Query-Specific Visual Semantic Spaces for Web Image Re-ranking

Xiaogang Wang[1]
[1] Department of Electronic Engineering
The Chinese University of Hong Kong
xgwang@ee.cuhk.edu.hk

Ke Liu[2]
[2] Department of Information Engineering
The Chinese University of Hong Kong
lk009@ie.cuhk.edu.hk

Xiaoou Tang[2,3]
[3] Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences
xtang@ie.cuhk.edu.hk

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. Given a query keyword, a pool of images are first retrieved by the search engine 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. On the other hand, 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 visual semantic spaces for different query keywords through keyword expansions. The visual features of images are projected into their related visual semantic spaces to get semantic signatures. At the online stage, images are re-ranked by comparing their semantic signatures obtained from the visual semantic space specified by the query keyword. The new approach significantly improves 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* $20\% - 35\%$ *relative improvement has been achieved on re-ranking precisions compared with the state-of-the-art methods.*

1. Introduction

Web-scale image search engines mostly use keywords as queries and rely on surrounding text to search images. It is well known that they suffer from the ambiguity of query keywords. For example, using "apple" as query, the retrieved images belong to different categories, such as "red apple", "apple logo", and "apple laptop". Image re-ranking [4] is an effective way to improve the search results. Its diagram is shown in Figure 1. Given a query keyword input by a user, according to a stored word-image index file, a pool of images relevant to the query keyword are retrieved by the search engine. By asking a 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 visual features of images are pre-computed offline and stored by the search engine. The main online computational cost of image re-ranking is on comparing visual features. In order to achieve high efficiency, the visual feature vectors need to be short and their matching needs to be fast.

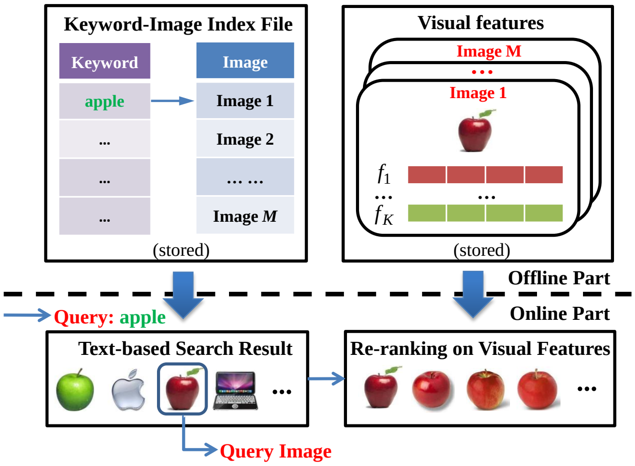


Figure 1. The conventional image re-ranking framework.

Another major challenge is that the similarities of low-level visual features may not well correlate with images' high-level semantic meanings which interpret users' search intention. To narrow down this semantic gap, for offline image recognition and retrieval, there have been a number of studies to map visual features to a set of predefined concepts or attributes as semantic signature [10, 7, 6]. However, these approaches are only applicable to closed image sets of relatively small sizes. They are not suitable for online web-based image re-ranking. According to our empirical study, images retrieved by 120 query keywords alone include more than 1500 concepts. Therefore, it is difficult and inefficient to design a huge concept dictionary to characterize highly diverse web images.

1.1. Our Approach

In this paper, a novel framework is proposed for web image re-ranking. Instead of constructing a universal concept dictionary, it learns different visual semantic spaces for different query keywords individually and automatically. We believe that 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 semantic concepts of "mountains" and "Paris" are unlikely to be relevant and can be ignored. Instead, the semantic concepts of "computers" and "fruit" will be used to learn the visual semantic space related to "apple". The query-specific visual semantic spaces can more accurately model the images to be re-ranked, since they have removed other potentially unlimited number of non-relevant concepts, which serve as noise and deteriorate the performance of re-ranking in terms of both accuracy and computational cost. The visual features of images are then projected into their related visual semantic spaces to get semantic signatures. At the online stage, images are re-ranked by comparing their semantic signatures obtained from the visual semantic space of the query keyword.

