Web Image Re-ranking Using Query-Specific Semantic Signatures

Xiaogang Wang, *Member, IEEE,* Shi Qiu, Ke Liu, and Xiaoou Tang, *Fellow, IEEE*

Abstract—Image re-ranking, as an effective way to improve the results of web-based image search, has been adopted by current commercial search engines such as Bing and Google. Given a query keyword, a pool of images are first retrieved based on textual information. By asking the user to select a query image from the pool, the remaining images are re-ranked based on their visual similarities with the query image. A major challenge is that the similarities of visual features do not well correlate with images' semantic meanings which interpret users' search intention. Recently people proposed to match images in a semantic space which used attributes or reference classes closely related to the semantic meanings of images as basis. However, learning a universal visual semantic space to characterize highly diverse images from the web is difficult and inefficient. In this paper, we propose a novel image re-ranking framework, which automatically offline learns different semantic spaces for different query keywords. The visual features of images are projected into their related semantic spaces to get semantic signatures. At the online stage, images are re-ranked by comparing their semantic signatures obtained from the semantic space specified by the query keyword. The proposed query-specific semantic signatures significantly improve both the accuracy and efficiency of image re-ranking. The original visual features of thousands of dimensions can be projected to the semantic signatures as short as 25 dimensions. Experimental results show that $25\% - 40\%$ relative improvement has been achieved on re-ranking precisions compared with the state-of-the-art methods.

Index Terms—Image search, Image re-ranking, Semantic space, Semantic signature, Keyword expansion

1 Introduction

Web-scale image search engines mostly use keywords as queries and rely on surrounding text to search images. They suffer from the ambiguity of query keywords, because it is hard for users to accurately describe the visual content of target images only using keywords. For example, using "apple" as a query keyword, the retrieved images belong to different categories (also called concepts in this paper), such as "red apple", "apple logo", and "apple laptop". In order to solve the ambiguity, content-based image retrieval [1], [2] with relevance feedback [3]–[5] is widely used. It requires users to select multiple relevant and irrelevant image examples, from which visual similarity metrics are learned through online training. Images are re-ranked based on the learned visual similarities. However, for web-scale commercial systems, users' feedback has to be limited to the minimum without online training.

Online image re-ranking [6]–[8], which limits users' effort to just one-click feedback, is an effective way to improve search results and its interaction is simple enough. Major web image search engines have adopted this strategy [8]. Its diagram is shown in Figure 1. Given a query keyword input by a user, a pool of images relevant to the query keyword are retrieved by the search engine according to a stored word-image index file. Usually the size of the returned image pool is fixed, e.g. containing $1,000$ images. By asking the user to select a query image, which reflects the user's search intention, from the pool, the remaining images in the pool are re-ranked based on their visual similarities with the query image. The word-image index file and visual features of images are pre-computed offline and stored[1]. The main online computational cost is on comparing visual features. To achieve high efficiency, the visual feature vectors need to be short and their matching needs to be fast. Some popular visual features are in high dimensions and efficiency is not satisfactory if they are directly matched.

Another major challenge is that, without online training, the similarities of low-level visual features may not well correlate with images' high-level semantic meanings which interpret users' search intention. Some examples are shown in Figure 2. Moreover, low-level features are sometimes inconsistent with visual perception. For example, if images of the same object are captured from different viewpoints, under different lightings or even with different compression artifacts, their low-level features may change significantly, although humans think the visual content does not change much. To reduce this

- *X. Wang is with the Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong.*
- *S. Qiu and K. Liu are with the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong.*
- *X. Tang is with the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong, and the Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, CAS, China.*

[1] Visual features must be saved. The web image collection is dynamically updated. If the visual features are discarded and only the similarity scores of images are stored, whenever a new image is added into the collection and we have to compute its similarities with existing images, whose visual features need be computed again.

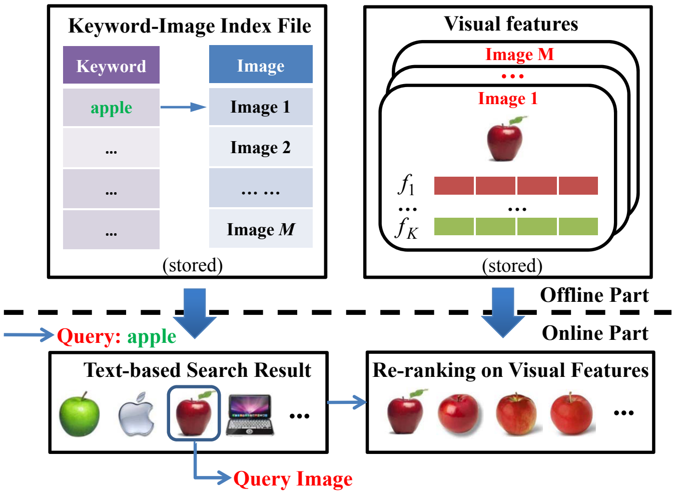


Fig. 1. The conventional image re-ranking framework.

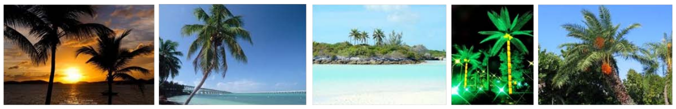


Fig. 2. All the images shown in this figure are related to palm trees. They are different in color, shape, and texture.

semantic gap and inconsistency with visual perception, there have been a number of studies to map visual features to a set of predefined concepts or attributes as semantic signatures [9]–[12]. For example, Kovashka *et al.* [12] proposed a system which refined image search with relative attribute feedback. Users described their search intention with reference images and a set of pre-defined attributes. These concepts and attributes are pre-trained offline and have tolerance with variation of visual content. However, these approaches are only applicable to closed image sets of relatively small sizes, but not suitable for online web-scale image re-ranking. According to our empirical study, images retrieved by 120 query keywords alone include more than 1500 concepts. It is difficult and inefficient to design a huge concept dictionary to characterize highly diverse web images. Since the topics of web images change dynamically, it is desirable that the concepts and attributes can be automatically found instead of being manually defined.

1.1 Our Approach

In this paper, a novel framework is proposed for web image re-ranking. Instead of manually defining a universal concept dictionary, it learns different semantic spaces for different query keywords individually and automatically. The semantic space related to the images to be re-ranked can be significantly narrowed down by the query keyword provided by the user. For example, if the query keyword is "apple", the concepts of "mountain" and "Paris" are irrelevant and should be excluded. Instead, the concepts of "computer" and "fruit" will be used as dimensions to learn the semantic space related to "apple". The query-specific semantic spaces can more accurately model the images to be re-ranked, since they have excluded other potentially unlimited number of irrelevant concepts, which serve only as noise and deteriorate the re-ranking performance on both accuracy and computational cost. The visual and textual features of images are then projected into their related semantic spaces to get semantic signatures. At the online stage, images are re-ranked by comparing their semantic signatures obtained from the semantic space of the query keyword. The semantic correlation between concepts is explored and incorporated when computing the similarity of semantic signatures.

Our experiments show that the semantic space of a query keyword can be described by just $20-30$ concepts (also referred as "reference classes"). Therefore the semantic signatures are very short and online image re-ranking becomes extremely efficient. Because of the large number of keywords and the dynamic variations of the web, the semantic spaces of query keywords are automatically learned through keyword expansion.

We introduce a large scale benchmark database[2] with manually labeled ground truth. It includes $120,000$ images retrieved by the Bing Image Search using 120 query keywords. Experiments on this database show that $25\% - 40\%$ relative improvement has been achieved on re-ranking precisions with around 70 times speedup, compared with the state-of-the-art methods.

The proposed query-specific semantic signatures are also effective on image re-ranking without query images being selected [13]–[32]. The effectiveness is shown in Section 7 through evaluation on the MSRA-MM dataset [33] and comparison with the state-of-the-art methods.

1.2 Discussion on Search Scenarios

We consider the following search scenarios when designing the system and doing evaluation. When a user inputs a textual query (e.g. "Disney") and starts to browse the text-based research result, he or she has a search intention, which could be a particular target image or images in a particular category (e.g. images of Cinderella Castle). Once the user finds a candidate image similar to the target image or belonging to the category of interest, the re-ranking function is used by choosing that candidate image as a query image. Certain criteria should be considered in these search scenarios. (1) In both cases, we expect the top ranked images are in the same semantic category as the query image (e.g. images of princesses and Disney logo are considered as irrelevant). (2) If the search intention is to find a target image, we expect that images visually similar to the query image should have higher ranks. (3) If the search intention is to browse images of a particular semantic category, diversity of candidate images may also be considered.

The first two criteria have been considered in our system design. Our query-specific semantic signatures effectively reduce the gap between low-level visual features

[2] http://mmlab.ie.cuhk.edu.hk/CUHKSR/Dataset.htm

and semantic categories, and also make image matching more consistent with visual perception. Details in later sections will show that if a candidate image is very similar to the query image, the distance of their semantic signatures will be close to zero and the candidate image will have a high rank. To evaluate the first criterion in experiments, we manually label images into categories according to their semantic meanings, and compare with re-ranking results in Section 6.1-6.6. It is measured with precisions. The second criterion is more subjective. We conduct a user study in Section 6.7, where the subjects were informed to consider the first two criteria.

