Efficient Re-ranking in Vocabulary Tree based Image Retrieval

Xiaoyu Wang[2], Ming Yang[1], Kai Yu[1]
[1]NEC Laboratories America, Inc. [2]Dept. of ECE, Univ. of Missouri
Cupertino, CA 95014 Columbia, MO 65211
{myang,kyu}@sv.nec-labs.com xw9x9@missouri.edu

***Abstract*—Image retrieval using a large vocabulary tree of local invariant features can efficiently handle databases with millions of images. However, a costly re-ranking step is generally required to re-order the top candidate images to enforce spatial consistency among local features. In this paper, we propose an efficient re-ranking approach which takes advantage of the vocabulary tree quantization to conduct fast feature matching. The proposed re-ranking algorithm involves no operations in the high-dimensional feature space and does not assume a global transform between a pair of images, thus, it not only dramatically reduces the computational complexity but also improves the retrieval precision, which is validated using 1.26 million images in the public *ImageNet* dataset and the *San Francisco Landmark* dataset including 1.7 million images.**

I. Introduction

Large-scale image retrieval from millions of images has been an active research topic for decades [3], [8], which attracts increasing interests due to its applications in web and mobile image search [6], [7], [11], [16], [12], [1], [14]. In terms of the methodologies and features, recent large-scale image retrieval algorithms may be categorized into two lines: 1) compact hashing of global features; and 2) efficient indexing of local features by a vocabulary tree. Global features such as GIST features [9] or color histograms delineate the holistic contents of images, which can be compactly indexed by binary codes [11] or hashing functions [15]. Thus, the retrieval is very efficient on both computation and memory usage though it is unable to attend to the details of images. In the other line of research, images are represented by a bag of local invariant features [5] which are quantized into visual words by a huge vocabulary tree [6]. This vocabulary tree based retrieval is very capable of finding near-duplicate images, *i.e.*, images of the same objects or scenes undergoing different capturing conditions, at the cost of memory usage for the inverted indexes of a large number of visual words.

In the vocabulary tree based image retrieval, since images are essentially represented by a bag of orderless visual words, the geometric relations of the local features or their spatial layout are largely ignored. Therefore, a post re-ranking procedure [7], [4] is often employed to re-order the retrieved candidate images by verifying the geometrical consistency against the query image in order to further improve the retrieval precision. Usually, in the geometrical re-ranking, the local feature descriptors of two images are first matched reliably using the method in [5], then a RANSAC procedure [7] is employed to fit a global affine transform. The candidate images are re-ranked according to the number of inliers in the RANSAC or fitting errors. This conventional re-ranking approach confronts by two issues. First, this procedure is generally computational intensive because it operates on the high dimensional descriptors. The running time could be even longer than the retrieval. Second, the assumption of a global affine transform between two image may not hold, *e.g.*, for images of a 3D object from different view angles.

In viewing of these issues, we develop an efficient re-ranking method for vocabulary tree based retrieval that takes advantage of the tree quantization to select a small set of matched local features and verify the consistency of their individual spatial local neighborhoods. The matched local features with a more consistent neighborhood shall contribute more to the matching score to re-rank the candidate images. In this procedure, we do not resort to the high dimensional descriptors, thus it is very efficient. In addition, we do not assume a global transform between a candidate image to the query, so it is potentially more general than the RANSAC based method. The proposed re-ranking method is particularly beneficial to a recent large-scale image retrieval algorithm [14] where the spatial neighbors of the local features in the query has been pre-calculated in spatial contextual weighting. We validate the effectiveness of this re-ranking on the *UKbench* dataset [6] using 1.26 million images in the *ImageNet* Challenge [2] as the distractors and the landmark retrieval problem on the *San Francisco Landmark* dataset [1] including 1.7 million images. All the images in the experiments are publicly available.

After briefly reviewing some related work in Sec. II, we introduce the vocabulary tree based image retrieval [6] and the variant using spatial contextual weighting [14] in Sec. III. Following the same notations, we present the proposed efficient re-ranking method in Sec. IV, with experiments on two large datasets in Sec. V and the concluding remarks in Sec. VI.

