Another way is content-based image retrieval with relevance feedback. Users label multiple positive and

- *X. Tang and K. Liu are with the Department of Information Engineering, the Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: xtang@ie.cuhk.edu.hk, liuke87@gmail.com.*
- *J. Cui is with the Department of Electrical Engineering, Stanford University, 87 Hulme, Apt 718, Stanford, CA 94305. E-mail: jycui@stanford.edu.*
- *F. Wen is with Microsoft Research Asia, Building 2, No. 5, Danling Street, Haidian District, Beijing 100080, P.R. China. E-mail: fangwen@microsoft.com.*
- *X. Wang is with the Department of Electronic Engineering, the Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: xgwang@ee.cuhk.edu.hk.*

Manuscript received 1 July 2010; revised 11 Apr. 2011; accepted 19 Oct. 2011; published online 8 Dec. 2011.
Recommended for acceptance by A. Leonardis.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-07-0496.
Digital Object Identifier no. 10.1109/TPAMI.2011.242.

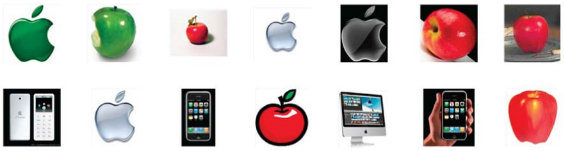
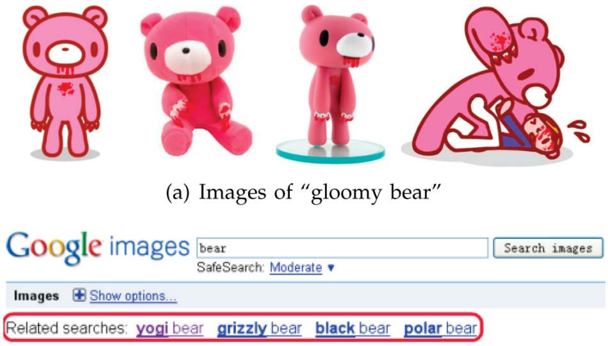


Fig. 1. Top-ranked images returned from Bing image search using "apple" as query.



(a) Images of "gloomy bear"

(b) Google Related Searches of query "bear".

Fig. 2. (a) Images of "gloomy bear." (b) Google related searches of query "bear."

negative image examples. A query-specific visual similarity metric is learned from the selected examples and used to rank images. The requirement of more users' effort makes it unsuitable for web-scale commercial systems like Bing image search and Google image search in which users' feedback has to be minimized.

We do believe that adding visual information to image search is important. However, the interaction has to be as simple as possible. The absolute minimum is One-Click. In this paper, we propose a novel Internet image search approach. It requires the user to give only one click on a query image and images from a pool retrieved by text-based search are reranked based on their visual and textual similarities to the query image. We believe that users will tolerate one-click interaction, which has been used by many popular text-based search engines. For example, Google requires a user to select a suggested textual query expansion by one-click to get additional results. The key problem to be solved in this paper is how to capture user intention from this one-click query image. Four steps are proposed as follows:

1. **Adaptive similarity.** We design a set of visual features to describe different aspects of images. How to integrate various visual features to compute the similarities between the query image and other images is an important problem. In this paper, an *Adaptive Similarity* is proposed, motivated by the idea that a user always has specific intention when submitting a query image. For example, if the user submits a picture with a big face in the middle, most probably he/she wants images with similar faces and using face-related features is more appropriate. In our approach, the query image is first categorized into one of the predefined adaptive weight categories, such as "portrait" and "scenery." Inside each category, a specific pretrained weight schema is used to combine visual features adapting to this kind of images to better rerank the text-based search result. This correspondence between the query image and its proper similarity measurement reflects the user intention. This initial reranking result is not good enough and will be improved by the following steps.
2. **Keyword expansion.** Query keywords input by users tend to be short and some important keywords may be missed because of users' lack of knowledge on the textual description of target images. In our approach, query keywords are expanded to capture users' search intention, inferred from the visual content of query images, which are not considered in traditional keyword expansion approaches. A word w is suggested as an expansion of the query if a cluster of images are visually similar to the query image and all contain the same word w.[1] The expanded keywords better capture users' search intention since the consistency of both visual content and textual description is ensured.
3. **Image pool expansion.** The image pool retrieved by text-based search accommodates images with a large variety of semantic meanings and the number of images related to the query image is small. In this case, reranking images in the pool is not very effective. Thus, more accurate query by keywords is needed to narrow the intention and retrieve more relevant images. A naive way is to ask the user to click on one of the suggested keywords given by traditional approaches only using text information and to expand query results like in Google "related searches." This increases users' burden. Moreover, the suggested keywords based on text information only are not accurate to describe users' intention. Keyword expansions suggested by our approach using both visual and textual information better capture users' intention. They are automatically added into the text query and enlarge the image pool to include more relevant images. Feedback from users is not required. Our experiments show that it significantly improves the precision of top ranked images.
4. **Visual query expansion.** One query image is not diverse enough to capture the user's intention. In Step 2, a cluster of images all containing the same expanded keywords and visually similar to the query image are found. They are selected as expanded positive examples to learn visual and textual similarity metrics, which are more robust and more specific to the query, for image reranking. Compared with the weight schema in Step 1, these similarity metrics reflect users' intention at a finer level since every query image has different metrics. Different from relevance feedback, this visual expansion does not require users' feedback.

All four of these steps are automatic with only one click in the first step without increasing users' burden. This makes it possible for Internet scale image search by both textual and visual content with a very simple user interface.

[1] The word w does not have to be contained by the query image.

Our one-click intentional modeling in Step 1 has been proven successful in industrial applications [2], [3] and is now used in the Bing image search engine [4].[2] This work extends the approach with Steps 2-4 to further improve the performance greatly.

2 Related Work

2.1 Image Search and Visual Expansion

Many Internet scale image search methods [5], [6], [7], [8], [9] are text-based and are limited by the fact that query keywords cannot describe image content accurately. Content-based image retrieval [10] uses visual features to evaluate image similarity. Many visual features [11], [12], [13], [14], [15], [16], [17] were developed for image search in recent years. Some were global features, such as GIST [11] and HOG [12]. Some quantized local features, such as SIFT [13], into visual words, and represented images as bags-of-visual-words (BoV) [14]. In order to preserve the geometry of visual words, spatial information was encoded into the BoV model in multiple ways. For example, Zhang et al. [17] proposed geometry-preserving visual phases which captured the local and long-range spatial layouts of visual words.

One of the major challenges of content-based image retrieval is to learn the visual similarities which reflect the semantic relevance of images well. Image similarities can be learned from a large training set where the relevance of pairs of images is known [18]. Deng et al. [19] learned visual similarities from a hierarchical structure defined on semantic attributes of training images. Since web images are highly diversified, defining a set of attributes with hierarchical relationships for them is challenging. In general, learning a universal visual similarity metric for generic images is still an open problem to be solved.