In this paper, we do not consider increasing the diversity of search result by removing near-duplicate or very similar images, which is another important issue in web image search and has a lot of existing works [34], [35]. We re-rank the first $1,000$ candidate images returned by the commercial web image search engine, which has considered the diversity issue and removed many near-duplicate images. The query-specific semantic signature is proposed to reduce semantic gap but cannot directly increase the diversity of search result. We do not address the diversity problem to make the paper focused on semantic signatures. However, we believe that the two aspects can incorporated in multiple possible ways.

2 Related Work

The key component of image re-ranking is to compute visual similarities reflecting semantic relevance of images. Many visual features [36]–[40] have been developed in recent years. However, for different query images, the effective low-level visual features are different. Therefore, Cui *et al.* [6], [7] classified query images into eight predefined intention categories and gave different feature weighting schemes to different types of query images. But it was difficult for the eight weighting schemes to cover the large diversity of all the web images. It was also likely for a query image to be classified to a wrong category. In order to reduce the semantic gap, query-specific semantic signature was first proposed in [41]. Kuo *et al.* [42] recently augmented each image with relevant semantic features through propagation over a visual graph and a textual graph which were correlated.

Another way of learning visual similarities without adding users' burden is pseudo relevance feedback [43]–[45]. It takes the top N images most visually similar to the query image as expanded positive examples to learn a similarity metric. Since the top N images are not necessarily semantically-consistent with the query image, the learned similarity metric may not reliably reflect the semantic relevance and may even deteriorate re-ranking performance. In object retrieval, in order to purify the expanded positive examples, the spatial configurations of local visual features are verified [46]–[48]. But it is not applicable to general web image search, where relevant images may not contain the same objects.

There is a lot of work [13]–[32] on using visual features to re-rank images retrieved by initial text-only search, however, without requiring users to select query images. Tian *et al.* [24] formulated image re-ranking with a Bayesian framework. Hsu *et al.* [15] used the Information Bottleneck (IB) principle to maximize the mutual information between search relevance and visual features. Krapac *et al.* [26] introduced generic classifiers based on query-relative features which could be used for new query keywords without additional training. Jing *et al.* [21] proposed VisualRank to analyze the visual link structures of images and to find the visual themes for re-ranking. Lu *et al.* [31] proposed the deep context to refine search results. Cai *et al.* [32] re-ranked images with attributes which were manually defined and learned from manually labeled training samples. These approaches assumed that there was one major semantic category under a query keyword. Images were re-ranked by modeling this dominant category with visual and textual features. In Section 7, we show that the proposed query-specific semantic signature is also effective in this application, where it is crucial to reduce the semantic gap when computing the similarities of images. Due to the ambiguity of query keywords, there may be multiple semantic categories under one keyword query. Without query images selected by users, these approaches cannot accurately capture users' search intention.

Recently, for general image recognition and matching, there have been a number of works on using projections over predefined concepts, attributes or reference classes as image signatures. The classifiers of concepts, attributes, and reference classes are trained from known classes with labeled examples. But the knowledge learned from the known classes can be transferred to recognize samples of novel classes which have few or even no training samples. Since these concepts, attributes, and reference classes are defined with semantic meanings, the projections over them can well capture the semantic meanings of new images even without further training. Rasiwasia *et al.* [9] mapped visual features to a universal concept dictionary for image retrieval. Attributes [49] with semantic meanings were used for object detection [10], [50], [51], object recognition [52]–[60], face recognition [58], [61], [62], image search [60], [63]–[67], action recognition [68], and 3D object retrieval [69]. Lampert *et al.* [10] predefined a set of attributes on an animal database and detected target objects based on a combination of human-specified attributes instead of training images. Sharmanska *et al.* [50] augmented this representation with additional dimensions and allowed a smooth transition between zero-shot learning, unsupervised training and supervised training. Parikh and Grauman [58] proposed relative attributes to indicate the strength of an attribute in an image with respect to other images. Parkash and Parikh [60] used attributes to guide active learning. In order to detect objects of many categories or even unseen categories, instead of building a new detector for each category, Farhadi *et al.* [51] learned part and attribute detectors which were shared across categories and modeled the correlation

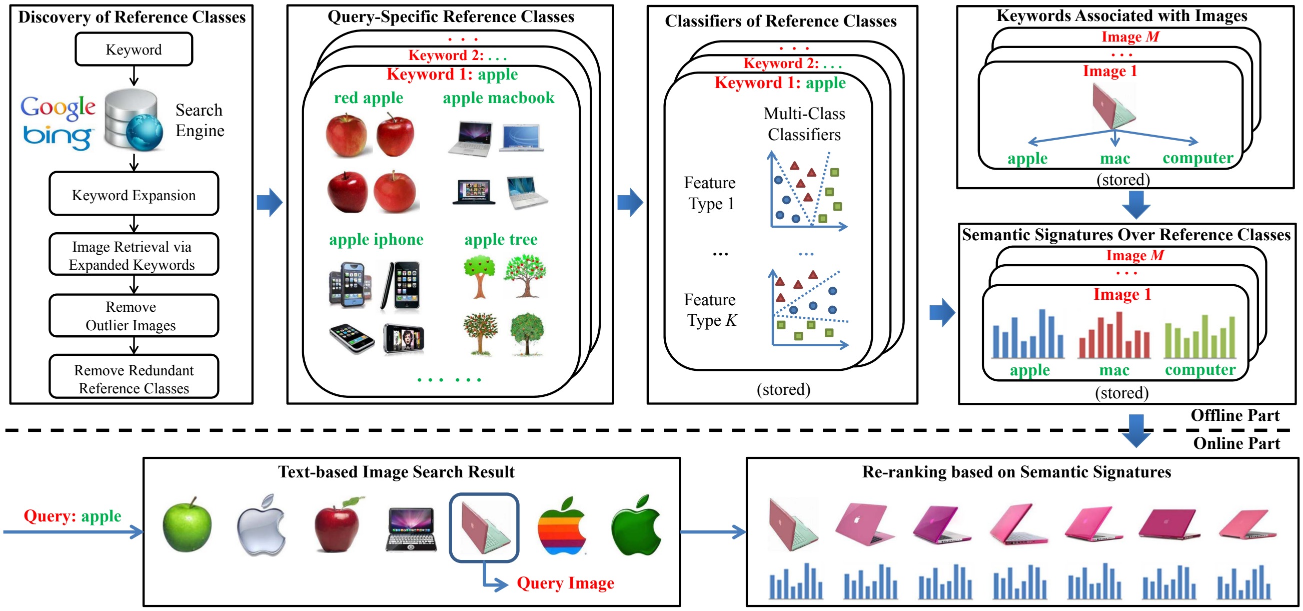


Fig. 3. Diagram of our new image re-ranking framework.

among attributes. Some approaches [11], [54], [70], [71] transferred knowledge between object classes by measuring the similarities between novel object classes and known object classes (called reference classes). For example, Torresani *et al.* [71] proposed an image descriptor which was the output of a number of classifiers on a set of known image classes, and used it to match images of other unrelated visual classes. In the current approaches, all the concepts/attributes/reference-classes are universally applied to all the images and they are manually defined. They are more suitable for offline databases with lower diversity (such as animal databases [10], [54] and face databases [11]), since image classes in these databases can better share similarities. To model all the web images, a huge set of concepts or reference classes are required, which is impractical and ineffective for online image re-ranking. Intuitively, only a small subset of the concepts are relevant to a specific query. Many concepts irrelevant to the query not only increase the computational cost but also deteriorate the accuracy of re-ranking. However, how to automatically find such relevant concepts and use them for online web image re-ranking was not well explored in previous studies.

3 Approach Overview

The diagram of our approach is shown in Figure 3. It has offline and online parts. At the offline stage, the reference classes (which represent different concepts) related to query keywords are automatically discovered and their training images are automatically collected in several steps. For a query keyword (e.g. “apple”), a set of most relevant keyword expansions (such as “red apple” and “apple macbook”) are automatically selected utilizing both textual and visual information. This set of keyword expansions defines the reference classes for the query keyword. In order to automatically obtain the training examples of a reference class, the keyword expansion (e.g. “red apple”) is used to retrieve images by the search engine based on textual information again. Images retrieved by the keyword expansion (“red apple”) are much less diverse than those retrieved by the original keyword (“apple”). After automatically removing outliers, the retrieved top images are used as the training examples of the reference class. Some reference classes (such as “apple laptop” and “apple macbook”) have similar semantic meanings and their training sets are visually similar. In order to improve the efficiency of online image re-ranking, redundant reference classes are removed. To better measure the similarity of semantic signatures, the semantic correlation between reference classes is estimated with a web-based kernel function.

For each query keyword, its reference classes forms the basis of its semantic space. A multi-class classifier on visual and textual features is trained from the training sets of its reference classes and stored offline. Under a query keyword, the semantic signature of an image is extracted by computing the similarities between the image and the reference classes of the query keyword using the trained multi-class classifier. If there are K types of visual/textual features, such as color, texture, and shape, one could combine them together to train a single classifier, which extracts one semantic signature for an image. It is also possible to train a separate classifier for each type of features. Then, the K classifiers based on different types of features extract K semantic signatures, which are combined at the later stage of image matching. Our experiments show that the latter strategy can increase the re-ranking accuracy at the cost of storage and online matching efficiency because of the increased size of semantic signatures.