Our experiments show that the semantic space of a query keyword can be described by just $20 - 30$ concepts (also referred as "reference classes" in our paper). 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 visual semantic spaces of query keywords need to be automatically learned. Instead of manually defined, under our framework this is done through keyword expansions.

Another contribution of the paper is to introduce a large scale benchmark database[1] with manually labeled ground truth for the performance evaluation of image re-ranking. It includes $120,000$ labeled images of around 1500 categories (which are defined by semantic concepts) retrieved by the Bing Image Search using 120 query keywords. Experiments on this benchmark database show that $20\% - 35\%$ relative improvement has been achieved on re-ranking precisions with much faster speed by our approach, compared with the state-of-the-art methods.

1.2. Related Work

Content-based image retrieval uses visual features to calculate image similarity. Relevance feedback [12, 13] was widely used to learn visual similarity metrics to capture users' search intention. However, it required more users' effort to select multiple relevant and irrelevant image examples and often online training. For a web-scale commercial system, users' feedback has to be limited to the minimum with no online training. Cui et al. [4] proposed an image re-ranking approach which limited users' effort to just one-click feedback. Such simple image re-ranking approach has been adopted by popular web-scale image search engines such as Bing and Google recently.

The key component of image re-ranking is to compute the visual similarities between images. Many image features [8, 5, 2, 9] have been developed in recent years. However, for different query images, low-level visual features that are effective for one image category may not work well for another. To address this, Cui et al. [4] classified the query images into eight predefined intention categories and gave different feature weighting schemes to different types of query images. However, it was difficult for only 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.

Recently, for general image recognition and matching, there have been a number of works on using predefined concepts or attributes as image signature. Rasiwasia et al. [10] mapped visual features to a universal concept dictionary. Lampert et al. [7] used predefined attributes with semantic meanings to detect novel object classes. Some approaches [1, 6, 11] transferred knowledge between object classes by measuring the similarities between novel object classes and known object classes (called reference classes). All these concepts/attributes/reference-classes were universally applied to all the images and their training data was manually selected. They are more suitable for offline databases with lower diversity (such as animal databases [7, 11] and face databases [6]) such that object classes 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.

2. Approach Overview

The diagram of our approach is shown in Figure 2. At the offline stage, the reference classes (which represent different semantic concepts) of query keywords are automatically discovered. For a query keyword (e.g. "apple"), a set of most relevant keyword expansions (such as "red apple", "apple macbook" and "apple iphone") are automatically selected considering 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. 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