Some visual features may be more effective for certain query images than others. In order to make the visual similarity metrics more specific to the query, relevance feedback [20], [21], [22], [23], [24], [25], [26] was widely used to expand visual examples. The user was asked to select multiple relevant and irrelevant image examples from the image pool. A query-specific similarity metric was learned from the selected examples. For example, in [20], [21], [22], [24], and [25], discriminative models were learned from the examples labeled by users using support vector machines or boosting, and classified the relevant and irrelevant images. In [26], the weights of combining different types of features were adjusted according to users' feedback. Since the number of user-labeled images is small for supervised learning methods, Huang et al. [27] proposed probabilistic hypergraph ranking under the semi-supervised learning framework. It utilized both labeled and unlabeled images in the learning procedure. Relevance feedback required more users' effort. For a web-scale commercial system, users' feedback has to be limited to the minimum, such as one-click feedback.

In order to reduce users' burden, pseudorelevance feedback [28], [29] expanded the query image by taking the top N images visually most similar to the query image as positive examples. However, due to the well-known semantic gap, the top N images may not be all semantically consistent with the query image. This may reduce the performance of pseudorelevance feedback. Chum et al. [30] used RANSAC to verify the spatial configurations of local visual features and to purify the expanded image examples. However, it was only applicable to object retrieval. It required users to draw the image region of the object to be retrieved and assumed that relevant images contained the same object. Under the framework of pseudorelevance feedback, Ah-Pine et al. [31] proposed transmedia similarities which combined both textual and visual features. Krapac et al. [32] proposed the query-relative classifiers, which combined visual and textual information, to rerank images retrieved by an initial text-only search. However, since users were not required to select query images, the users' intention could not be accurately captured when the semantic meanings of the query keywords had large diversity.

We conducted the first study that combines text and image content for image search directly on the Internet in [33], where simple visual features and clustering algorithms were used to demonstrate the great potential of such an approach. Following our intent image search work in [2] and [3], a visual query suggestion method is developed in [34]. Its difference from [2] and [3] is that instead of asking the user to click on a query image for reranking, the system asks users to click on a list of keyword-image pairs generated offline using a data set from Flickr and search images on the web based on the selected keyword. The problem with this approach is that, on one hand, the data set from Flickr is too small compared with the entire Internet and thus cannot cover the unlimited possibility of Internet images and, on the other hand, the keyword-image suggestions for any input query are generated from the millions of images of the whole data set and thus are expensive to compute and may produce a large number of unrelated keyword-image pairs.

Besides visual query expansion, some approaches [35], [36] used concept-based query expansions through mapping textual query keywords or visual query examples to high-level semantic concepts. They needed a predefined concept lexicons whose detectors were offline learned from fixed training sets. These approaches were suitable for closed databases but not for web-based image search, since the limited number of concepts cannot cover the numerous images on the Internet. The idea of learning example specific visual similarity metric was explored in previous work [37], [38]. However, they required training a specific visual similarity for every example in the image pool, which is assumed to be fixed. This is impractical in our application where the image pool returned by text-based search constantly changes for different query keywords. Moreover, text information, which can significantly improve visual similarity learning, was not considered in previous work.

2.2 Keyword Expansion

In our approach, keyword expansion is used to expand the retrieved image pool and to expand positive examples. Keyword expansion was mainly used in document retrieval. Thesaurus-based methods [39], [40] expanded query keywords with their linguistically related words such as synonyms and hypernyms. Corpus-based methods, such as

[2] The "similar images" function of http://www.bing.com/images/.

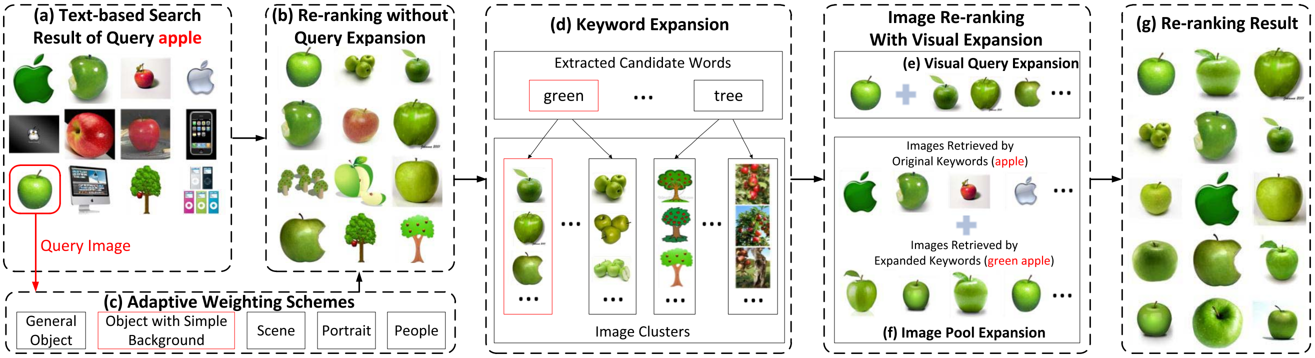

(a) Text-based Search Result of Query apple
Query Image
(b) Re-ranking without Query Expansion
(c) Adaptive Weighting Schemes
General Object
Object with Simple Background
Scene
Portrait
People
(d) Keyword Expansion
Extracted Candidate Words
green
...
tree
Image Clusters
Image Re-ranking With Visual Expansion
(e) Visual Query Expansion
Images Retrieved by Original Keywords (apple)
Images Retrieved by Expanded Keywords (green apple)
(f) Image Pool Expansion
(g) Re-ranking Result

Fig. 3. An example to illustrate our algorithm. The details of steps (c)-(f) are given in Sections 3.3, 3.4, 3.5, and 3.6.

well-known term clustering [41] and Latent Semantic Indexing [42], measured the similarity of words based on their co-occurrences in documents. Words most similar to the query keywords were chosen as textual query expansion. Some image search engines have the feature of expanded keywords suggestion. They mostly use surrounding text.

Some algorithms [43], [44] generated tag suggestions or annotations based on visual content for input images. Their goal is not to improve the performance of image reranking. Although they can be viewed as options of keyword expansions, some difficulties prevent them from being directly applied to our problem. Most of them assumed fixed keyword sets, which are hard to obtain for image reranking in the open and dynamic web environment. Some annotation methods required supervised training, which is also difficult for our problem. Different than image annotation, our method provides extra image clusters during the procedure of keyword expansions, and such image clusters can be used as visual expansions to further improve the performance of image reranking.

3 Method

3.1 Overview

The flowchart of our approach is shown in Fig. 3. The user first submits query keywords q. A pool of images is retrieved by text-based search[3] (Fig. 3a). Then the user is asked to select a query image from the image pool. The query image is classified as one of the predefined adaptive weight categories. Images in the pool are reranked (Fig. 3b) based on their visual similarities to the query image and the similarities are computed using the weight schema (Fig. 3c described in Section 3.3) specified by the category to combine visual features (Section 3.2).