II. Related Work

In recent years, large-scale near-duplicate image retrieval based on local invariant features [5] has been significantly improved by large hierarchical vocabulary trees [6], where local features are encoded into a bag-of-words (BoW) histogram [10] with millions of visual words. Therefore, this histogram is so sparse that inverted indexes are well suited to implement the indexing and searching efficiently. Moreover, the tree structure substantially reduces the computation to quantize a feature descriptor into one of millions of words, *e.g.*, for a tree with 7 layers and branch factor 10, only 7×10 inner products of 128-D SIFT descriptors are needed, instead of 10^7 for a flat structure. Visual words are weighted by the TF-IDF (term frequency-inverse document frequency) [10], [6], where the IDF reflects their discriminative abilities in database images and the TF indicates their importance in a query image.

To take into account the spatial layout of local features, [7] proposed a post spatial verification stage to re-rank the top-1000 candidates by a variant of RANSAC algorithm assuming a global transform with 3-5 degree of freedom between the query and candidate images. The spatial verification stage substantially improves the mean average precision (mAP) of the retrieval, however, the induced computational costs could be far more expensive than the the retrieval procedure itself. This geometric re-ranking procedure has been widely adopted in many following up work.

The most related work to ours is [12] where the matching feature pairs between a query and candidate image are identified using the descriptor quantization path in the vocabulary tree. Then the geometric similarity scores of those matching feature pairs, *i.e.*, the maximal number of features with similar location, orientation and scale ratio differences, are used to re-rank the candidates. This method essentially combines 3 one-order Hough transforms, which also assumes a single global transform between the query and candidate. We employ a similar way to identify the subset of matched local features. The major differences are that we do not rely on the global transform assumption and we augment the matching scores of those features with consistent local spatial neighbors.

III. Vocabulary tree based image retrieval

In the section, we present the framework of the vocabulary tree based image retrieval [6] shown in Fig. 1 and introduce a recent work [14] which proposed to employ two types of image-dependent contextual information in weighting the local features besides the TF-IDF weighting scheme.

SIFT features $\mathbf{f} = \{\mathbf{x}, \mathbf{u}, s, \theta\}$ are first extracted from both query and database images, which include the descriptor $\mathbf{x} \in \mathbb{R}^D$, location $\mathbf{u}$, characteristic scale s in log domain, and orientation θ. Then, a vocabulary tree T is offline trained by hierarchical K-means clustering of local descriptors with a branch factor K and depth L. Each node v^{ℓ,h_ℓ} (v for short) in the tree represents a visual word, where ℓ indicates its layer and h_ℓ is the node index at that layer.

A query image q is represented by a bag I_q of local descriptors $\{\mathbf{x}_i\}_{i \in I_q}$, where $\mathbf{x}_i \in \mathbb{R}^D$ and $D = 128$ for SIFT features. Each $\mathbf{x}_i$ is mapped to a path of visual words from the root to a leaf of T, resulting in the quantization $T(\mathbf{x}_i) = \{v_i^{\ell,h_\ell}\}_{\ell=1}^{L_i}$. Thus, a query image is eventually represented by the set of node paths obtained from its descriptors, *i.e.*, $\{T(\mathbf{x}_i)\}_{i \in I_q}$. The database images are denoted by $\{d^m\}_{m=1}^M$. Using the same hierarchical quantization procedure, the local descriptors $\mathbf{y}_j^m$ in d^m are mapped to the collection of node paths $\{T(\mathbf{y}_j^m)\}_{j \in I_{d^m}}$. The similarity score $sim(q, d^m)$ between the query q and a database image d^m is given by the average matching score among all pairs of descriptors passing the same node paths.

$$sim(q, d^m) \doteq \frac{1}{|I_q||I_{d^m}|} \sum_{i \in I_q, j \in I_d} w(v_i) \mathbb{1}(v_i = v_j). \quad (1)$$

In particular, the matching score of two features quantized to one node v is specified as the IDF term of this node [6]:

$$w(v) = idf(v) = \log\left(\frac{M}{M_v}\right), \quad (2)$$

where M is the total number of database images and M_v is the number of images containing at least one descriptor that quantizes to the node v.