According to the word-image index file, an image may be associated with multiple query keywords, which have different semantic spaces. Therefore, it may have different semantic signatures. The query keyword input by the user decides which semantic signature to choose. As an example shown in Figure 3, an image is associated with three keywords “apple”, “mac” and “computer”. When using any of the three keywords as query, this image will be retrieved and re-ranked. However, under different query keywords, different semantic spaces are used. Therefore an image could have several semantic signatures obtained in different semantic spaces. They all need to be computed and stored offline.

At the online stage, a pool of images are retrieved by the search engine according to the query keyword. Since all the images in the pool are associated with the query keyword according to the word-image index file, they all have pre-computed semantic signatures in the same semantic space specified by the query keyword. Once the user chooses a query image, these semantic signatures are used to compute image similarities for re-ranking. The semantic correlation of reference classes is incorporated when computing the similarities.

3.1 Discussion on Computational Cost and Storage

Compared with the conventional image re-ranking diagram in Figure 1, our approach is much more efficient at the online stage, because the main online computational cost is on comparing visual features or semantic signatures and the lengths of semantic signatures are much shorter than those of low-level visual features. For example, the visual features used in [6] are of more than $1,700$ dimensions. According to our experiments, each keyword has 25 reference classes on average. If only one classifier is trained combining all types of visual features, the semantic signatures are of 25 dimensions on average. If separate classifiers are trained for different types of visual features, the semantic signatures are of $100 - 200$ dimensions[3]. Our approach does not involve online training as required by pseudo relevance feedback [43]–[45]. It also provides much better re-ranking accuracy, since offline training the classifiers of reference classes captures the mapping between visual features and semantic meanings. Experiments show that semantic signatures are effective even if images do not belong to any of the found reference classes.

However, in order to achieve significant improvement of online efficiency and accuracy, our approach does need extra offline computation and storage, which come from collecting the training examples and reference classes, training the classifiers of reference classes and computing the semantic signatures. According to our experimental study, it takes 20 hours to learn the semantic spaces of 120 keywords using a machine with Intel Xeon W5580 3.2G CPU. The total cost linearly increases with the number of query keywords, which can be processed in parallel. Given 1000 CPUs[4], we will be able to process 100,000 query keywords in one day. With the fast growth of GPUs, it is feasible to process the industrial scale queries. The extra storage of classifiers and semantic signatures are comparable or even smaller than the storage of visual features of images. In order to periodically update the semantic spaces, one could repeat the offline steps. However, a more efficient way is to adopt the framework of incremental learning [72]. Our experimental studies show that the leaned semantic spaces are still effective without being updated for several months or even one year.

4 Discovery of Reference Classes

4.1 Keyword Expansion

For a keyword q, we define its reference classes by finding a set of keyword expansions $E(q)$ most relevant to q. To achieve this, a set of images $S(q)$ are retrieved by the search engine using q as query based on textual information. Keyword expansions are found from words extracted from images in $S(q)$[5], according to a very large dictionary used by the search engine. A keyword expansion $e \in E(q)$ is expected to frequently appear in $S(q)$. In addition, in order for reference classes to well capture the visual content of images, we require that there are subsets of images which all contain e and have similar visual content. Based on these considerations, $E(q)$ is found in a search-and-rank way as follows.

For each image $I \in S(q)$, all the images in $S(q)$ are re-ranked according to their visual similarities to I. Here, we use the visual features and visual similarities introduced in [6]. The T most frequent words $W_I = \{w_I^1, w_I^2, \cdots, w_I^T\}$ among top D re-ranked images (most visually similar to I) are found. $\{w_I^1, w_I^2, \cdots, w_I^T\}$ are sorted by the frequency of words appearing among the D images from large to small. If a word w is among the top ranked image, it has a ranking score $r_I(w)$ according to its ranking order; otherwise $r_I(w) = 0$,

$$r_I(w) = \begin{cases} T - j & w = w_I^j \\ 0 & w \notin W_I. \end{cases} \tag{1}$$

The overall score of a word w is its accumulated ranking scores over all the images,

$$r(w) = \sum_{I \in S} r_I(w). \tag{2}$$

A large $r_I(w)$ indicates that w appears in a good number of images visual similar to I. If w only exists in a small number of images or the images containing w are visually dissimilar to one another, $r_I(w)$ would be

[3] In our experiments, 120 query keywords are considered. But keyword expansions, which define reference classes, are from a very large dictionary used by the web search engine. They could be any words beyond the 120 ones. Different query keywords are processed independently. If more query keywords are considered, the dimensions of semantic signatures of each query keyword will not increase.

[4] Computational power of such a scale or even larger is used by industry. Jing and Baluja [21] used 1000 CPUs to process images offline.

[5] The words are extracted from filenames, ALT tags and surrounding text of images, after being stemmed and removing stop words.

zero for most I. Therefore, if w has a high accumulated ranking score $r(w)$, it should be found among a large number of images in $S(q)$ and some images with w are visually similar in the meanwhile. The P words with highest scores are selected to form the keyword expansions $E(q)$, which define the reference classes. We choose $T = 3$, $D = 16$, $P = 30$, and the size of $S(q)$ is $1,000$.

An intuitive way of finding keyword expansions could be first clustering images with visual/textual features and then finding the most frequent word in each cluster as the keyword expansion. We do not adopt this approach for two reasons. Images belonging to the same semantic concept (e.g. "apple laptop") have certain visual diversity (e.g. due to variations of viewpoints and colors of laptops). Therefore, one keyword expansion falls into several image clusters. Similarly, one image cluster may have several keyword expansions with high frequency, because some concepts have overlaps on images. For examples, an image may belong to "Paris Eiffel tower", "Paris nights" and "Paris Album". Since the one-to-one mapping between clusters and keyword expansions do not exist, a post processing step similar to our approach is needed to compute the scores of keywords selected from multiple clusters and fuse them. The multimodal and overlapping distributions of concepts can be well handled by our approach. Secondly, clustering web images with visual and textual features is not an easy task especially with the existence of many outliers. Bad clustering result greatly affects later steps. Since we only need keyword expansions, clustering is avoided in our approach. For each image I, our approach only considers its D nearest neighbors and is robust to outliers.

4.2 Training Images of Reference Classes

In order to automatically obtain the training images of reference classes, each keyword expansion e combined with the orginal keyword q is used as query to retrieve images from the search engine and top K images are kept. Since the expanded keywords e have less semantic ambiguity than the original keyword q, the images retrieved by e are much less diverse. After removing outliers by k-means clustering, these images are used as the training examples of the reference class. The cluster number of k-means is set as 20 and clusters of sizes smaller than 5 are removed as outliers.

4.3 Redundant Reference Classes

Some reference classes, e.g. "apple laptop" and "apple macbook", are pair-wisely similar in both semantics and visual appearance. To reduce computational cost we remove some redundant ones, which cannot increase the discriminative power of the semantic space. To compute the similarity between two reference classes, we use half data in both classes to train a binary SVM classifier to classify the other half data. If they can be easily separated, the two classes are considered not similar.

P reference classes are obtained from previous steps. The training images of reference class i are randomly split into two sets, A_i^1 and A_i^2. To measure the distinctness $D(i,j)$ between two reference classes i and j, a SVM is trained from A_i^1 and A_j^1. For each image in A_i^2, the SVM outputs a score of its probability of belonging to class i. Assume the average score over A_i^2 is $\bar{p}_i$. Similarly, the average score $\bar{p}_j$ over A_j^2 is also computed. Then

$$D(i,j) = h((\bar{p}_i + \bar{p}_j)/2), \tag{3}$$

where h is a monotonically increasing function. In our approach, it is defined as

$$h(\bar{p}) = 1 - e^{-\beta(\bar{p}-\alpha)}, \tag{4}$$

where β and α are two constants. When $(\bar{p}_i + \bar{p}_j)/2$ goes below the threshold α, $h(\bar{p})$ decreases very quickly so as to penalize pairwisely similar reference classes. We empirically choose $\alpha = 0.6$ and $\beta = 30$.

4.4 Reference Class Selection

We finally select a set of reference classes from the P candidates. The keyword expansions of the selected reference classes are most relevant to the query keyword q. The relevance is defined by Eq (2). Meanwhile, we require that the selected reference classes are dissimilar with each other such that they are diverse enough to characterize different aspects of its keyword. The distinctiveness is measured by the $P \times P$ matrix D defined in Section 4.3. The two criteria are simultaneously satisfied by solving the following optimization problem.

We introduce an indicator vector $y \in \{0,1\}^P$ such that $y_i = 1$ indicates reference class i is selected and $y_i = 0$ indicates it is removed. y is estimated by solving,

$$\arg \max_{y \in \{0,1\}^P} \left\{ \lambda R y + y^T D y \right\}. \tag{5}$$

Let e_i be the keyword expansion of reference class i. $R = (r(e_1), \ldots, r(e_P))$, where $r(e_i)$ is defined in Eq (2). λ is the scaling factor used to modulate the two criterions. Since integer quadratic programming is NP hard, we relax y to be in $\mathbb{R}^P$ and select reference classes i whose $y_i \geq 0.5$.