In the keyword expansion step (Fig. 3d described in Section 3.4), words are extracted from the textual descriptions (such as image file names and surrounding texts in the html pages) of the top k images most similar to the query image, and the *tf-idf* method [45] is used to rank these words. To save computational cost, only the top m words are reserved as candidates for further processing. However, because the initial image reranking result is still ambiguous and noisy, the top k images may have a large diversity of semantic meanings and cannot be used as visual query expansion. The word with the highest tf-idf score computed from the top k images is not reliable to be chosen as keyword expansion either. In our approach, reliable keyword expansions are found through further image clustering. For each candidate word w_i, we find all the images containing w_i and group them into different clusters $\{c_{i,1}, c_{i,2}, \ldots, c_{i,t_i}\}$ based on visual content. As shown in Fig. 3d, images with the same candidate word may have a large diversity in visual content. Images assigned to the same cluster have higher semantic consistency since they have high visual similarity to one another and contain the same candidate word. Among all the clusters of different candidate words, cluster $c_{i,j}$ with the largest visual similarity to the query image is selected as visual query expansion (Fig. 3d, described in Section 3.5), and its corresponding word w_i is selected to form keyword expansion $q' = q + w_i$.

A query specific visual similarity metric (Section 3.5) and a query specific textual similarity metric (Section 3.7) are learned from both the query image and the visual query expansion. The image pool is enlarged through combining the original image pool retrieved by the query keywords q provided by the user and an additional image pool retrieved by the expanded keywords q' (Fig. 3f, described in Section 3.6). Images in the enlarged pool are reranked using the learned query-specific visual and textual similarity metrics (Fig. 3g). The size of the image cluster selected as visual query expansion and its similarity to the query image indicate the confidence that the expansion captures the user's search intention. If they are below certain thresholds, expansion is not used in image reranking.

3.2 Visual Feature Design

We design and adopt a set of features that are both effective in describing the visual content of images from different aspects, and efficient in their computational and storage complexity. Some of them are existing features proposed in recent years. Some new features are first proposed by us or extensions of existing features. It takes an average of 0.01 ms to compute the similarity between two features on a machine of 3.0 GHz CPU. The total space to store all features for an image is 12 KB. More advanced visual features developed in recent years or in the future can also be incorporated into our framework.

3.2.1 Existing Features

- **Gist.** Gist [11] characterizes the holistic appearance of an image, and works well for scenery images.
- **SIFT.** We adopt 128-dimension SIFT [13] to describe regions around Harris interest points. SIFT

[3] In this paper, it is from Bing image search [4].

descriptors are quantized according to a codebook of 450 words.

- **Daubechies Wavelet.** We use the second-order moments of wavelet coefficients in various frequency bands (DWave) to characterize the texture properties in the image [46].
- **Histogram of Gradient (HoG).** HoG [12] reflects distributions of edges over different parts of an image, and is especially effective for images with strong long edges.

3.2.2 New Features

- **Attention Guided Color Signature.** Color signature [47] describes the color composition of an image. After clustering colors of pixels in the LAB color space, cluster centers and their relative proportions are taken as the signature. We propose a new Attention Guided Color Signature (ASig) as a color signature that accounts for varying importance of different parts of an image. We use an attention detector [48] to compute a saliency map for the image, and then perform k-Means clustering weighted by this map. The distance between two ASigs can be calculated efficiently using the Earth Mover Distance algorithm [47].
- **Color Spatialet (CSpa).** We design a novel feature, Color Spatialet, to characterize the spatial distribution of colors. An image is divided into $n \times n$ patches. Within each patch, we calculate its main color as the largest cluster after k-Means clustering. The image is characterized by CSpa, a vector of n^2 color values. In our experiments, we take $n = 9$. We account for some spatial shifting and resizing of objects in the images when calculating the distance of two CSpas A and B:?rlv

$$d(A, B) = \sum_{i=1}^{n} \sum_{j=1}^{n} \min[d(A_{i,j}, B_{i\pm1,j\pm1})],$$

where $A_{i,j}$ denotes the main color of the (i, j)th block in the image. Color Spatialet describes color spatial configuration. By capturing only the main color, it is robust to slight color changes due to lighting, white balance, and imaging noise. Since local shift is allowed when calculating distance between two Color Spatialets, the feature is shift invariant to some extent and has resilience to misalignment.

- **Multilayer Rotation Invariant (EOH).** Edge Orientation Histogram [49] describes the histogram of edge orientations. We incorporate rotation invariance when comparing two EOHs, rotating one of them to best match the other. This results in a Multilayer Rotation Invariant EOH (MRI-EOH). Also, when calculating MRI-EOH, a threshold parameter is required to filter out the weak edges. We use multiple thresholds to get multiple EOHs to characterize image edge distributions at different scales.
- **Facial Feature.** Face existence and their appearances give clear semantic interpretations of the image. We apply face detection algorithm [50] to each image, and obtain the number of faces, face sizes, and positions as features to describe the image from a "facial" perspective.

3.3 Adaptive Weight Schema

Humans can easily categorize images into high-level semantic classes, such as scene, people, or object. We observed that images inside these categories usually agree on the relative importance of features for similarity calculations. Inspired by this observation, we assign the query images into several typical categories, and adaptively adjust feature weights within each category.

Suppose an image i from query category Q_q is characterized using F visual features, the adaptive similarity between image i and j is defined as $s^q(i, j) = \sum_{m=1}^{F} \alpha_m^q s_m(i, j)$, where $s_m(i, j)$ is the similarity between image i and j on feature m, and α_m^q expresses the importance of feature m for measuring similarities for query images from category Q_q. We further constrain $\alpha_m^q \geq 0$ and $\sum_m \alpha_m^q = 1$.

3.3.1 Query Categorization

The query categories we considered are: General Object, Object with Simple Background, Scenery Images, Portrait, and People. We use 500 manually labeled images, 100 for each category, to train a C4.5 decision tree for query categorization.

The features we used for query categorization are: existence of faces, the number of faces in the image, the percentage of the image frame taken up by the face region, the coordinate of the face center relative to the center of the image, Directionality (Kurtosis of Edge Orientation Histogram, Section 3.2), Color Spatial Homogeneousness (variance of values in different blocks of Color Spatialet, Section 3.2), total energy of edge map obtained from Canny operator, and Edge Spatial Distribution (the variance of edge energy in a 3×3 regular block of the image, characterizing whether edge energy is mainly distributed at the image center).

3.3.2 Feature Fusion

In each query category Q_q, we pretrain a set of optimal weights α_m^q based on the RankBoost framework [51]. For a query image i, a real-valued feedback function $\Phi_i(j, k)$ is defined to denote preference between image j and k. We set $\Phi_i(j, k) > 0$ if image k should be ranked above image j, and 0 otherwise. The feedback function Φ_i induces a distribution over all pairs of images:

$$(j, k) : D_i(j, k) = \frac{\Phi_i(j, k)}{\sum_{j,k} \Phi_i(j, k)},$$

and the ranking loss for query image i using similarity measurement $s^q(i, \cdot)$ is $L_i = \Pr_{(j,k)\sim D_i}[s^q(i, k) \leq s^q(i, j)]$.