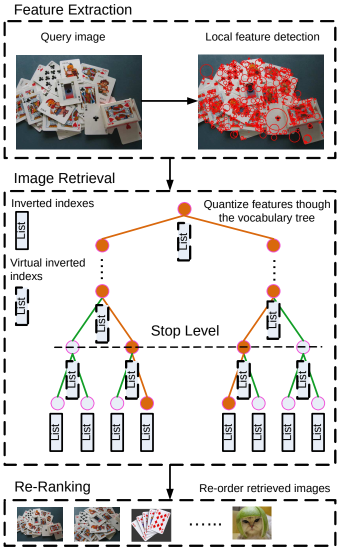

Fig. 1. The framework of vocabulary tree based image retrieval.

To enhance the discriminability of single local features, Wang *et. al.* [14] proposed a *descriptor contextual weighting* (DCW) and a *spatial contextual weighting* (SCW) of local features to improve the IDF weighting scheme. The SCW calculates spatial contextual statistics for each local feature, including the density of its adjacent features and their scales and orientations, where the computation to determine the spatial neighborhoods is reusable in the proposed re-ranking method. Please refer to [14] for the details.

IV. Proposed efficient re-ranking

Given the top candidate images, we develop an efficient re-ranking method to re-order them according to the local geometrical consistency of $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_j^m\}$. First we obtain a small subset of "matched" local descriptors *w.r.t.* the vocabulary tree quantization, namely, we call the descriptors in two images as "matched" if they are quantized to a unique tree node at the deep levels of the tree. Afterwards, we construct two sub-graphs of these matched features based on their spatial neighborhood relations. The IDF weights of matched features in the intersection of two sub graphs are weighted by the ratio of common neighbor features and added to the overall image similarity score again to re-order the top candidates.

Specifically, for the query q and a candidate d^m, we calculate the intersection of all node paths $\{\mathbf{v}_i\}_{i \in I_q}$ and $\{\mathbf{v}_j^m\}_{j \in I_d^m}$, and only select those nodes (with depth $l > L - 3$) whose $n_I^q(v^{l,h_l}) = 1$ and $n_I^d(v^{l,h_l}) = 1$, denoted by $\{v^{l\prime,h\prime}\}$. Here n_I^q and n_I^d indicate the number of local features quantized to a particular tree node. As the vocabulary tree is quite deep, the subsets of descriptors $\{\mathbf{x}_i'\}$ and $\{\mathbf{y}_j'\}$ that correspond to $\{v^{l\prime,h\prime}\}$ are regarded roughly matched and only around $10\% - 20\%$ of the initial feature sets. Then we build two graphs from matched $\{\mathbf{x}_i'\}$ and $\{\mathbf{y}_j'\}$ where $\mathbf{x}_i'$ links to $\mathbf{x}_i''$ which is in its spatial neighborhood $C(\mathbf{x}_i')$. Here, let $C(\mathbf{f})$ denote the neighborhood of one feature given by the disc $(\mathbf{u}, R)$. Empirically we set the radius $R = 12 \times 2^s$ (maximum 150 pixels). Finally, we calculate the intersection of these 2 graphs and add the weighted $idf(v^{l\prime,h\prime})$ to the matching score $sim(q, d^m)$ to re-order the top returned images. The final similarity score of two images is defined as

$$\overline{sim}(q, d^m) \doteq sim(q, d^m) + \sum_{\{\mathbf{x}_i'\}} \alpha(\mathbf{x}_i') idf(v^{l\prime,h\prime}), \qquad (3)$$

where $\alpha(\mathbf{x}_i') = \frac{|\{\mathbf{x}_i'' | \mathbf{x}_i'' \in C(\mathbf{x}_i') \text{ and } \mathbf{y}_i'' \in C(\mathbf{y}_i')\}|}{|C(\mathbf{x}_i')|}$ and $\mathbf{x}_i''$ matches to $\mathbf{y}_i''$, the ratio of common neighbors of $\mathbf{x}_i'$ in the query and its matched feature $\mathbf{y}_i'$ in the database image.