5 Semantic Signatures

Given M reference classes for keyword q and their training images, a multi-class classifier on the visual features of images is trained and it outputs an M-dimensional vector p, indicating the probabilities of a new image I belonging to different reference classes. p is used as the semantic signature of I. The distance between two images I^a and I^b are measured as the L_1-distance between their semantic signatures p^a and p^b,

$$d(I^a, I^b) = \left\| p^a - p^b \right\|_1. \tag{6}$$

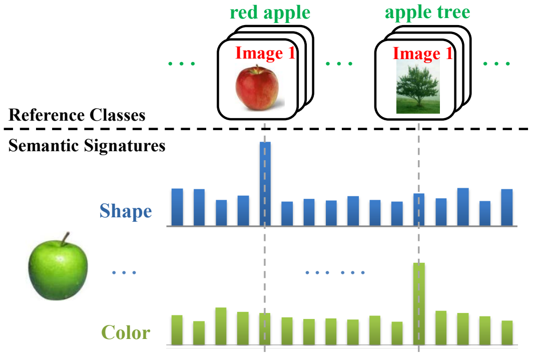


Fig. 4. Describe "green apple" with reference classes. Its shape is captured by the shape classifier of "red apple" and its color is captured by the color classifier of "apple tree".

5.1 Combined Features vs Separate Features

In order to train the SVM classifier, we adopt six types of visual features used in [6]: attention guided color signature, color spatialet, wavelet [73], multi-layer rotation invariant edge orientation histogram, histogram of oriented gradients [37], and GIST [74]. They characterize images from different perspectives of color, shape, and texture. The total dimensionality around $1,700$.

A natural idea is to combine all the visual features to train a single powerful SVM better distinguishing reference classes. However, the purpose of using semantic signatures is to capture the visual content of an image, which may belong to none of the reference classes, instead of classifying it into one of the reference classes. If there are K types of independent visual features, it is more effective to train separate SVM classifiers on different types of features and to combine the K semantic signatures $\{p^k\}_{k=1}^K$ from the outputs of the K classifiers. The K semantic signatures describe the visual content from different aspects (e.g. color, texture, and shape) and can better characterize images outside the reference classes. For example, in Figure 4, "red apple" and "apple tree" are two reference classes. A new image of "green apple" can be well characterized by two semantic signatures from two classifiers trained on color features and shape features separately, since "green apple" is similar to "red apple" in shape and similar to "apple tree" in color. If the color and shape features are combined to compute a single semantic signature, it cannot well characterize the image of "green apple". Since the "green apple" is dissimilar to any reference class when jointly considering color and shape, the semantic signature has low distributions over all the reference classes.

Then the distance between two images I^a and I^b is,

$$d(I^a, I^b) = \sum_{k=1}^{K} w_k \left\| p^{a,k} - p^{b,k} \right\|_1, \tag{7}$$

where w_k is the weight on different semantic signatures and it is specified by the query image I^a selected by the user. w_k is decided by the entropy of $p^{a,k}$,

$$w_k = \frac{1}{1 + e^{H(p^{a,k})}}, \tag{8}$$

$$H(p^{a,k}) = -\sum_{i=1}^{M} p_i^{a,k} \ln p_i^{a,k}. \tag{9}$$

If $p^{a,k}$ uniformly distributes over reference classes, the kth type of visual features of the query image cannot be well characterized by any of the reference classes and we assign a low weight to this semantic signature.

5.2 Incorporating Textual Features

Our approach provides a natural way to integrate visual and textual features. Semantic signatures can also be computed from textual features and combined with those from visual features. Visual and textual features are in different modalities. However, after projecting into the same semantic space, they have the same representation. The semantic signatures from textual features are computed as follows. Let $E = \{d_i, \ldots, d_m\}$ be the training examples of a reference class. d_i contains the words extracted from image i and is treated as a document. In principle, any document classifier can be used here. We adopt a state-of-the-art approach proposed in [45] to learn a word probability model $p(w|\theta)$, which is a discrete distribution, from E. θ is the parameter of the discrete distribution of words over the dictionary and it is learned by maximizing the observed probability,

$$\prod_{d_i \in E} \prod_{w \in d_i} (0.5p(w|\theta) + 0.5p(w|C))^{n_w^i}. \tag{10}$$

n_w^i is the frequency of word w in d_i, and $p(w|C)$ is the word probability built upon the whole repository C,

$$p(w|C) = \frac{\sum_{d_i} n_w^i}{|C|}. \tag{11}$$

Once θ is learned with EM, the textual similarity between an image d_j and the reference class is defined as

$$\sum_{w \in d_j} p(w|\theta) n_w^j. \tag{12}$$

After normalization, the similarities to reference classes are used as semantic signatures.

5.3 Incorporating Semantic Correlations

Eq (6) matches two semantic signatures along each dimension separately. It assumes the independency between reference classes, which are in fact semantically related. For example, "apple macbook" is more related to "apple ipad" than to "apple tree". This indicates that in order to more reasonably compute image similarities, we should take into account such semantic correlations, and allow one dimension in the semantic signature (e.g. "apple macbook") to match with its correlated dimensions (e.g. "apple ipod"). We further improve the image similarity proposed in Eq(6) with a bilinear form,

$$s(I_a, I_b) = \sum_{i,j} p^{a,i} C_{ij} p^{b,j} = p^{aT} C p^b, \quad (13)$$

where C is an M by M matrix, whose (i, j)th entry C_{ij} denotes the strength of semantic correlation between the ith and jth reference classes. If multiple semantic signatures are used, we compute the bilinear similarity on each type of semantic signatures and combine them using the same weights as in Eq (7).

We adopt the web-based kernel function [75] to compute the semantic correlations between reference classes. For each reference class i, the expanded keywords $e_i + q$ is used as an input to the Google web search, and the top 50 Google snippets[6], denoted as $S(e_i)$, are collected. After removing the original keyword q from the snippets, the term frequency (TF) vector of $S(e_i)$ is computed, and the top 100 terms with the highest TFs in $S(e_i)$ are reserved. Each e_i has a different set of top 100 terms. We L2-normalize the truncated vector, and denote the result vector as $ntf(e_i)$. The dimensionality of $ntf(e_i)$ is equal to the size of the dictionary. However, only the top 100 terms of e_i with highest TFs have non-zero values. The semantic correlation between the ith and jth reference classes, i.e. e_i and e_j, is computed as

$$C_{i,j} = \text{cosine}(ntf(e_i), ntf(e_j)). \quad (14)$$

6 Experimental Results

The images for testing the performance of re-ranking and the training images of reference classes can be collected at different time (since the update of reference classes may be delayed) and from different search engines. Given a query keyword, 1000 images are retrieved from the whole web using a search engine. As summarized in Table 1, we create three data sets to evaluate the performance of our approach in different scenarios. In data set I, $120,000$ testing images for re-ranking were collected from the Bing Image Search with 120 query keywords in July 2010. These query keywords cover diverse topics including animals, plants, food, places, people, events, objects, and scenes, etc. The training images of reference classes were also collected from the Bing Image Search around the same time. Data set II uses the same testing images as in data set I. However, its training images of reference classes were collected from the Google Image Search also in July 2010. In data set III, both testing and training images were collected from the Bing Image Search but at different time (eleven months apart)[7]. All the testing images for re-ranking are manually labeled, while the images of reference classes, whose number is much larger, are not labeled.

6.1 Re-ranking precisions

We invited five labelers to manually label testing images under each query keyword into different categories according to semantic meanings. Image categories were carefully defined by the five labelers through inspecting all the testing images under a query keyword. Defining image categories was completely independent of discovering reference classes. The labelers were unaware of what reference classes have been discovered by our system. The number of image categories is also different than the number of reference classes. Each image was labeled by at least three labelers and its label was decided by voting. Some images irrelevant to query keywords were labeled as outliers and not assigned to any category.

Averaged top m precision is used as the evaluation criterion. Top m precision is defined as the proportion of relevant images among top m re-ranked images. Relevant images are those in the same category as the query image. Averaged top m precision is obtained by averaging over all the query images. For a query keyword, each of the $1,000$ images retrieved only by keywords is used as a query image in turn, excluding outlier images. We do not adopt the precision-recall curve, since in image re-ranking the users are more concerned about the qualities of top ranked images instead of the number of relevant images returned in the whole result set.

We compare with two image re-ranking approaches used in [6], which directly compare visual features, and two approaches of pseudo-relevance feedback [43], [44], which online learns visual similarity metrics.

- **Global Weighting**. Fixed weights are adopted to fuse the distances of different visual features [6].
- **Adaptive Weighting**. [6] proposed adaptive weights for query images to fuse the distances of different visual features. It is adopted by Bing Image Search.
- **PRF**. The pseudo-relevance feedback approach proposed in [43]. It used top-ranked images as positive examples to train a one-class SVM .
- **NPRF**. The pseudo-relevance feedback approach proposed in [44]. It used top-ranked images as positive examples and bottom-ranked images as negative examples to train a SVM.