We use an adaptation of the RankBoost framework (Algorithm 1) to minimize the loss function with respect to α_m^q, which is the optimal weight schema for category Q_q. Steps 2, 3, 4, and 8 differ from the original RankBoost algorithm to accommodate to our application, and are detailed as follows:

Algorithm 1. Feature weight learning for a certain query category

```
1. Input: Initial weight $D_i$ for all query images $i$ in the
   current intention category $Q_q$, similarity matrices $s_m(i, \cdot)$
   for all query image $i$ and feature $m$;
```

```
2. Initialize: Set step $t = 1$, set $D_i^1 = D_i$ for all $i$;
while not converged do
    for each query image $i \in Q_q$ do
        3. Select best feature $m_t$ and the corresponding similarity $s_{m_t}(i, \cdot)$ for current re-ranking problem under weight $D_i^t$;
        4. Calculate ensemble weight $\alpha_t$ according to Equation 1;
        5. Adjust weight $D_i^{t+1}(j,k) \propto D_i^t(j,k)\exp\{\alpha_t[s_{m_t}(i,j) - s_{m_t}(i,k)]\}$;
        6. Normalize $D_i^{t+1}$ to make it a distribution;
        7. $t{+}{+}$;
    end for
end while
8. Output: Final optimal similarity measure for current intention category: $s^q(\cdot,\cdot) = \frac{\sum_t \alpha_t s_{m_t}(\cdot,\cdot)}{\sum_t \alpha_t}$, and the weight for feature $m$: $a_m^q = \frac{\sum_{m_t=m}\alpha_t}{\sum_t \alpha_t}$.
```

Step 2: Initialization. The training images are categorized into the five main classes. Images within each main class are further categorized into subclasses. Images in each subclass are visually similar. Besides, a few images are labeled as noise (irrelevant images) or neglect (hard to judge relevancy). Given a query image i, we define four image sets: $S1_i$ includes images within the same subclass as i, $S2_i$ includes images within the same main class as i, excluding those in $S1_i$, $S3_i$ includes images labeled as "neglect," and $S4_i$ includes images labeled as "noise." For any image $j \in S1_i$ and any image $k \in S2_i \cup S4_i$, we set $\Phi(k,j) = 1$. In all other cases, we set $\Phi(k,j) = 0$.

Step 3: Select best feature. We need to select a feature that performs best under current weight D_i^t for query image i. This step is very efficient since we constrain our weak ranker to be one of the F similarity measurements $s_m(\cdot,\cdot)$. The best feature m_t is found by enumerating all F features.

Step 4: Calculate ensemble weight. It is proven in [51] that minimizing $\widehat{Z}^t = (\frac{1-r^t}{2})e^{\alpha^t} + (\frac{1+r^t}{2})e^{-\alpha^t}$ in each step of boosting is approximately equivalent to minimizing the upper bound of the rank loss, where $r^t = \sum_{j,k} D_i^t(j,k)[s_{m_t}(i,k) - s_{m_t}(i,j)]$. Since we are looking for a single weighting scheme for each category, variations in α_t obtained for different query images are penalized by an additional smoothness term. The objective becomes $\widehat{Z}^t = (\frac{1-r^t}{2})e^{\alpha^t} + (\frac{1+r^t}{2})e^{-\alpha^t} + \frac{\lambda}{2}(e^{\alpha^t-\alpha^{t-1}} + e^{\alpha^{t-1}-\alpha^t})$, where λ is a hyperparameter to balance the new term and the old terms. In our implementation, $\lambda = 1$ is used. Note that the third term takes minimum value if and only if $\alpha^t = \alpha^{t-1}$, so by imposing this new term, we are looking for a common α_m^q for all query images i in current intention category while trying to reduce all the losses $L_i, i \in Q_q$. Letting $\frac{\partial \widehat{Z}^t}{\partial \alpha^t} = 0$, we know that $\widehat{Z}^t$ is minimized when

$$\alpha^t = \frac{1}{2}\ln\left(\frac{1 + r^t + e^{\alpha^{t-1}}}{1 - r^t + e^{-\alpha^{t-1}}}\right). \tag{1}$$

Step 8: Output final weight for feature fusion. The final output of the new RankBoost algorithm is a linear combination of all the base rankers generated in each step. However, since there are actually F base rankers, the output is equivalent to a weighted combination of the F similarity measurements.

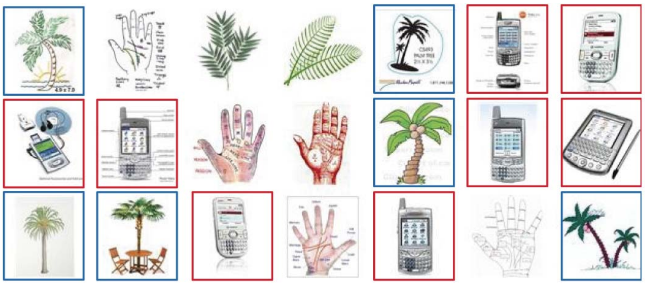


Fig. 4. An example of content-based image ranking result with many irrelevant images among top-ranked images. The query keyword is "palm" and the query image is the top leftmost image of "palm tree."

3.4 Keyword Expansion

Once the top k images most similar to the query image are found according to the visual similarity metric introduced in Sections 3.2 and 3.3, words from their textual descriptions[4] are extracted and ranked, using the *term frequency-inverse document frequency* (tf-idf) [45] method. The top m ($m = 5$ in our experiments) words are reserved as candidates for query expansion.

Because of the semantic diversity of the top k images, the word with the highest *tf-idf* score may not capture the user's search intention. Some image annotation algorithms utilized the visual content of the top k images for word expansion. For instance, Wang et al. [44] gave each image i a weight $weight(i)$ according to its visual distance $d(i)$ to the query image,

$$weight(i) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-d^2(i)/2\sigma^2}, \tag{2}$$

and the scores of words were calculated as weighted sum of tf-idf values. If there are many irrelevant images among the top k images, the performance of these methods is degraded. Fig. 4 shows such an example. The query keyword is "palm" and the query image is the top leftmost image of "palm tree." Its top-ranked images using the adaptive weight schema are shown from left to right and from top to bottom. They include images of "palm tree" (marked by blue rectangles), "palm treo" (marked by red rectangles), "palm leaves," and "palm reading." There are more images of "palm treo" than those of "palm tree," and some images of "palm tree" are ranked in low positions. Thus, the word "treo" gets the highest score calculated either by tf-idf values or tf-idf value weighted by visual distance.

We do keyword expansion through image clustering. For each candidate word w_i, all the images containing w_i in the image pool are found. However, they cannot be directly used as the visual representations of w_i for two reasons. First, there may be a number of noisy images irrelevant to w_i. Second, even if these images are relevant to w_i semantically, they may have quite different visual content. Fig. 5 shows such an example. In order to find images with similar visual content as the query example and remove noisy images, we divide these images into different clusters

[4] The file names and the surrounding texts of images.

Fig. 5. Examples of images containing the same word "palm tree" but with different visual content.

using k-Means. The number of clusters is empirically set to be $n/6$, where n is the number of images to cluster.