[1] http://137.189.35.203/WebUI/CVPR2011/DatasetDescription.htm

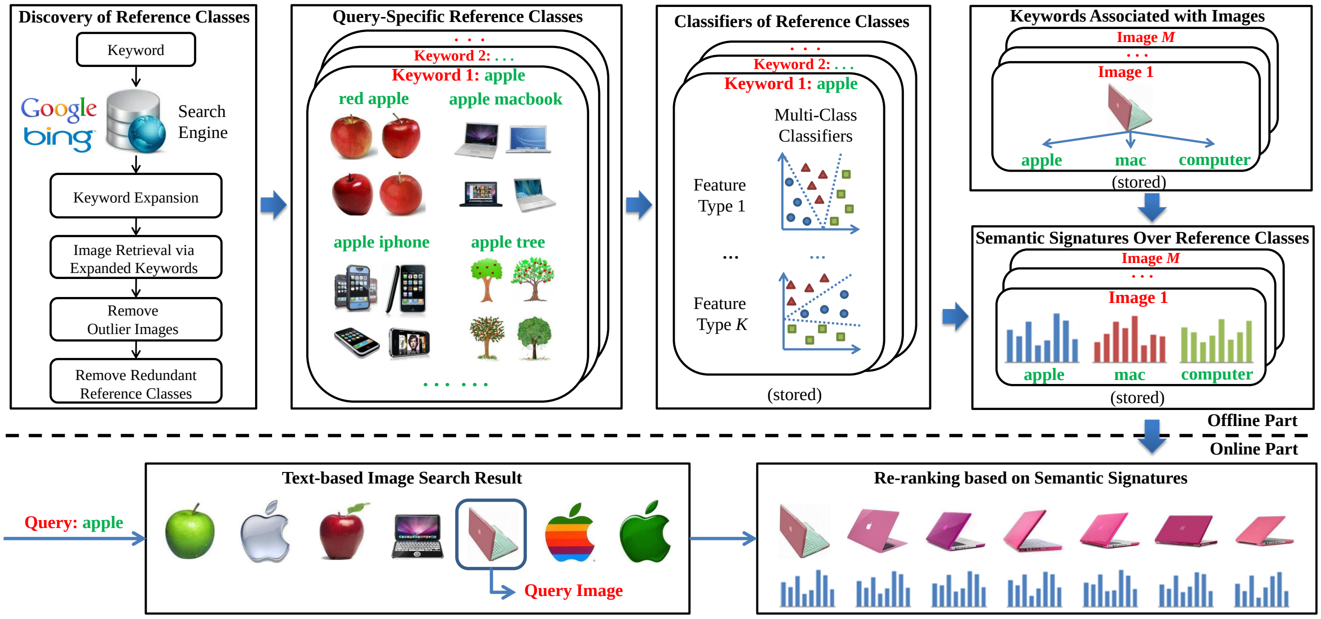


Figure 2. Diagram of our new image re-ranking framework.

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.

For each query keyword, a multi-class classifier on low-level visual features is trained from the training sets of its reference classes and stored offline. If there are K types of visual features, one could combine them to train a single classifier. It is also possible to train a separate classifier for each type of features. Our experiments show that the latter choice can increase the re-ranking accuracy but will also increase storage and reduce the online matching efficiency because of the increased size of semantic signatures.

An image may be relevant to multiple query keywords. Therefore it could have several semantic signatures obtained in different semantic spaces. According to the word-image index file, each image in the database is associated with a few relevant keywords. For each relevant keyword, a semantic signature of the image is extracted by computing the visual similarities between the image and the reference classes of the keyword using the classifiers trained in the previous step. The reference classes form the basis of the semantic space of the keyword. If an image has N relevant keywords, then it has N semantic signatures 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 input by a user. Since all the images in the pool are relevant to the query keyword, they all have pre-computed semantic signatures in the semantic space of the query keyword. Once the user chooses a query image, all the images are re-ranked by comparing similarities of the semantic signatures.

2.1. Discussion on Computational Cost and Storage

Compared with the conventional image re-ranking diagram shown in Figure 1, our approach is much more efficient at the online stage, because the main computational cost of online image re-ranking is on comparing visual features or semantic signatures and the lengths of semantic signatures are much short than those of low-level visual features. For example, the visual features used in [4] are of more than $1,700$ dimensions. Based on our experimental results, 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[2]. However, our approach needs extra offline computation and storage. 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, we will be able to process 100,000 query keywords in one day. With the fast growth of GPUs, which achieve hundreds of speedup than CPU, it is feasible

[2] In our experiments, 120 query keywords are considered. However, the keyword expansions, which define the reference classes, are from a very large dictionary used by the web search engine. They could be any words and are not limited to the 120 ones. Different query keywords are processed independently. Therefore, even if more query keywords are considered, the averaged dimensions of semantic signatures of each query keyword will not increase.

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 [3]. This will be left to the future work.

3. Discovery of Reference Classes

3.1. Keyword Expansion

For a keyword q, we automatically define its reference classes through 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 the words extracted from the images in $S(q)$[3]. A keyword expansion $e \in E_q$ is expected to frequently appear in $S(q)$. In order for reference classes to well capture the visual content of images, we require that there is a subset of images which all contain e and have similar visual content. Based on these considerations, keyword expansions are found in a search-and-rank way as follows.