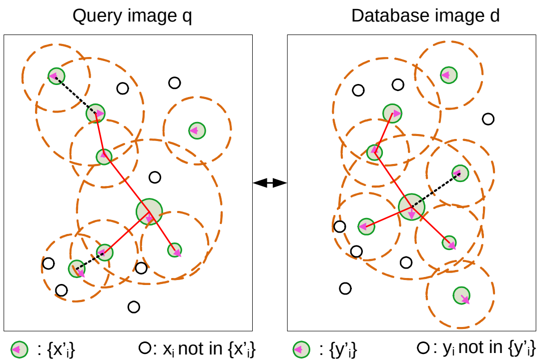

Fig. 2. Illustration of the proposed re-ranking process.

Figure 2 shows an illustrative example. The green circles indicate the matched SIFT features by the tree quantization and they have the one-on-one correspondences in q and d. According to their spatial neighborhoods drawn as orange dash circles, the features with larger common neighborhood contribute more to the final re-ranking similarity score. When the global transform assumption does not hold, the number of inliers in the RANSAC could be small and unable to reliably re-rank the candidates. In such cases, this re-ranking method allows local deformation and promotes those candidates with large consistent local image regions. Furthermore, the computation of the tree based quantization and finding out the spatial neighborhoods of local features are shared with the retrieval using spatial contextual weighting, thus, the induced computation overhead is quite limited.

V. Experiments

We re-rank the retrieval results of [14] to validate the proposed method using 3 public datasets, *Ukbench*, *ImageNet*, and *SFLandmark*. In all the experiments, we extract up to 2500 SIFT features for each image using the VLFeat library [13]. The system was implemented in C++ and all timings of the average retrieval time t_r, including the re-ranking but not the feature extraction, are based on a *single* core of an 8-core Xeon 2.44GHz blade server with 32G RAM. Next we introduce each dataset and the evaluation criteria, then present the large-scale retrieval performance.

A. The datasets

We conduct two large-scale image retrieval experiments where one experiment focuses on finding the same object in the *UKbench* dataset with 1.26M images in the *ImageNet* as the distractors, and the other searches for the same landmark or building in the *San Francisco Landmark* dataset including 1.7M images. To our best knowledge, these are among the largest public datasets available for image retrieval.

UKbench was collected by [6], and includes 2550 different objects or scenes. Each one has 4 images taken from different viewpoints. All the 10200 images are both indexed as database images and used as queries. The retrieval performance is measured by $4\times$ recall at the top-4 candidate images, referred as N-S score [6] (maximum is 4), and the mean average precision (mAP).

ImageNet is an image database crawled from internet according to the WordNet hierarchy [2]. The *ImageNet* Large Scale Visual Recognition Challenge 2010 (ILSVRC2010) clearly specifies 1.26M images of 1000 categories as the training set (denoted by *ImageNet-T*) and 50K images as the validation set (denoted by *ImageNet-V*).

SFLandmark (the San Francisco Landmark dataset [1]) contains 150K panoramic images collected in San Francisco which were converted to 1.06 million perspective central images (PCIs) and 638 thousand perspective frontal images (PFIs). The query images are 803 pictures captured by mobile phones. All the images are associated with one or several building IDs and true or approximate GPS locations. We evaluate the recall rate of the top-1 candidate image in terms of building IDs, denoted by τ_1.