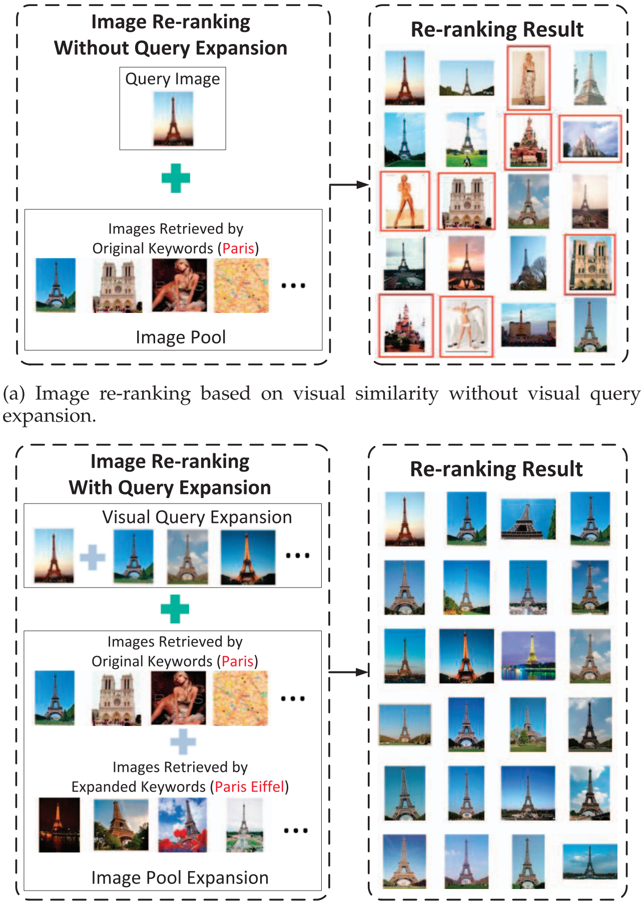
Each word w_i has t_i clusters $C(w_i) = \{c_{i,1}, \ldots, c_{i,t_i}\}$. The visual distance between the query image and a cluster c is calculated as the mean of the distances between the query image and the images in c. The cluster $c_{i,j}$ with the minimal distance is chosen as visual query expansion and its corresponding word w_i, combined with the original keyword query q, is chosen as keyword expansion q'. See the example in Fig. 3. If the distance between the closest cluster and the query image is larger than a threshold ρ, it indicates that there is no suitable image cluster and word to expand the query, and thus query expansion will not be used.

3.5 Visual Query Expansion

So far we only have one positive image example which is the query image. The goal of visual query expansion is to obtain multiple positive example images to learn a visual similarity metric which is more robust and more specific to the query image. An example in Fig. 6 explains the motivation. The query keyword is "Paris" and the query image is an image of "eiffel tower." The image reranking result based on visual similarities without visual expansion is shown in Fig. 6a and there are many irrelevant images among the top-ranked images. This is because the visual similarity metric learned from one query example image is not robust enough. By adding more positive examples to learn a more robust similarity metric, such irrelevant images can be filtered out. In a traditional way, adding additional positive examples was typically done through relevance feedback, which required more users' labeling burden. We aim at developing an image reranking method, which only requires one click on the query image and thus positive examples have to be obtained automatically. The cluster of images chosen in Section 3.4 has the closest visual distance to the query example and have consistent semantic meanings. Thus, they are used as additional positive examples for visual query expansion. We adopt the one-class SVM [22] to refine the visual similarity in Section 3.3. The one-class SVM classifier is trained from the additional positive examples obtained by visual query expansion. It requires defining the kernel between images, and the kernel is computed from the similarity introduced in Section 3.3. An image to be reranked is input to the one-class SVM classifier and the output is used as the similarity (sim_V) to the query image. Notice the effect of this step is similar to relevance feedback [52]. However, the key difference is that instead of asking users to add the positive samples manually, our method is fully automatic.

3.6 Image Pool Expansion

Considering efficiency, image search engines such as Bing image search only rerank the top N images of the text-based image search result. If the query keywords do not capture the user's search intention accurately, there are only a small number of relevant images with the same semantic meanings as the query image in the image pool. This can significantly degrade the ranking performance. In Section 3.3, we rerank the top N retrieved images by the original keyword query based on their visual similarities to the query image. We remove the $N/2$ images with the lowest ranks from the image pool. Using the expanded keywords as query, the top $N/2$ retrieved images are added to the image pool. We believe that there are more relevant images in the image pool with the help of expanded query keywords. The reranking result by extending image pool and positive example images is shown in Fig. 6b, which is significantly improved compared with Fig. 6a.



(a) Image re-ranking based on visual similarity without visual query expansion.

(b) Image re-ranking by extending the image pool and positive example images.

Fig. 6. An example of image reranking using "Paris" as query keyword and an image of "eiffel tower" as query image. Irrelevant images are marked by red rectangles.

3.7 Combining Visual and Textual Similarities

Learning a query specific textual similarity metric from the positive examples $E = \{e_1, \ldots, e_j\}$ obtained by visual query expansion and combining it with the query specific visual similarity metric introduced in Section 3.5 can further improve the performance of image reranking. For a selected query image, a word probability model is trained from E and used to compute the textual distance $dist_T$. We adopt the approach in [31]. Let θ be the parameter of a discrete distribution of words over the dictionary. Each image i is regarded as a document d_i where the words are extracted from its textual descriptions (see definition in 3.4). θ is learned by maximizing the observed probability

$$\Pi_{e_i \in E} \Pi_{w \in d_i} (\lambda p(w|\theta) + (1-\lambda) p(w|C))^{n_w^i},$$

where λ is a fixed parameter set to be 0.5, w is a word, and n_w^i is the frequency of w in d_i. $p(w|C)$ is the word probability built upon the whole repository C:

$$p(w|C) = \frac{\sum_{d_i} n_w^i}{|C|}.$$

θ can be learned by the Expectation-Maximization algorithm. Once θ is learned, for an image k its textual distance to the positive examples is defined by cross-entropy function:

$$dist_T(k) = -\sum_w p(w|d_k) \log(w|\theta).$$

Here, $p(w|d_i) = n_w^k / |d_k|$. At last, this textual distance can be combined with the visual similarity sim_V obtained in Section 3.5 to rerank images:

$$-\alpha \cdot sim_V(k) + (1-\alpha) \cdot dist_T(k).$$

α is a fixed parameter and set as 0.5.

3.8 Summary

The goal of the proposed framework is to capture user intention and is achieved in multiple steps. The user intention is first roughly captured by classifying the query image into one of the coarse semantic categories and choosing a proper weight schema accordingly. The adaptive visual similarity obtained from the selected weight schema is used in all the following steps. Then according to the query keywords and the query image provided by the user, the user intention is further captured in two aspects: 1) finding more query keywords (called keyword expansion) describing user intention more accurately 2) and in the meanwhile finding a cluster of images (called visual query expansion) which are both visually and semantically consistent with the query image. The keyword expansion frequently co-occurs with the query keywords and the visual expansion is visually similar to the query image. Moreover, it is required that all the images in the cluster of visual query expansion contain the same keyword expansion. Therefore, the keyword expansion and visual expansion support each other and are obtained simultaneously. In the later steps, the keyword expansion is used to expand the image pool to include more images relevant to user intention, and the visual query expansion is used to learn visual and textual similarity metrics which better reflect user intention.

4 Experimental Evaluation

In the first experiment, 300,000 web images are manually labeled into different classes (images that are semantically similar) as ground truth. Precisions of different approaches are compared. The semantic meanings of images are closely related to users' intention. However, they are not exactly the same. Images of the same class (thus with similar semantic meanings) can be visually quite different. Thus, a user study is conducted in the second experiment to evaluate whether the search results capture the users' intentions well. Running on a machine of 3 GHz CPU and without optimizing the code, it needs less than half second computation for each query.