For each image $I \in S(q)$, all the images in $S(q)$ are reranked according to their visual similarities (defined in [4]) to I. The T most frequent words $W_I = \{w_I^1, w_I^2, \cdots, w_I^T\}$ among top D re-ranked images are found. If a word w is among the top ranked image, it has a ranking score $r_I(w)$ according to its ranking order; otherwise $r_I(w) = 0$,

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

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

$$r(w) = \sum_{I \in S} r_I(w) \tag{1}$$

The P words with highest scores are selected and combined with the original keyword q to form keyword expansions, which define the reference classes. In our experiment, $T = 3$, $D = 16$ and $P = 30$.

3.2. Training Images of Reference Classes

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

3.3. Redundant Reference Classes

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

Suppose n reference classes are obtained from the previous steps. The training images of reference class i are split into two sets, A_i^1 and A_i^2. In order to measure the distinctness $D(i, j)$ between two reference classes i and j, a two-class SVM is trained from A_i^1 and A_j^1. For each image in A_i^2, the SVM classifier output a score indicating its probability of belonging to class i. Assume the averaging score over A_i^2 is $\bar{p}_i$. Similarly, the averaging score $\bar{p}_j$ over A_j^2 is also computed. Then $D(i, j) = h((\bar{p}_i + \bar{p}_j)/2)$, where h is a monotonically increasing function. In our approach, it is defined as

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

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 pair-wisely similar reference classes. We empirically choose $\alpha = 0.6$ and $\beta = 30$.

3.4. Reference Class Selection

We finally select a set of reference classes from the n candidates. The keyword expansions of the selected reference classes are most revelant to the query keyword q. The relevance is defined by Eq (1) in Section 3.1. In the 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 $n \times n$ matrix D defined in Section 3.3. The two criterions are simultaneously satisfied by solving the following optimization problem.

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

$$\arg \max_{y \in \{0,1\}^n} \left\{ \lambda R y + y^T D y \right\} \tag{2}$$

Let e_i be the keyword expansion of reference class i. $R = (r(e_1), \ldots, r(e_n))$, where $r(e_i)$ is defined in Eq (1). λ is the scaling factor used to modulate the two criterions. Since

[3] Words are extracted from filenames, ALT tags and surrounding text of images. They are stemmed and stop words are removed

integer quadratic programming is NP hard, we relax y to be in $\mathbb{R}^n$ and select reference classes i whose $y_i \geq 0.5$.

4. Semantic Signatures

Given M reference classes for keyword q and their training images automatically retrieved, 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. Then p is used as 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,

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

4.1. Combined Features vs Separate Features

In order to train the SVM classifier, we adopt six types of visual features used in [4]: attention guided color signature, color spatialet, wavelet, multi-layer rotation invariant edge orientation histogram, histogram of gradients and GIST. They characterize images from different perspectives of color, shape, and texture. The combined features have around $1,700$ dimensions in total.

A natural idea is to combine all types of visual features to train a single powerful SVM classifier which better distinguish different 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 N types of independent visual features, it is actually more effective to train separate SVM classifiers on different types of features and to combine the N semantic signatures $\{p^n\}_{n=1}^N$ from the outputs of N classifiers. The N semantic signatures describe the visual content of an image from different aspects (e.g. color, texture and shape) and can better characterize images outside the reference classes. For example, in Figure 3, "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.

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

$$d(I^a, I^b) = \sum_{n=1}^{N} w_n \left\|p^{a,n} - p^{b,n}\right\|_1 ,$$

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

$$w_n = \frac{1}{1 + e^{H(p^{a,n})}}$$

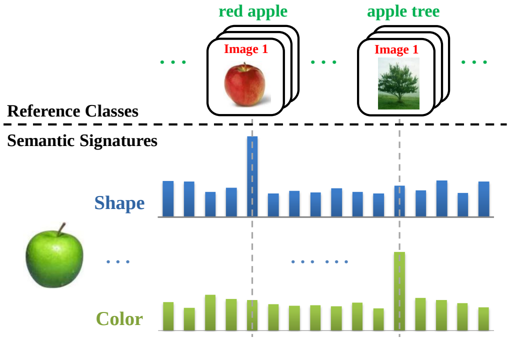


Figure 3. Describe "green apple" using reference classes. Its shape is captured by shape classifier of "red apple" and its color is captured by color classifier of "apple tree".