B. Large-scale retrieval performance

In all the experiments, we employ a vocabulary tree with depth $L = 7$ and branch factor $K = 10$. For the experiments on the *UKbench* dataset, the tree is trained using 8 million randomly selected SIFT features from the *ImageNet-V* dataset which has no overlap with *UKbench* and *ImageNet-T*. Following the practice in [1], the vocabulary trees for the *SFLandmark* dataset are trained for PCIs and PFIs separately.

We compare the retrieval performance among the *baseline* method in [6], the contextual weighting (*CW*) algorithm [14], our re-ranking method over the CW algorithm (denoted by *CW+TreeRank*), and the conventional re-ranking method using the RANSAC [7] (denoted by *CW+RANSAC*) in which the

TABLE I
IMAGE RETRIEVAL PERFORMANCE ON *UKbench*.

	UKbench		*UKbench+ImageNet-T*		
Method	N-S	mAP(%)	N-S	mAP (%)	t_r (ms)
Baseline [6]	3.38	87.76	2.89	75.17	311
CW [14]	3.56	91.70	3.30	85.17	676
CW+TreeRank	3.59	92.06	3.39	87.98	708
CW+RANSAC	3.61	92.27	3.36	87.53	2728

number of inliers in fitting a 6D Affine model is used to re-rank the candidates.

For the *UKbench*, besides the common experiment setting, *i.e.*, using the 10200 images as both query and database images, we also index the images in both *UKbench* and *ImageNet-T* into the database to perform the large-scale experiment. We re-rank the top-10 candidates given by the *CW* algorithm in both *CW+TreeRank* and *CW+RANSAC*. The retrieval performance is shown in Tab. I, from which we observe that the proposed re-ranking method improves the N-S score and mAP considerably over the *CW* algorithm with much less computational overhead than the RANSAC method. Interestingly, in the large-scale case, it even outperforms the RANSAC method, probably because the images of some 3D objects viewed from different angles cannot be modeled by Affine transforms. Note, we choose to re-rank only the top 10 candidates, while more candidates [4] may further boost the performance.

For the *SFlandmark*, the goal is to retrieve and recognize the buildings in the photos captured by mobile phones. We follow the same pre-processing procedure in [1] to perform the histogram equalization to the PFIs and PCIs before extracting SIFT features and re-rank the top 50 candidates. In [1], the upright SIFT features, *i.e.*, ignoring the characteristic orientation, are observed to yield better performance than the original SIFT since the buildings are mostly upright in the query and database images. We choose to use the original SIFT since the query images captured by mobile phones could be in any possible orientations. The retrieval results are shown in Tab. II and Fig. 3. The recall of the RANSAC method is higher than the proposed re-ranking, perhaps because the panoramic pictures of buildings were captured at a distance so that the Affine transform assumption is approximately satisfied. Our performance compares favorably against that of re-ranking with RANSAC using the "oriented" SIFT in [1] whose recall rates at the top-1 candidate are about 50% and 48% for PCIs and PFIs respectively. Note [1] employs a hard threshold on the number of inliers to guarantee the precision at the top-1 candidate with GPS is 95%, which may reduce the recall rates. Following [1], in Tab. II we also list the retrieval performance leveraging the GPS information to filter out locations larger than 300 meters away from the query.

Some examples of the retrieval results after re-ranking are shown in Fig. 4, Fig. 5 and Fig. 6 where the first column are the queries. Note on the *UKbench*, the top-1 candidate is always the query image itself.

TABLE II
IMAGE RETRIEVAL PERFORMANCE ON *SFLandmark*.

	PCIs		*PFIs*	
Method	τ_1 (%)	t_r (ms)	τ_1 (%)	t_r (ms)
Baseline [6]	46.45	302	36.24	201
CW [14]	59.40	645	45.58	467
CW+TreeRank	62.76	806	50.31	620
CW+RANSAC	69.36	12389	54.79	11913
CW+RANSAC+GPS	74.22	12389	62.39	11913

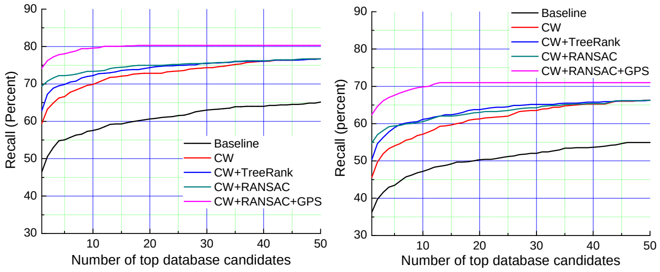

Fig. 3. The recall in terms of buildings versus number of top database candidates for the 803 queries on the 1.06M PCIs (on the left) and on the 638k PFIs (on the right).