4.1 Experiment One: Evaluation with Ground Truth

Fifty query keywords are chosen for evaluation. Using each keyword as query, the top 1,000 images are crawled from Bing image search. These images are manually labeled into different classes. For example, for query "apple," its images are labeled as "red apple," "apple ipod," "apple pie," etc. There are totally 700 classes for all 50 query keywords. Another 500 images are crawled from Bing image search using each keyword expansion as query. These images are also manually labeled. There are in total around 300,000 images in our data set. A small portion of them are outliers and are not assigned to any category (e.g., some images are irrelevant to the query keywords). The threshold ρ (in Section 3.4) is chosen as 0.3 through cross-validation measuring and is fixed in all the experiments. The performance is stable when ρ varies between 0.25 and 0.35. Some examples of image reranking results are shown in the supplemental material, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.242.

4.1.1 Precisions on Different Steps of Our Framework

Top m precision, the proportion of relevant images among the top m ranked images, is used to evaluate the performance of image reranking. Images are considered to be relevant if they are labeled as the same class. For each query keyword, image reranking repeats many times by choosing different query images. Except for those outlier images not being assigned to any category, every image returned by keyword query has been chosen as the query image. In order to evaluate the effectiveness of different steps of our proposed image reranking framework, we compare the following approaches. From 1 to 8, more and more steps in Fig. 3 are added in.

1. **Text based**: Text-based search from Bing. It is used as the baseline.
2. **GW**: Image reranking using global weights to combine visual features (Global Weight).
3. **AW**: Image reranking using adaptive weight schema to combine visual features (Section 3.3).
4. **ExtEg**: Image reranking by extending positive examples only, from which the query specific visual similarity metric is learned and used (Section 3.5).
5. **GW + Pool**: Image reranking by extending the image pool only (Section 3.6) while using global weights to combine visual features.
6. **ExtPool**: Similar to GW + Pool, however using adaptive weight schema to combine visual features.
7. **ExtBoth(V)**: Image reranking by extending both the image pool and positive example images. Only the query specific visual similarity metric is used.
8. **ExtBoth(V + T)**: Similar to ExtBoth, however combining query specific visual and textual similarity metrics (Section 3.7). This is the complete approach proposed by us.

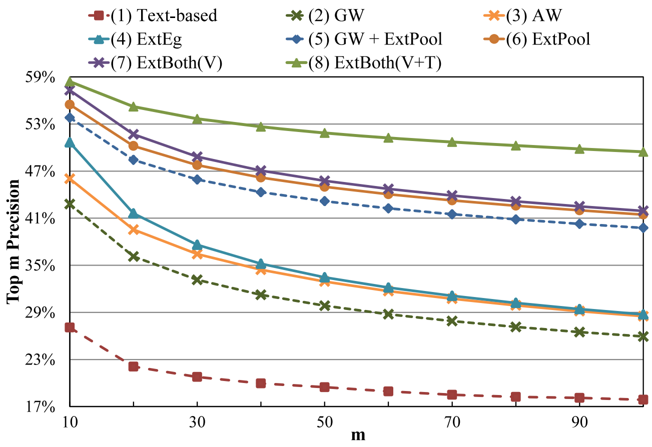


Fig. 7. Comparison of averaged top m precisions on different steps.

The averaged top m precisions are shown in Fig. 7. Approaches (2)-(7) only use visual similarity. Approach (8) uses both visual and textual similarities. Approaches (2) and (3) are initial image reranking based on the text-based search results in (1). Their difference is to combine visual features in different ways. We can see that by using a single query image we can significantly improve the text-based image search result. The proposed adaptive weight schema, which reflects user intention at a coarse level, outperforms the global weight. After initial reranking using adaptive weight, the top 50 precision of text-based research is improved from 19.5 to 32.9 percent. Approaches (4), (6), (7), and (8) are based on the initial image reranking result in (3). We can clearly see the effectiveness of expanding image pool and expanding positive image examples through keyword expansion and image clustering. These steps capture user intention at a finer level since, for every query image, the image pool and positive examples are expanded differently. Using expansions, the top 50 precision of initial reranking using adaptive weight is improved from 32.9 to 51.9 percent.

Our keyword expansion step (Section 3.4) could be replaced by other equivalent methods, such as image annotation methods. As discussed in Section 2.2, many existing image annotation methods cannot be directly applied to our problem. We compare with two image annotation approaches, which do not require a fixed set of keywords, by replacing the keyword expansion step with them. One is to choose the word with the highest tf-idf score as keyword expansion to extend image pool (**ExtPoolByTfIdf**). The other uses the method proposed in [44] which weighted the tf-idf score by images' visual similarities to extend image pool (**ExtPoolByWTfIdf**). Fig. 8 shows the result. Our ExtPool has a better performance. Moreover, image annotation only aims at ranking words and cannot automatically provide visual query expansion as our method does. Therefore, our ExtBoth has a even better performance than the other two.

4.1.2 Comparison with Other Methods

In this section, we compare with several existing approaches [29], [31], [53] which can be applied to image reranking with only one-click feedback as discussed in Section 2.2.

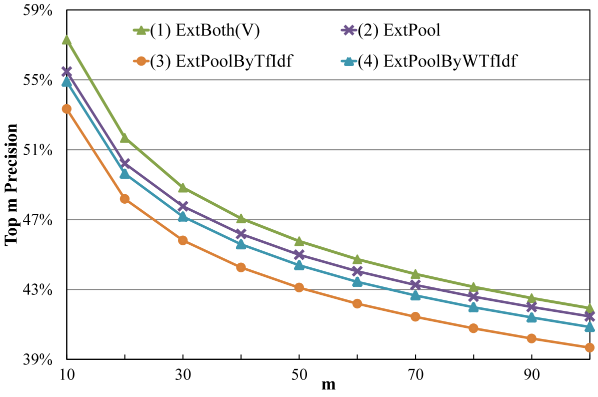


Fig. 8. Comparison of averaged top m precisions of keyword expansion through image clustering with the other two methods.

1. **ExtBoth (V + T)**: Our approach.
2. **CrossMedia**: Image reranking by transmedia distances defined in [31]. It combined both visual and textual features under the pseudorelevance feedback framework.
3. **NPRF**: Image reranking by the pseudorelevance feedback approach proposed in [29]. It used top-ranked images as positive examples and bottom-ranked images as negative examples to train an SVM.
4. **PRF**: Image reranking by the pseudorelevance feedback approach proposed in [53]. It used top-ranked images as positive examples to train a one-class SVM.

Fig. 9 shows the result. Our algorithm outperforms others, especially when m is large.

4.2 Experiment Two: User Study

The purpose of the user study is to evaluate the effectiveness of visual expansions (expanding both the image pool and positive visual examples) to capture user intention. Forty users are invited. For each query keyword, the user is asked to browse the images and to randomly select an image of interest as a query example. We show them the initial image reranking results of using adaptive weight schema (Section 3.3) and the results of extending both the image pool and positive example images. The users are asked to do the following:

- Mark irrelevant images among the top 10 images.
- Compare the top 50 retrieved images by both results and choose whether the final result with visual

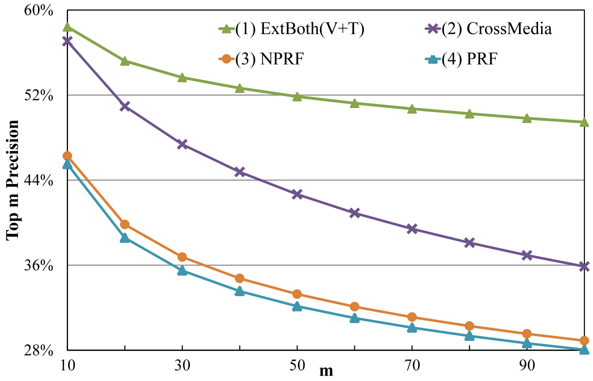


Fig. 9. Comparison of averaged top m precisions with existing methods.