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

If $p^{a,n}$ uniformly distributes over reference classes, the nth 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. Experimental Results

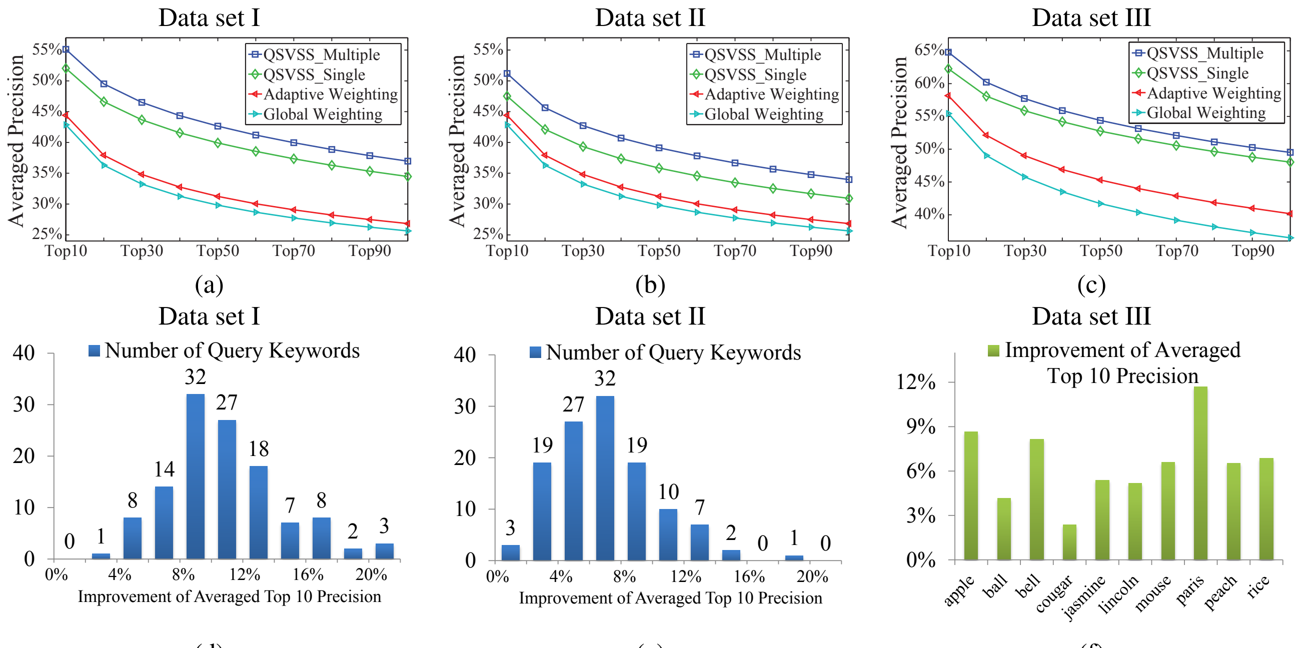
The images for testing the performance of re-ranking and the images of reference classes can be collected at different time[4] and from different search engines. Given a query keyword, 1000 images are retrieved from the whole web using certain search engine. As summarized in Table 1, we create three 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 using 120 query keywords in July 2010. These query keywords cover diverse topics including animal, plant, food, place, people, event, object, scene, etc. The images of reference classes were also collected from the Bing Image Search around the same time. Data set II use the same testing images for re-ranking as in data set I. However, its images of reference classes were collected from the Google Image Search also in July 2010. In data set III, both testing images and images of reference classes were collected from the Bing Image Search but at different time (eleven months apart)[5]. All testing images for re-ranking are manually labeled, while images of reference classes, whose number is much larger, are not labeled.