For our approach, two different ways of computing semantic signatures in Section 5.1 are compared.

- *Query-specific visual semantic space using single signatures (**QSVSS_Single**)*. For an image, a single semantic signature is computed from one SVM classifier trained by combining all types of visual features.
- *Query-specific visual semantic space using multiple signatures (**QSVSS_Multiple**)*. For an image, multiple semantic signatures are computed from multiple SVM classifiers, each of which is trained on one type of visual features separately.

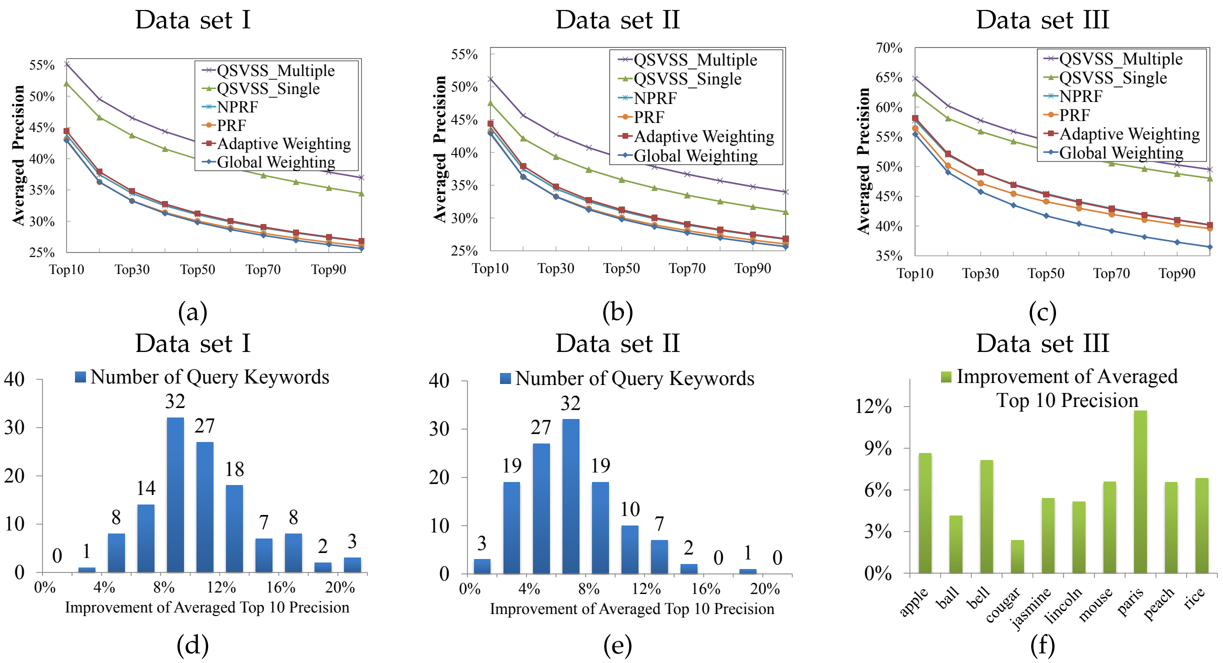
QSVSS_Single and QSVSS_Multiple do not use textual features to compute semantic signatures and do not incorporate semantic correlation between reference classes. The two improvements are evaluated in Section 6.5 and

[6] Google snippet is a short summary generated by Google for each search result item in response to the query.

[7] It would be closer to the scenario of real applications if test images were collected later than the images of reference classes. However, such data set is not available for now. Although data set III is smaller than data set I, it is comparable with the data set used in [6].

TABLE 1. Descriptions of data sets

Data set	Images for re-ranking				Images of reference classes	
	# Keywords	# Images	Collecting date	Search engine	Collecting date	Search engine
I	120	120,000	July 2010	Bing Image Search	July 2010	Bing Image Search
II					July 2010	Google Image Search
III	10	10,000	August 2009	Bing Image Search	July 2010	Bing Image Search



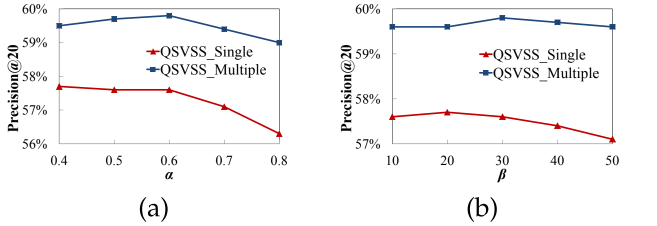
(a) (b) (c)

(d) (e) (f)

Fig. 5. (a)-(c): averaged top m precisions on data set I, II, III. (d)-(e): histograms of improvements of averaged top 10 precisions on data sets I and II by comparing QSVSS_Multiple with Adaptive Weighting. (f): improvements of averaged top 10 precisions on the 10 query keywords on data set III by comparing QSVSS_Multiple with Adaptive Weighting.

6.6. The visual features in all the six approaches above are the same as [6]. The parameters of our approaches mentioned in Section 4 and 5 are tuned in a small separate dataset and fixed in all the experiments.

The averaged top m precisions on data sets I-III are shown in Figure 5 (a)-(c). Our approach significantly outperforms Global Weighting and Adaptive Weighting, which directly compare visual features. On data set I, the averaged top 10 precision is enhanced from 44.41% (Adaptive Weighting) to 55.12% (QSVSS_Multiple). 24.1% relative improvement is achieved. Figure 5 (d) and (e) show the histograms of improvements of averaged top 10 precision of the 120 query keywords on data set I and II by comparing QSVSS_Multiple with Adaptive Weighting. Figure 5 (f) shows the improvements on the 10 query keywords on data set III. Our approach also outperforms pseudo-relevance feedback.



(a) (b)

Fig. 6. Averaged top 20 precisions on Data set III when (a) choosing different α while fixing $\beta = 30$; and (b) choosing different β while fixing $\alpha = 0.6$.

Computing multiple semantic signatures from separate visual features has higher precisions than computing a single semantic signature from combined features. It costs more online computation since the dimensionality of multiple semantic signatures is higher. Figure 6 shows the sensitivity of QSVSS_Multiple and QSVSS_Single to the choice of parameters α and β in Eq. (4). They are robust in certain ranges. Comparing Figure 5 (a) and (b), if the testing images for re-ranking and the images of reference classes are collected from different search engines, the performance is slightly lower than the case when they are collected from the same search engine. But it is still much higher than matching visual features. It indicates that we can utilize images from various sources to learn query-specific semantic spaces. As shown in Figure 5 (c), even if the testing images and the images of reference classes are collected eleven months apart, query-specific semantic spaces are still effective. Compared with Adaptive Weighting, the averaged top 10 precision has been improved by 6.6% and the averaged top 100 precision has been improved by 9.3%. This indicates that once the query-specific semantic spaces are learned, they can remain effective for a long time.

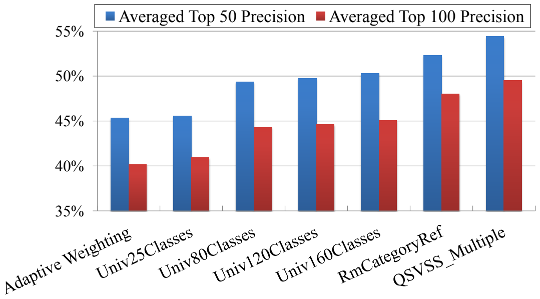


Fig. 7. Comparisons of averaged top m precisions of re-ranking images outside reference classes and using universal semantic space on data set III.

6.2 Online efficiency

The online computational cost depends on the length of visual feature (if matching visual features) or semantic signatures (if using our approach). In our experiments, the visual features have around $1,700$ dimensions, and the averaged number of reference classes per query is 25. Thus the length of QSVSS_Single is 25 on average. Since six types of visual features are used, the length of QSVSS_Multiple is 150. It takes 12ms to re-rank 1000 images matching visual features, while QSVSS_Multiple and QSVSS_Single only need 1.14ms and 0.2ms. Given the large improvement on precisions, our approach also improves the efficiency by 10 to 60 times.

6.3 Re-ranking images outside reference classes

It is interesting to know whether the query-specific semantic spaces are effective for query images outside reference classes. We design an experiment to answer this question. If the category of an query image corresponds to a reference class, we deliberately delete this reference class and use the remaining reference classes to train SVM classifiers and to compute semantic signatures when comparing this query image with other images. We repeat this for every image and calculate the average top m precisions. This evaluation is denoted as **RmCategoryRef** and is done on data set III[8]. QSVSS_Multiple is used. The results are shown in Figure 7. It still greatly outperforms the approaches of directly comparing visual features. It can be explained from two aspects. (1) As discussed in Section 5.1, QSVSS_Multiple can characterize the visual content of images outside reference classes. (2) Many negative examples (belonging to different categories than the query image) are well modeled by the reference classes and are therefore pushed backward on the ranking list. Therefore query-specific semantic spaces are effective for query images outside reference classes.