TABLE 1
Average Number of "Irrelevant" Images

	No visual expansions	With visual expansions
Top 10 images	4.26	2.33
Top 5 images	1.67	0.91
Top 3 images	0.74	0.43

expansions is "much better," "somewhat better," "similar," "somewhat worse," or "much worse" than that of initial reranking using an adaptive weight schema.

Each user is assigned five query keywords from all 50 keywords. Given each query keyword, the user is asked to choose 30 different query images and compare their reranking results. As shown in Table 1, visual expansions significantly reduce the average numbers of irrelevant images among top 10 images. Fig. 10 shows that in most cases (> 67 percent) the users think visual expansions improve the result.

4.3 Discussion

In our approach, the keyword expansion (Section 3.4), visual query expansion (Section 3.5), and image pool expansion (Section 3.6) all affect the quality of initial image reranking result (Section 3.3). According to our experimental evaluation and user study, if the quality of initial image reranking is reasonable, which means that there are a few relevant examples among top-ranked images, the following expansion steps can significantly improve the reranking performance. Inappropriate expansions which significantly deteriorate the performance happen in the cases when the initial rerank result is very poor. The chance is lower than 2 percent according to our user study in Fig. 10.

In this paper, it is assumed that an image captures user intention when it is both semantically and visually similar to the query image. However, in some cases user intention cannot by well expressed by a single query image. For instance, the user may be interested in only part of the image. In those cases, more user interactions, such as labeling the regions that the user thinks are "important," have to be allowed. However, more user burden has to be added and it is not considered in this paper.

5 Conclusion

In this paper, we propose a novel Internet image search approach which only requires one-click user feedback. Intention specific weight schema is proposed to combine visual features and to compute visual similarity adaptive to query images. Without additional human feedback, textual and visual expansions are integrated to capture user intention. Expanded keywords are used to extend positive example images and also enlarge the image pool to include more relevant images. This framework makes it possible for industrial scale image search by both text and visual content. The proposed new image reranking framework consists of multiple steps, which can be improved separately or replaced by other techniques equivalently effective. In future work, this framework can be further improved by making use of the query log data, which provides valuable co-occurrence information of keywords, for keyword expansion. One shortcoming of the current system is that sometimes duplicate images show up as similar images to the query. This can be improved by including duplicate detection in the future work. Finally, to further improve the quality of reranked images, we intend to combine this work with photo quality assessment work in [54], [55], and [56] to rerank images not only by content similarity but also by the visual quality of the images.

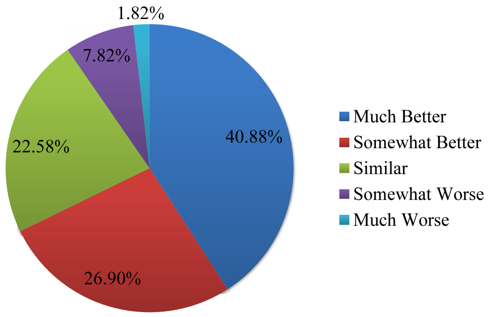


Fig. 10. Percentages of cases when users think the result of visual expansions is "much better," "somewhat better," "similar," "somewhat worse," and "much worse" than the result of only using intention weight schema.

Acknowledgments

This work is partially supported by Hong Kong SAR through RGC project 416510 and by Guangdong Province through Introduced Innovative R&D Team of Guangdong Province 201001D0104648280.

References

[1] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W. Ma, "Igroup: Web Image Search Results Clustering," *Proc. 14th Ann. ACM Int'l Conf. Multimedia,* 2006.
[2] J. Cui, F. Wen, and X. Tang, "Real Time Google and Live Image Search Re-Ranking," *Proc. 16th ACM Int'l Conf. Multimedia,* 2008.
[3] J. Cui, F. Wen, and X. Tang, "IntentSearch: Interactive On-Line Image Search Re-Ranking," *Proc. 16th ACM Int'l Conf. Multimedia,* 2008.
[4] "Bing Image Search," http://www.bing.com/images, 2012.
[5] N. Ben-Haim, B. Babenko, and S. Belongie, "Improving Web-Based Image Search via Content Based Clustering," *Proc. Int'l Workshop Semantic Learning Applications in Multimedia,* 2006.
[6] R. Fergus, P. Perona, and A. Zisserman, "A Visual Category Filter for Google Images," *Proc. European Conf. Computer Vision,* 2004.
[7] G. Park, Y. Baek, and H. Lee, "Majority Based Ranking Approach in Web Image Retrieval," *Proc. Second Int'l Conf. Image and Video Retrieval,* 2003.
[8] Y. Jing and S. Baluja, "Pagerank for Product Image Search," *Proc. Int'l Conf. World Wide Web,* 2008.
[9] W.H. Hsu, L.S. Kennedy, and S.-F. Chang, "Video Search Reranking via Information Bottleneck Principle," *Proc. 14th Ann. ACM Int'l Conf. Multimedia,* 2006.
[10] R. Datta, D. Joshi, and J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Computing Surveys,* vol. 40, pp. 1-60, 2007.
[11] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-Based Vision System for Place and Object Recognition," *Proc. Int'l Conf. Computer Vision,* 2003.
[12] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2005.

[13] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.

[14] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," *Proc. Int'l Conf. Computer Vision,* 2003.

[15] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Spatial-Bag-of-Features," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2010.

[16] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Descriptor Learning for Efficient Retrieval," *Proc. European Conf. Computer Vision,* 2010.

[17] Y. Zhang, Z. Jia, and T. Chen, "Image Retrieval with Geometry-Preserving Visual Phrases," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2011.

[18] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large Scale Online Learning of Image Similarity through Ranking," *J. Machine Learning Research,* vol. 11, pp. 1109-1135, 2010.

[19] J. Deng, A.C. Berg, and L. Fei-Fei, "Hierarchical Semantic Indexing for Large Scale Image Retrieval," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2011.

[20] K. Tieu and P. Viola, "Boosting Image Retrieval," *Int'l J. Computer Vision,* vol. 56, no. 1, pp. 17-36, 2004.

[21] S. Tong and E. Chang, "Support Vector Machine Active Learning for Image Retrieval," *Proc. ACM Multimedia,* 2001.

[22] Y. Chen, X. Zhou, and T. Huang, "One-Class SVM for Learning in Image Retrieval," *Proc. IEEE Int'l Conf. Image Processing,* 2001.

[23] Y. Lu, H. Zhang, L. Wenyin, and C. Hu, "Joint Semantics and Feature Based Image Retrieval Using Relevance Feedback," *IEEE Trans. Multimedia,* vol. 5, no. 3, pp. 339-347, Sept. 2003.