VI. CONCLUSION

In this paper, we proposed an efficient re-ranking method for vocabulary tree based image retrieval. Utilizing the tree quantization to select a small subset of matched local features, we verify the consistency of their spatial neighborhoods and re-order the candidates by augmenting their matching similarity scores. The method improves the retrieval precision with limited computational overhead. The experiments on two large-scale datasets, retrieving the same objects or landmark buildings demonstrate promising performance.

ACKNOWLEDGEMENT

This work was done during the internship of the first author at NEC Laboratories America in Cupertino, CA.

REFERENCES

[1] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale lanrkmark indentification on mobile devices. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 21-23, 2011.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, FL, June 20-26, 2009.

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *Computer*, 28(9):22–32, Sept. 1995.

[4] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-feature for large scale image search. *Int'l Journal of Computer Vision*, 87(3):316–336, 2010.

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

[6] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, New York City, June 17-22, 2006.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, June 17-22, 2007.

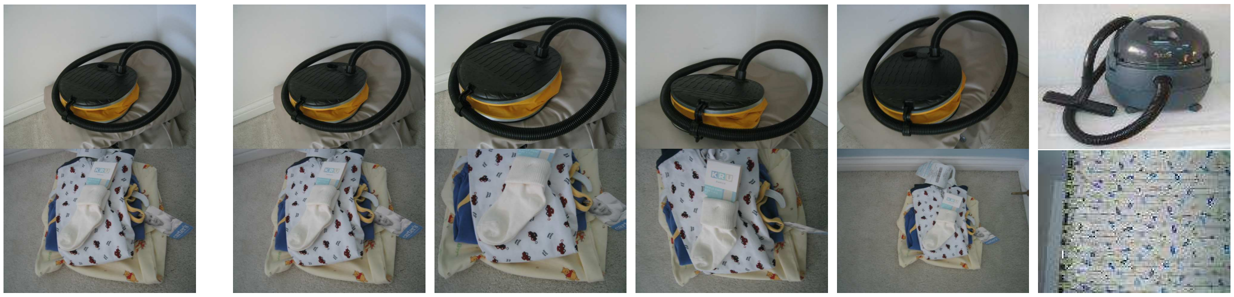

Fig. 4. Sample retrieval results in *UKbench* and *ImageNet-T*. The query image is shown in the first column followed by top-5 candidates after re-ranking.