[4] The update of reference classes may be delayed.

[5] It would be closer to the scenario of real applications if the testing 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 [4].

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

Table 1. Descriptions of data sets



(a) (b) (c)

(d) (e) (f)

Figure 4. (a)-(c): comparisons of 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 of the 10 query keywords on data set III by comparing QSVSS_Multiple with Adaptive Weighting.

5.1. Re-ranking precisions

We invited five labelers to manually label testing images under each query keywords into different categories according to their semantic meanings. Image categories were carefully defined by the five labelers through inspecting all the testing images under a query keyword. Each image was labeled by at least three labelers and its label was decided by voting. A small portion of the images are labeled as outliers and not assigned to any category (e.g., some images are irrelevant to the query keywords).

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 top m precision for every query image (excluding outliers). We adopt this criterion instead of the precision-recall curve since in image re-ranking, the users are more concerned about the qualities of top retrieved images instead of number of relevant images returned in the whole result set.

We compare with two benchmark image re-ranking approaches used in [4]. They directly compare visual features. (1) **Global Weighting**. Predefined fixed weights are adopted to fuse the distances of different low-level visual features. (2) **Adaptive Weighting**. [4] proposed adaptive weights for query images to fuse the distances of different low-level visual features. It is adopted by Bing Image Search.

For our new approaches, two different ways of computing semantic signatures as discussed in Section 4.1 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.

Some parameters used in our approach as mentioned in Sections 3 and 4 are tuned in a small separate data set and they are fixed in all the experiments.

The averaged top m precisions on data sets I-III are shown in Figure 4 (a)-(c). Our approach significantly outperforms Global Weighting and Adaptive Weighting, which directly compare visual features. On data set I, our approach enhances the averaged top 10 precision from 44.41% (Adaptive Weighting) to 55.12% (QSVSS_Multiple). 24.1% relative improvement has been achieved. Figure 4 (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 4 (f) shows the improvements of averaged top 10 precision of the 10 query keywords on data set III.

In our approach, computing multiple semantic signatures from separate visual features has higher precisions than computing a single semantic signature from combined features. However, it costs more online computation since the dimensionality of multiple semantic signatures is higher. Comparing Figure 4 (a) and Figure 4 (b), if the testing images for re-ranking and 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. However, it is still much higher than directly comparing visual features. This indicates that we can utilize images from various sources to learn query-specific semantic spaces. As shown in Figure 4 (c), even if the testing images and images of reference classes are collected at different times (eleven months apart), query specific semantic spaces still can effectively improve re-ranking. 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 and do not have to be updated very frequently.

5.2. Online efficiency

The online computational cost of image re-ranking depends on the length of visual feature (if directly comparing 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. Therefore the length of the single semantic signature (QSVSS_Single) is 25 on average. Since six types of visual features are used, the length of the multiple semantic signatures (QSVSS_Multiple) is 150. It takes 12ms to re-rank 1000 images matching the visual features, while QSVSS_Multiple and QSVSS_Single only need 1.14ms and 0.2ms respectively. Given the large improvement of precisions our approach has achieved, it also improves the efficiency by around 10 to 60 times compared with matching visual features.

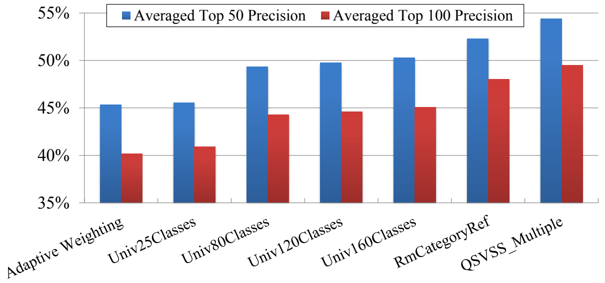


Figure 5. Comparisons of averaged top m precisions of re-ranking images outside reference classes and using universal semantic space on data set III.