6.4 Query-specific vs. universal semantic spaces

In previous works [9]–[11], [54], [70], a universal set of reference classes or concepts were used to map visual features to a semantic space for object recognition or image retrieval on closed databases. We evaluate whether it is applicable to web-based image re-ranking. We randomly select M reference classes from the whole set of reference classes of all the 120 query keywords in data set I. The M selected reference classes are used to train a universal semantic space in a way similar to Section 5.1. Multiple semantic signatures are obtained from different types of features separately. This universal semantic space is applied to data set III. The averaged top m precisions are shown in Figure 7. M is chosen as 25, 80, 120 and 160[9]. This method is denoted as **Univ*M*Classes**. When the universal semantic space chooses the same number (25) of reference classes as our query-specific semantic spaces, its precisions are no better than visual features. Its precisions increase when a larger number of reference classes are selected. However, the gain increases very slowly when M is larger than 80. Its best precisions (when $M = 160$) are much lower than QSVSS_Multiple and RmCategoryRef, even though the length of its semantic signatures is five times larger.

6.5 Incorporating textual features

As discussed in Section 5.2, semantic signatures can be computed from textual features and combined with those from visual features. Figure 8 compares the averaged top m precisions of QSVSS_Multiple with

- *Query-specific textual and visual semantic space using multiple signatures (**QSTVSS_Multiple**).* For an image, multiple semantic signatures are computed from multiple classifiers, each of which is trained on one type of visual or textual features separately.
- *Textual feature alone (**Text**).* The cross-entropy between the word histograms of two images is used to compute the similarity.

It shows that incorporating textual features into the computation of semantic signatures significantly improves the performance. Moreover, the weights of combining visual semantic signatures and textual semantic signatures can be automatically decided by Eq (8).

6.6 Incorporating Semantic Correlations

As discussed in Section 5.3, we can further incorporate semantic correlations between reference classes when computing image similarities. For each type of semantic signatures obtained above, i.e., QSVSS_Single, QSVSS_Multiple, and QSTVSS_Multiple, we compute the image similarity with Eq (13), and name the corresponding results as QSVSS_Single_Corr, QSVSS_Multiple_Corr, and QSTVSS_Multiple_Corr respectively. Fig. 9 shows the re-ranking precisions for all types of semantic signatures on the three data sets. Notably, QSVSS_Single_Corr achieves around 10% relative improvement compared with QSVSS_Single, reaching the performance of QSVSS_Multiple despite its signature is six times shorter.

[8] We did not test on dataset I or II since it is very time consuming. For every query image, SVM classifiers have to be re-trained.

[9] We stop evaluating larger M because training a multi-class SVM classifier on hundreds of classes is time consuming.

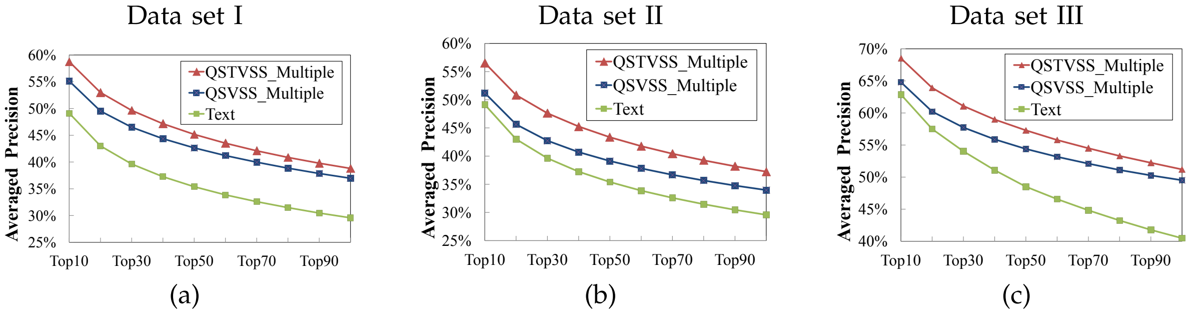


Fig. 8. Averaged top m precisions incorporating textual features.

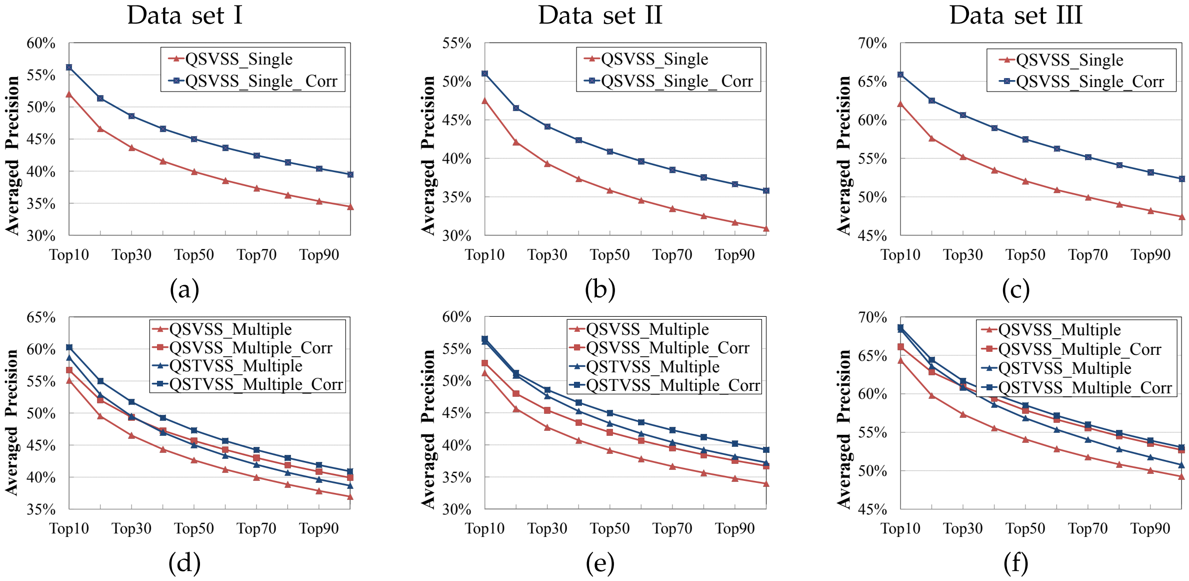


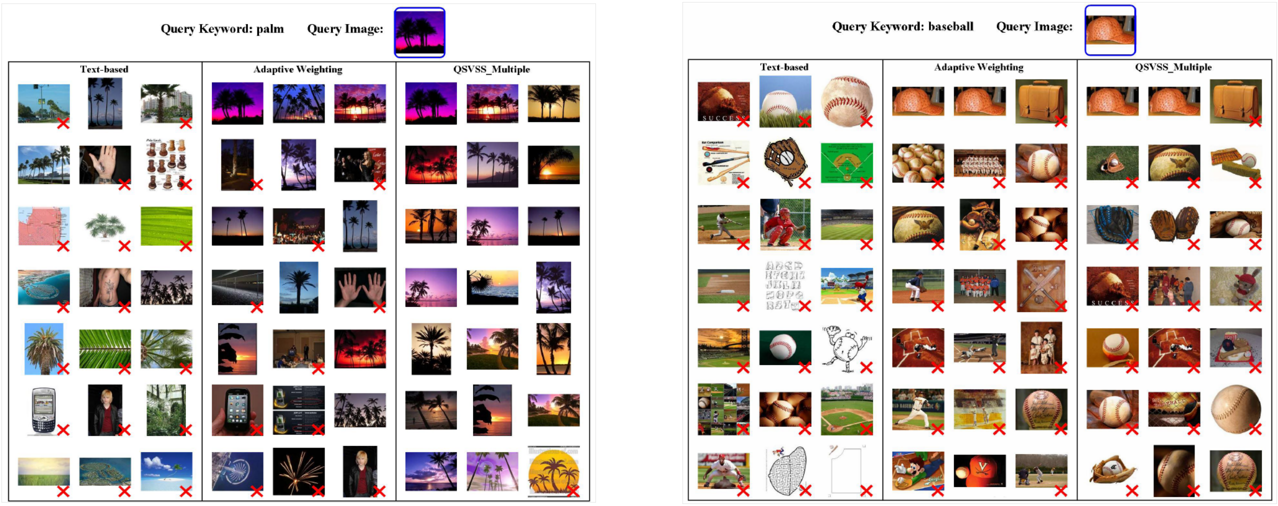
Fig. 9. Incorporating semantic correlations among reference classes. (a)-(c): single visual semantic signatures with/without sematic correlation. (d)-(f): multiple visual & textual semantic signatures with/without sematic correlation.

6.7 User study

In order to fully reflect the extent of users' satisfaction, user study is conducted to compare the results of QSVSS_Multiple[10] and Adaptive Weighting on data set I. Twenty users are invited. Eight of them are familiar with image search and the other twelve are not. We ensure that all the participants do not have any knowledge about current approaches for image re-ranking, and they are not told which results are from which methods. Each user is assigned 20 queries and is asked to randomly select 30 images per query. Each selected image is used as a query image and the re-ranking results of Adaptive Weighting and our approach are shown to the user. The user is required to indicate whether our re-ranking result is "Much Better", "Better", "Similar", "Worse", or "Much Worse" than that of Adaptive Weighting. The evaluation criteria are (1) the top ranked images belong to the same semantic category as the query image; and (2) candidate images which are more visual similar to the query image have higher ranks. $12,000$ user comparison results are collected and shown in Figure 10. In over 55% cases our approach delivers better results. Ours is worse only in fewer than 18% cases, which are often the noisy cases with few images relevant to the query image.