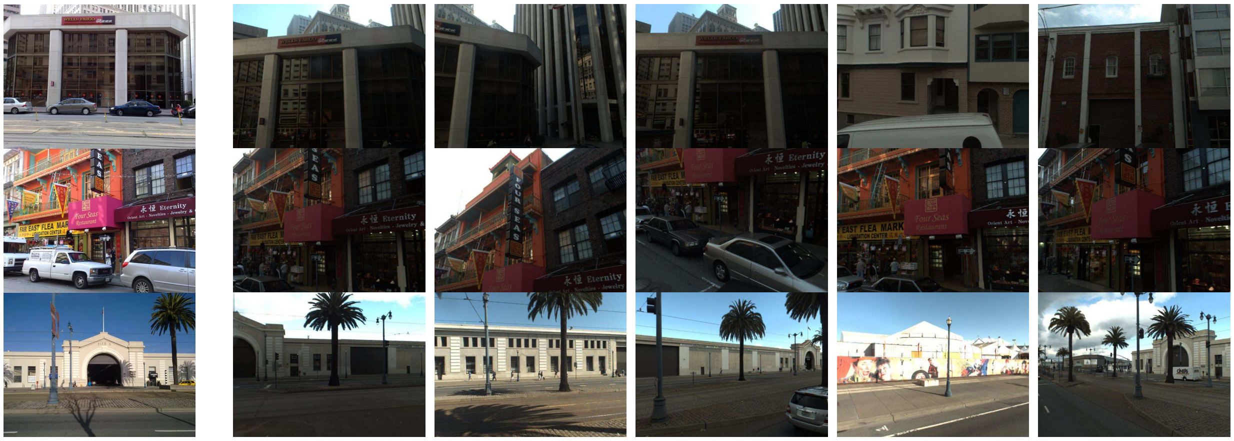

Fig. 5. Sample retrieval results on PCIs in *SFLandmark*. The query image is shown in the first column followed by top-5 candidates after re-ranking.

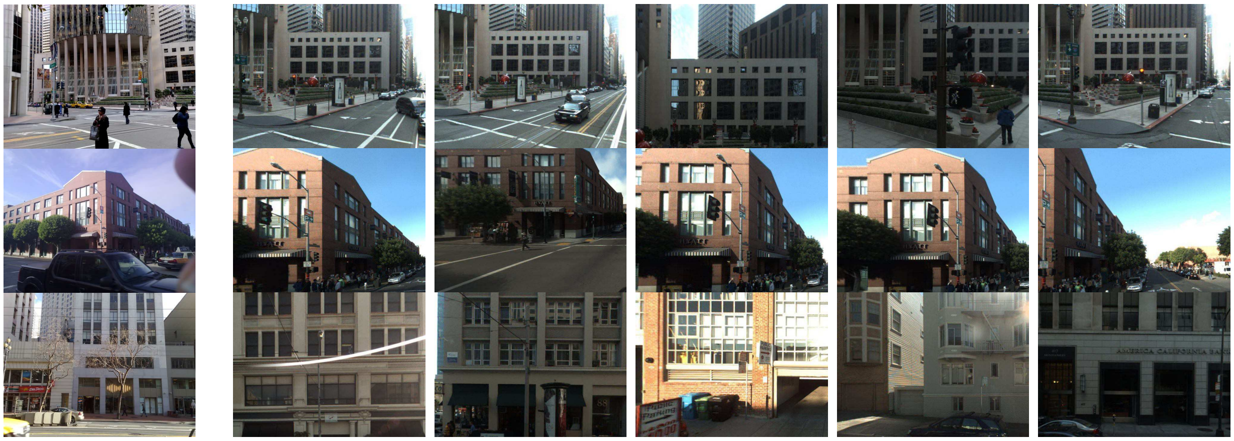

Fig. 6. Sample retrieval results on PFIs in *SFLandmark*. The query image is shown in the first column followed by top-5 candidates after re-ranking.

[8] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. Circuits Syst. Video Technol.*, 8(5):644–655, Sept. 1998.

[9] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans. Pattern Anal. Machine Intell.*, 29(2):300–312, Feb. 2007.

[10] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE Int'l Conf. on Computer Vision*, volume 2, Nice, France, Oct. 13-16, 2003.

[11] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, June 23-28, 2008.

[12] S. S. Tsai, D. Chen, C. Takacs, V. Chandrasekhar, Ramakrishna, Vedantham, R. Grzeszczuk, and B. Girod. Fast geometric re-ranking for image-based retrieval. In *IEEE Int'l Conf. on Image Processing*, Hong Kong, Sept. 26-29, 2010.

[13] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[14] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han. Contextual weighting for vocabulary tree based image retrieval. In *IEEE Int'l Conf. on Computer Vision*, Barcelona, Spain, Nov. 6-13, 2011.

[15] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, Dec. 9-11, 2008.

[16] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian. Spatial coding for large scale partial-duplicate web image search. In *ACM Multimedia*, Florence, Italy, Oct. 25-29, 2010.