5.3. Re-ranking images outside the reference classes

It is interesting to know whether the learned query-specific semantic spaces are effective for query images which are outside the reference classes. 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[6]. Multiple semantic signatures (QSVSS_Multiple) are used. The results are shown in Figure 5. It still greatly outperforms the approaches of directly comparing visual features. This result can be explained from two aspects. (1) As discussed in Section 4.1, the multiple semantic signatures obtained from different types of visual features separately have the capability to characterize the visual content of images outside the reference classes. (2) Many negative examples (images belonging to different categories than the query image) are well modeled by the reference classes and are therefore pushed backward on the ranking list.

5.4. Query-specific semantic space vs. universal semantic space

In previous works [10, 7, 1, 6, 11], 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. In this experiment, we evaluate whether this approach is applicable to web-based image re-ranking and compare it with our approach. 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

[6] We did not do this evaluation on the larger data set I or II because it is very time consuming. For every query image, the SVM classifiers have to be re-trained.

selected reference classes are used to train a universal semantic space in a way similar to Section 4.1. Multiple semantic signatures are obtained from different types of features separately. This universal semantic space is applied to data set III for image re-ranking. The averaged top m precisions are shown in Figure 5. M is chosen as 25, 80, 120 and 160[7]. 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 even lower than RmCategoryRef, given that the length of its semantic signatures is five times larger than ours.

5.5. User study

User experience is critical for web-based image search. In order to fully reflect the extent of users' satisfaction, user study is conducted to compare the results of our approach (QSVSS_Multiple) compared with Adaptive Weighting on data set I. Twenty users are invited. Eight of them are familiar with image search and the other twelve are not. To avoid bias on the evaluation, we ensure that all the participants do not have any knowledge about the 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. $12,000$ user comparison results are collected. The comparison results are shown in Figure 6. In over 55% cases our approach delivers better results than Adaptive Weighting and only in less than 18% cases ours is worse, which are often the noisy cases with few images relevant to the query image exists.

Please find examples of search results of different re-ranking methods from the project web page [8].

6. Conclusion

We propose a novel image re-ranking 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 visual semantic spaces automatically learned through keyword expansions at the offline stage. The extracted semantic signatures can be 70 times shorter than the original visual feature on average, while achieve $20\% - 35\%$ relative improvement on re-ranking precisions over state-of-the-art methods.

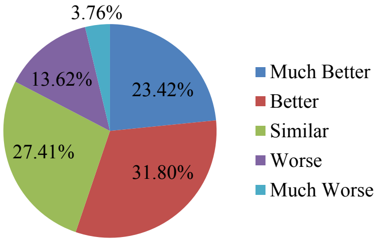


Figure 6. Comparison results of user study on data set I.

References

[1] E. Bart and S. Ullman. Single-example learning of novel classes using representation by similarity. In *Proc. BMVC*, 2005. 2, 7

[2] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-bag-of-features. In *Proc. CVPR*, pages 3352–3359, 2010. 2

[3] G. Cauwenberghs and T. Poggio. Incremental and decremental support vector machine learning. 2001. 4

[4] J. Cui, F. Wen, and X. Tang. Real time google and live image search re-ranking. In *Proc. ACM Multimedia*, 2008. 1, 2, 3, 4, 5, 6

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005. 2

[6] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *Proc. ICCV*, 2009. 1, 2, 7

[7] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. CVPR*, 2005. 1, 2, 7

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l Journal of Computer Vision*, 60(2):91–110, 2004. 2

[9] J. Philbin, M. Isard, J. Sivic, and A. Zisserman. Descriptor Learning for Efficient Retrieval. In *Proc. ECCV*, pages 677–691, 2010. 2

[10] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Trans. on Multimedia*, 9:923–938, 2007. 1, 2, 7

[11] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps wherevand why? semantic relatedness for knowledge transfer. In *Proc. CVPR*, 2010. 2, 7

[12] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 8:644–655, 1998. 2

[13] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8:536–544, 2003. 2

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

[8] http://137.189.35.203/WebUI/CVPR2011/ProjectWebPage.htm