Figure 11 (a) shows an example that QSVSS_Multiple provides much better results. The query keyword is "palm". The initial text-based search returns a pool of images with diverse semantic meanings, such as palm cell phones, palm centro and hands. The selected query image is about palm trees on beach. After re-ranking, QSVSS_Multiple returns many images which have large variance in visual content but are relevant to the query image in semantic meanings. These images cannot be found by directly matching visual features. Figure 11 (b) shows an example that QSVSS_Multiple provides worse results than Adaptive Weighting according to the user study. Actually, in this example there are very few images in the image pool relevant to the query image, which can be regarded as an outlier. Both approaches provide bad results. The user prefers the result of Adaptive Weighting perhaps because its result is more diverse, although not many relevant images are found either. Please find more examples in supplementary material.

[10] Since Adaptive Weighting only uses visual features, to make the comparison fair, textual features are not used to compute semantic signatures and sematic correlation between classes is not considered.



(a)

(b)

Fig. 11. Examples of results of initial text-based search, image re-ranking by Adaptive Weighting [6] and by QSVSS_Multiple. The red crosses indicate the images irrelevant to the query image. Examples that QSVSS_Multiple has a better (a) or worse (b) result than Adaptive Weighting according to the user study.

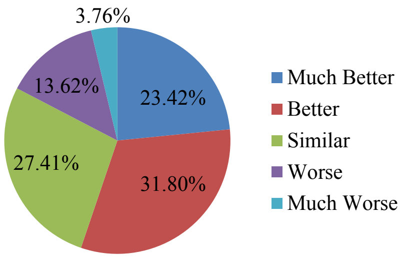


Fig. 10. Comparison results of user study on data set I.

7 Re-ranking without Query Images

Query-specific semantic signature can also be applied to image re-ranking without selecting query images. This application also requires the user to input a query keyword. But it assumes that images returned by initial text-only search have a dominant topic and images belonging to that topic should have higher ranks. A lot of related works are discussed in the third paragraph in Section 2. Existing approaches typically address two issues: (1) how to compute the similarities between images and reduce the semantic gap; and (2) how to find the dominant topic with ranking algorithms based on the similarities. Our query-specific semantic signature is effective in this application since it can improve the similarity measurement of images. In this experiment QSVSS_Multiple is used to compute similarities. We compare with the state-of-the-art methods on the public MSRA-MM V1.0 dataset [33]. This dataset includes 68 diverse yet representative queries collected from the query log of Bing, and contains $60,257$ images. Each image was manually labeled into three relevance levels and the Normalized Discounted Cumulated Gain (NDCG) [28] is used as the standard evaluation metric. NDCG at rank m is calculated as NDCG@m $= \frac{1}{Z}\sum_{j=1}^{m}\frac{2^{t_j}-1}{\log(1+j)}$, where t_j is the relevance level the jth image in the rank list and Z is a normalization constant to make NDCG@m be 1 for a perfect ranking. We adopt three re-ranking approaches by keeping their ranking algorithms while replacing their features with our query-specific semantic signatures: random walk (RWalk) [17], kernel-based re-ranking by taking top N images as confident samples (KernelTopN) [28], and kernel-based re-ranking by detecting confident samples based on bounded variable least square (KernelBVLS) [28]. The details of these ranking algorithms can be found in literature. Table 2 reports NDCG@m of initial text result, the three original approaches in [17], [28], their corresponding versions with our query-specific sematic signatures, Information Bottleneck (IB) [15] and Bayesian Visual Ranking (Bayesian) [24]. The NDCG@m improvements of these approaches over initial result are shown in Figure 12. It is observed that our query-specific semantic signatures are very effective. Compared with the initial result, the NDCG@m improvements of the three approaches in [17], [28] are 0.007, 0.008 and 0.021, while the improvements become 0.029, 0.052 and 0.067 when their features are placed with query-specific semantic signatures.

8 Conclusion and Future Work

We propose a novel framework, which learns query-specific semantic spaces to significantly improve the effectiveness and efficiency of online image re-ranking. The visual features of images are projected into their related semantic spaces automatically learned through keyword expansions offline. The extracted semantic signatures can be 70 times shorter than the original visual features, while achieve $25\% - 40\%$ relative improvement on re-ranking precisions over state-of-the-art methods.

In the future work, our framework can be improved along several directions. Finding the keyword expansions used to define reference classes can incorporate

TABLE 2. Performance of image re-ranking without selecting query images on the MSRA-MM V1.0 dataset. The values in the parentheses are the NDCG@m improvements over initial search.

	Initial	IB [15]	Bayesian [24]	RWalk [17]		KernelTopN [28]		KernelBVLS [28]	
				original	ours	original	ours	original	ours
NDCG@10	0.582	0.594(0.012)	0.592(0.010)	0.597(0.015)	**0.617(0.035)**	0.598(0.016)	**0.640(0.058)**	0.645(0.063)	**0.655(0.073)**
NDCG@50	0.556	0.570(0.014)	0.565(0.009)	0.563(0.007)	**0.585(0.029)**	0.564(0.008)	**0.608(0.053)**	0.587(0.031)	**0.623(0.067)**
NDCG@100	0.536	0.554(0.018)	0.559(0.023)	0.550(0.014)	**0.580(0.044)**	0.560(0.024)	**0.605(0.069)**	0.580(0.044)	**0.618(0.082)**

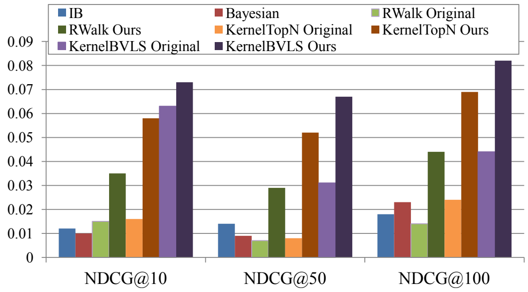


Fig. 12. The NDCG@m improvements over initial search, i.e. the difference between NDCG@m after re-ranking and that without re-ranking.

other metadata and log data besides the textual and visual features. For example, the co-occurrence information of keywords in user queries is useful and can be obtained in log data. In order to update the reference classes over time in an efficient way, how to adopt incremental learning [72] under our framework needs to be further investigated. Although the semantic signatures are already small, it is possible to make them more compact and to further enhance their matching efficiency using other technologies such as hashing [76].

Acknowledgement

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] R. Datta, D. Joshi, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, 2007.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval content-based image retrieval," *IEEE Trans. on PAMI*, vol. 22, p. 1349, 2000.

[3] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Trans. on CSVT*, 1998.

[4] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Systems*, 2003.

[5] 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. on PAMI*, 2006.

[6] J. Cui, F. Wen, and X. Tang, "Real time google and live image search re-ranking," in *Proc. ACM Multimedia*, 2008.

[7] ——, "Intentsearch: Interactive on-line image search re-ranking," in *Proc. ACM Multimedia*, 2008.

[8] X. Tang, K. Liu, J. Cui, F. Wen, and X. Wang, "Intentsearch:capturing user intention for one-click internet image search," *IEEE Trans. on PAMI*, vol. 34, pp. 1342–1353, 2012.

[9] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Trans. on Multimedia*, 2007.

[10] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. CVPR*, 2009.

[11] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *Proc. CVPR*, 2011.

[12] A. Kovashka, D. Parikh, and K. Grauman, "Whittlesearch: Image search with relative attribute feedback," in *Proc. CVPR*, 2012.

[13] R. Fergus, P. Perona, and A. Zisserman, "A visual category filter for google images," in *Proc. ECCV*, 2004.

[14] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *Proc. ICCV*, 2005.

[15] W. Hsu, L. Kennedy, and S. F. Chang, "Video search reranking via information bottleneck principle," in *Proc. ACM Multimedia*, 2006.

[16] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," in *Proc. ICCV*, 2007.

[17] W. Hsu, L. Kennedy, and S. F. Chang, "Video search reranking through random walk over document-level context graph," in *Proc. ACM Multimedia*, 2007.

[18] M. Fritz and B. Schiele, "Decomposition, discovery and detection of visual categories using topic models," in *Proc. CVPR*, 2008.

[19] T. Berg and D. Forsyth, "Animals on the web," in *Proc. CVPR*, 2008.

[20] D. Grangier and S. Bengio, "A discriminative kernel-based model to rank images from text queries," *IEEE Trans. on PAMI*, vol. 30, pp. 1371–1384, 2008.

[21] Y. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *IEEE Trans. on PAMI*, vol. 30, pp. 1877–1890, 2008.

[22] L. Yang and A. Hanjalic, "Supervised reranking for web image search," in *Proc. ACM Multimedia*, 2010.

[23] L. Chen, D. Xu, I. Tsang, and J. Luo, "Tag-based web photo retrieval improved by batch mode re-tagging," in *Proc. CVPR*, 2010.