[24] D. Tao and X. Tang, "Random Sampling Based SVM for Relevance Feedback Image Retrieval," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2004.

[25] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 7, pp. 1088-1099, July 2006.

[26] T. Quack, U. Monich, L. Thiele, and B. Manjunath, "Cortina: A System for Large-Scale, Content-Based Web Image Retrieval," *Proc. 12th Ann. ACM Int'l Conf. Multimedia,* 2004.

[27] Y. Huang, Q. Liu, S. Zhang, and D.N. Metaxas, "Image Retrieval via Probabilistic Hypergraph Ranking," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2011.

[28] R. Yan, E. Hauptmann, and R. Jin, "Multimedia Search with Pseudo-Relevance Feedback," *Proc. Int'l Conf. Image and Video Retrieval,* 2003.

[29] R. Yan, A.G. Hauptmann, and R. Jin, "Negative Pseudo-Relevance Feedback in Content-Based Video Retrieval," *Proc. 11th ACM Int'l Conf. Multimedia,* 2003.

[30] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[31] J. Ah-Pine, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, and J. Renders, "Crossing Textual and Visual Content in Different Application Scenarios," *Multimedia Tools and Applications,* vol. 42, pp. 31-56, 2009.

[32] J. Krapac, M. Allan, J. Verbeek, and F. Jurie, "Improving Web Image Search Results Using Query-Relative Classifiers," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2010.

[33] B. Luo, X. Wang, and X. Tang, "A World Wide Web Based Image Search Engine Using Text and Image Content Features," *Proc. IS&T/SPIE Electronic Imaging, Internet Imaging IV,* 2003.

[34] Z. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang, "Visual Query Expansion," *Proc. 17th ACM Int'l Conf. Multimedia,* 2009.

[35] A. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan, "Semantic Concept-Based Query Expansion and Re-Ranking for Multimedia Retrieval," *Proc. 15th Int'l Conf. Multimedia,* 2007.

[36] J. Smith, M. Naphade, and A. Natsev, "Multimedia Semantic Indexing Using Model Vectors," *Proc. Int'l Conf. Multimedia and Expo.,* 2003.

[37] A. Frome, Y. Singer, F. Sha, and J. Malik, "Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[38] Y. Lin, T. Liu, and C. Fuh, "Local Ensemble Kernel Learning for Object Category Recognition," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2007.

[39] S. Liu, F. Liu, C. Yu, and W. Meng, "An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases," *Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* 2004.

[40] S. Kim, H. Seo, and H. Rim, "Information Retrieval Using Word Senses: Root Sense Tagging Approach," *Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* 2004.

[41] K. Sparck Jones, *Automatic Keyword Classification for Information Retrieval.* Archon Books, 1971.

[42] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *J. Am. Soc. for Information Science,* vol. 41, no. 6, pp. 391-407, 1990.

[43] L. Wu, L. Yang, N. Yu, and X. Hua, "Learning to Tag," *Proc. Int'l Conf. World Wide Web,* 2009.

[44] C. Wang, F. Jing, L. Zhang, and H. Zhang, "Scalable Search-Based Image Annotation of Personal Images," *Proc. Eighth ACM Int'l Workshop Multimedia Information Retrieval,* 2006.

[45] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., 1999.

[46] M. Unser, "Texture Classification and Segmentation Using Wavelet Frames," *IEEE Trans. Image Processing,* vol. 4, no. 11, pp. 1549-1560, Nov. 1995.

[47] Y. Rubner, L. Guibas, and C. Tomasi, "The Earth Movers Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval," *Proc. ARPA Image Understanding Workshop,* 1997.

[48] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to Detect a Salient Object," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2007.

[49] W. Freeman and M. Roth, "Orientation Histograms for Hand Gesture Recognition," *Proc. Int'l Workshop Automatic Face and Gesture Recognition,* 1995.

[50] R. Xiao, H. Zhu, H. Sun, and X. Tang, "Dynamic Cascades for Face Detection," *Proc. Int'l Conf. Computer Vision,* 2007.

[51] Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer, "An Efficient Boosting Algorithm for Combining Features," *J. Machine Learning Research,* vol. 4, pp. 933-969, 2003.

[52] X.S. Zhou and T.S. Huang, "Relevance Feedback in Image Retrieval: A Comprehensive Review," *Multimedia Systems,* vol. 8, pp. 536-544, 2003.

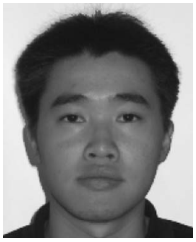
[53] J. He, M. Li, Z. Li, H. Zhang, H. Tong, and C. Zhang, "Pseudo Relevance Feedback Based on Iterative Probabilistic One-Class SVMs in Web Image Retrieval," *Proc. Pacific-Rim Conf. Multimedia,* 2004.

[54] Y. Ke, X. Tang, and F. Jing, "The Design of High-Level Features for Photo Quality Assessment," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2006.

[55] Y. Luo and X. Tang, "Photo and Video Quality Evaluation: Focusing on the Subject," *Proc. European Conf. Computer Vision,* 2008.

[56] W. Luo, X. Wang, and X. Tang, "Content-Based Photo Quality Assessment," *Proc. IEEE Int'l Conf. Computer Vision,* 2011.

Xiaoou Tang received the BS degree from the University of Science and Technology of China, Hefei, in 1990, the MS degree from the University of Rochester, Rochester, New York, in 1991, and the PhD degree from the Massachusetts Institute of Technology, Cambridge, in 1996. He is a professor in the Department of Information Engineering and an associate dean (research) of the Faculty of Engineering of the Chinese University of Hong Kong. He worked as the group manager of the Visual Computing Group at Microsoft Research Asia from 2005 to 2008. His research interests include computer vision, pattern recognition, and video processing. He received the Best Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009. He was a program chair of the IEEE International Conference on Computer Vision (ICCV) 2009 and is an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* and the *International Journal of Computer Vision (IJCV).* He is a fellow of the IEEE.



Ke Liu received the BS degree from Tsinghua University in computer science in 2009 and is currently working toward the MPhil degree in the Department of Information Engineering at the Chinese University of Hong Kong. His research interests include image search and computer vision.

Jingyu Cui received the BEng and MSc degrees in automation from Tsinghua University in 2005 and 2008, respectively, the MSc degree in electrical engineering in 2010 from Stanford University, where he is currently working toward the PhD degree. His research interests include visual recognition and retrieval, machine learning, and parallel computing. He is a student member of the IEEE.

Fang Wen received the BS degree in automation from Tsinghua University, and the MS and PhD degrees in pattern recognition and intelligent system in 1997 and 2003, respectively. Now she is a researcher in the Visual Computing Group at Microsoft Research Asia. Her research interests include computer vision, pattern recognition, and multimedia search. She is a member of the IEEE.

Xiaogang Wang received the BS degree in electrical engineering and information science from the University of Science and Technology of China in 2001, and the MS degree in information engineering from the Chinese University of Hong Kong in 2004, and the PhD degree in computer science from the Massachusetts Institute of Technology. He is currently an assistant professor in the Department of Electronic Engineering at the Chinese University of Hong Kong. His research interests include computer vision and machine learning. He is a member of the IEEE.