[24] X. Tian, L. Yang, J. Wang, X. Wu, and X. Hua, "Bayesian visual reranking," *IEEE Trans. on Multimedia*, vol. 13, pp. 639–652, 2010.

[25] B. Geng, L. Yang, C. Xu, and X. Hua, "Content-aware ranking for visual search," in *Proc. CVPR*, 2010.

[26] J. Krapac, M. Allan, J. Verbeek, and F. Jurie, "Improving web image search results using query-relative classifiers," in *Proc. CVPR*, 2010.

[27] W. Liu, Y. Jiang, J. Luo, and F. Chang, "Noise resistant graph ranking for improvedweb image search," in *Proc. CVPR*, 2011.

[28] N. Morioka and J. Wang, "Robust visual reranking via sparsity and ranking constraints," in *Proc. ACM Multimedia*, 2011.

[29] V. Jain and M. Varma, "Learning to re-rank: Query-dependent image re-ranking using click data," in *Proc. WWW*, 2011.

[30] J. Huang, X. Yang, X. Fang, and R. Zhang, "Integrating visual saliency and consistency for re-ranking image search results," *IEEE Trans. on Multimedia*, vol. 13, pp. 653–661, 2011.

[31] J. Lu, J. Zhou, J. Wang, X. Hua, and S. Li, "Image search results refinement via outlier detection using deep contexts," in *Proc. CVPR*, 2012.

[32] J. Cai, Z. Zha, W. Zhou, and Q. Tian, "Attribute-assisted reranking for web image retrieval," in *Proc. ACM Multimedia*, 2012.

[33] M. Wang, L. Yang, and X. Hua, "Msra-mm: Bridging research and industrial societies for multimedia information retrieval," Microsoft Research Asia, Tech. Rep., 2009.

[34] T. Deselaers, T. Gass, P. Dreuw, and H. Hey, "Jointly optimising relevance and diversity in image retrieval," in *Proc. ACM Int'l Conf. Image and Video Retrieval*, 2009.

[35] J. J. Foo, J. Zobel, and R. Sinha, "Clustering near-duplicate images in large collections," in *Proc. the Int'l Workshop on Multimedia Information Retrieval*, 2007.

[36] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int'l Journal of Computer Vision*, 2004.

[37] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. CVPR*, 2005.

[38] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang, “Spatial-bag-of-features,” in *Proc. CVPR*, 2010.

[39] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, “Descriptor Learning for Efficient Retrieval,” in *Proc. ECCV*, 2010.

[40] Y. Qiao, W. Wang, N. Minematsu, J. Liu, M. Takeda, and X. Tang, “A theory of phase singularities for image representation and its applications to object tracking and image matching,” *IEEE Trans. on Image Processing*, vol. 18, pp. 2153–2166, 2009.

[41] X. Wang, K. Liu, and X. Tang, “Query-specific visual semantic spaces for web image re-ranking,” in *Proc. CVPR*, 2010.

[42] Y. Kuo, W. Cheng, H. Lin, and W. Hsu, “Unsupervised semantic feature discovery for image object retrieval and tag refinement,” *IEEE Trans. on Multimedia*, vol. 14, pp. 1079–1090, 2012.

[43] R. Yan, E. Hauptmann, and R. Jin, “Multimedia search with pseudo-relevance feedback,” in *Proc. of Int'l Conf. on Image and Video Retrieval*, 2003.

[44] R. Yan, A. G. Hauptmann, and R. Jin, “Negative pseudo-relevance feedback in content-based video retrieval,” in *Prof. ACM Multimedia*, 2003.

[45] 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.

[46] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. zisserman, “Total recall: Automatic query expansion with a generative feature model for object retrieval,” in *Proc. CVPR*, 2007.

[47] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in *Proc. CVPR*, 2008.

[48] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, “Total recall ii: Query expansion revisited,” in *Proc. CVPR*, 2011.

[49] V. Ferrari and A. Zisserman, “Learning visual attributes,” in *Proc. NIPS*, 2007.

[50] V. Sharmanska, N. Quadrianto, and C. H. Lampert, “Augmented attribute representations,” in *Proc. ECCV*, 2012.

[51] A. Farhadi, I. Endres, and D. Hoiem, “Attribute-centric recognition for cross-category generalization,” in *Proc. CVPR*, 2010.

[52] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, “Describing objects by their attributes,” in *Proc. CVPR*, 2009.

[53] G. Wang and D. Forsyth, “Joint learning of visual attributes, object classes and visual saliency,” in *Proc. ICCV*, 2009.

[54] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, “What helps wherevand why? semantic relatedness for knowledge transfer,” in *Proc. CVPR*, 2010.

[55] Y. Wang and G. Mori, “A discriminative latent model of object classes and attributes,” in *Proc. ECCV*, 2010.

[56] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie, “Visual recognition with humans in the loop,” in *Proc. ECCV*, 2010.

[57] S. Hwang, F. Sha, and K. Grauman, “Sharing features between objects and their attributes,” in *Proc. CVPR*, 2011.

[58] D. Parikh and K. Grauman, “Relative attributes,” in *Proc. ICCV*, 2011.

[59] D. Mahajan, S. Sellamanickam, and V. Nair, “A joint learning framework for attribute models and object recognition,” in *Proc. ICCV*, 2011.

[60] A. Parkash and D. Parikh, “Attributes for classifier feedback,” in *Proc. ECCV*, 2012.

[61] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, “Attribute and simile classifiers for face verification.” in *Proc. ICCV*, 2009.

[62] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult, “Multi-attribute spaces: Calibration for attribute fusion and similarity search,” in *Proc. CVPR*, 2012.

[63] N. Kumar, P. Belhumeur, and S. Nayar, “A search engine for large collections of images with faces,” in *Proc. ECCV*, 2008.

[64] T. Berg, A. Berg, and J. Shih, “Automatic attribute discovery and characterization from noisy web data,” in *Proc. ECCV*, 2010.

[65] B. Siddiquie, S. Feris, and L. Davis, “Image ranking and retrieval based on multi-attribute queries,” in *Proc. CVPR*, 2011.

[66] M. Douze, A. Ramisa, and C. Schmid, “Combining attributes and fisher vectors for efficient image retrieval,” in *Proc. CVPR*, 2011.

[67] F. X. Yu, R. Ji, M. Tsai, G. Ye, and S. Chang, “Weak attributes for large-scale image search,” in *Proc. CVPR*, 2012.

[68] J. Lui, B. Kuipers, and S. Savarese, “Recognizing human actions by attributes,” in *Proc. CVPR*, 2011.

[69] B. Gong, J. Liu, X. Wang, and X. Tang, “3d object retrieval with semantic attributes,” in *Proc. ACM Multimedia*, 2011.

[70] E. Bart and S. Ullman, “Single-example learning of novel classes using representation by similarity,” in *Proc. BMVC*, 2005.

[71] L. Torresani, M. Szummer, and A. Fitzgibbon, “Efficient object category recognition using classemes,” in *Proc. ECCV*, 2010.

[72] G. Cauwenberghs and T. Poggio, “Incremental and decremental support vector machine learning,” in *Proc. NIPS*, 2001.

[73] M. Unser, “Texture classification and segmentation using wavelet frames,” *IEEE Trans. on image processing*, no. 11, 1995.

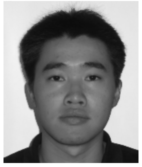
[74] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, “Context-based vision system for place and object recognition,” in *Proc. ICCV*, 2003.

[75] M. Sahami and T. D. Heilman, “A web-based kernel function for measuring the similarity of short text snippets,” in *Proc. WWW*, 2006.

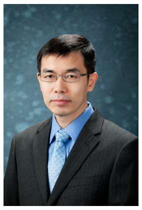
[76] A. Andoni and P. Indyk, “Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions,” *Communications of the ACM*, vol. 51, pp. 117–122, 2008.

Xiaogang Wang (S'03-M'10) received the B.S. degree from University of Science and Technology of China in 2001, the M.S. degree from 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. He received the Outstanding Young Researcher Award in Automatic Human Behaviour Analysis in 2011, and the Hong Kong Early Career Award in 2012. His research interests include computer vision and machine learning.

Shi Qiu received the B.S. degree in Electronic Engineering from Tsinghua University, China, in 2009. He is currently a PhD student in the Department of Information Engineering at the Chinese University of Hong Kong. His research interests include image search and computer vision.



Ke Liu received the B.S. degree in Computer Science from Tsinghua University in 2009. He is currently an M.Phil. student in the Department of Information Engineering at the Chinese University of Hong Kong. His research interests include image search and computer vision.



Xiaoou Tang (S'93-M'96-SM'02-F'09) received the B.S. degree from the University of Science and Technology of China, Hefei, in 1990, and the M.S. degree from the University of Rochester, Rochester, NY, in 1991. He received the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1996.

He is a Professor in the Department of Information Engineering and 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 the Microsoft Research Asia from 2005 to 2008. His research interests include computer vision, pattern recognition, and video processing.

Dr. Tang received the Best Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009. He is a program chair of the IEEE International Conference on Computer Vision (ICCV) 2009 and an Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) and International Journal of Computer Vision (IJCV). He is a Fellow of IEEE.