Generating Descriptive Visual Words and Visual Phrases for Large-Scale Image Applications

Shiliang Zhang, Qi Tian, *Senior Member, IEEE*, Gang Hua, *Member, IEEE*, Qingming Huang, *Senior Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*—Bag-of-visual Words (BoWs) representation has been applied for various problems in the fields of multimedia and computer vision. The basic idea is to represent images as visual documents composed of repeatable and distinctive visual elements, which are comparable to the text words. Notwithstanding its great success and wide adoption, visual vocabulary created from single-image local descriptors is often shown to be not as effective as desired. In this paper, descriptive visual words (DVWs) and descriptive visual phrases (DVPs) are proposed as the visual correspondences to text words and phrases, where visual phrases refer to the frequently co-occurring visual word pairs. Since images are the carriers of visual objects and scenes, a descriptive visual element set can be composed by the visual words and their combinations which are effective in representing certain visual objects or scenes. Based on this idea, a general framework is proposed for generating DVWs and DVPs for image applications. In a large-scale image database containing 1506 object and scene categories, the visual words and visual word pairs descriptive to certain objects or scenes are identified and collected as the DVWs and DVPs. Experiments show that the DVWs and DVPs are informative and descriptive and, thus, are more comparable with the text words than the classic visual words. We apply the identified DVWs and DVPs in several applications including large-scale near-duplicated image retrieval, image search re-ranking, and object recognition. The combination of DVW and DVP performs better than the state of the art in large-scale near-duplicated image retrieval in terms of accuracy, efficiency and memory consumption. The proposed image search re-ranking algorithm: DWPRank outperforms the state-of-the-art algorithm by 12.4% in mean average precision and about 11 times faster in efficiency.**

***Index Terms*—Image retrieval, image search re-ranking, object recognition, visual phrase, visual word.**

Manuscript received April 07, 2010; revised September 14, 2010, January 27, 2011; accepted February 03, 2011. Date of publication March 17, 2011; date of current version August 19, 2011.This work was supported in part by Microsoft Research Asia (MSRA), the National Science Foundation under Grant IIS 1052581, a Google Faculty Research Award, an FXPAL Faculty Research Award, the National Natural Science Foundation of China under Grant 61025011 and Grant 60833006, the National Basic Research Program of China (973 Program) under Grant 2009CB320906, and the Beijing Natural Science Foundation under Grant 4092042. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Min Wu.

S. Zhang and W. Gao are with the Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: slzhang@jdl.ac.cn; wgao@jdl.ac.cn).

Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

G. Hua is with the IBM Research T. J. Watson Center, NY, 10532 USA (e-mail: ganghua@gmail.com).

Q. Huang is with the Graduate University, Chinese Academy of Sciences, Beijing 100080, China (e-mail: qmhuang@jdl.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2011.2128333

I. Introduction

BAG-OF-VISUAL words (BoWs) image representation has been utilized for many multimedia and vision problems, including video event detection [36], [40], [46], object recognition [16], [17], [24], [25], [27], [30], [35], image segmentation [37], and large-scale image retrieval [10]–[12], [23], [31], [39]. Representing an image as a visual document composed of repeatable and distinctive visual elements that are indexable is very desirable. With such a representation, many matured techniques in information retrieval can be leveraged for vision tasks, such as visual search or recognition. Recently, it has been demonstrated that BoWs image representation is one of the most promising approaches for retrieval tasks in large-scale image and video databases [10]–[12], [23], [31], [39].

However, experimental results of reported works show that the commonly generated visual words [10], [21], [31], [37], are still not as expressive as the text words. Traditionally, the classic visual vocabulary is created by clustering a large number of local feature descriptors. The exemplar descriptor of each cluster is called a visual word, which is then indexed by an integer. In previous works [17], [23], [24], [27], [36], [39], [41], [43], [44], various numbers of visual words are generated for different tasks.

There are two general observations: 1) using more visual words results in better performance [17], [23], [27] and 2) however, the performance will be saturated when the number of visual words reaches certain levels [17], [23], [27]. Intuitively, a larger number of visual words indicates more fine-grained partitioning of the descriptor space. Hence, the visual words become more discriminative in representing certain visual contents. The second observation is that increasing the number of visual words to certain levels finally saturates the performance of vision vocabulary. Intuitively, dividing the feature space in finer scales increases the quantization error in visual vocabulary. This means local features near in the feature space might be quantized into different visual words.

These observations strongly imply the limited descriptive ability of the classic visual word. A toy example illustrating this finding is presented in Fig. 1. In the figure, SIFT descriptors are extracted on interest points detected by Difference of Gaussian (DoG) [20]. The three images are then represented as BoWs with a visual vocabulary containing 32 357 visual words, by replacing their SIFT descriptors with the indexes of the closest visual words. In the figure, two interest points are connected with a red line (online version) if they share the same visual word. As we can clearly observe, although the visual appearances of the plane and cat are very different, there are still many matched visual words between them. It can be inferred

Fig. 1. Matched visual words between the same and different objects.

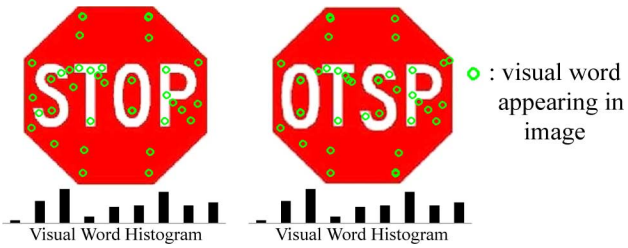


Fig. 2. Two images show different semantics. However, they contain the identical visual word histogram. Obviously, traditional BoW representation loses the spatial context in images.

that the visual word is noisy and indiscriminative, resulting in its ineffectiveness in measuring the similarity between the two images.

There are two problems in the classic visual words, which may be the main causes for their limited descriptive power.

1) Single visual word contains limited spatial contextual information, which has been proven important for visual matching and recognition [16], [17], [31], [39]. Thus, it is not effective in presenting the characteristics of objects and scenes. This can be explained by an analogy between basic English alphabets and single visual words. The English alphabets, which are also basic components of documents, present very limited ability for describing semantics, if they are not organized in specific orders. Similarly, the spatial layouts of different visual words need to be taken into consideration to make the classic visual words descriptive enough. Fig. 2 illustrates the importance of spatial context.
2) Previous K-means-based visual vocabulary generation cannot lead to very effective and compact visual vocabulary [23], [31], [39]. This is because simply clustering the local descriptors in unsupervised way generates lots of unnecessary and nondescriptive visual words in the cluttered background, e.g., the noisy mismatched visual words in Fig. 1.

Aiming at the first problem, many works are conducted to combine multiple visual words to model their spatial relationships [1], [3], [17], [21], [27], [36], [39], [41]–[45]. As for the second problem, novel feature quantization algorithms [14], [15], [19], [22], [26], [38] have been proposed, targeting for more discriminative visual vocabularies. We will review the related works and state the differences and advantages of our algorithm in detail in Section II.

In order to overcome the above two shortcomings and generate visual vocabulary that is as comparable to the text words as possible, descriptive visual words (DVWs) and descriptive visual phrases (DVPs) are proposed in this paper. DVWs are defined as the individual visual words specifically effective in describing certain objects or scenes. Similar to the semantic meaningful phrases in documents, DVPs are defined as the distinctive and commonly co-occurring visual word pairs in images. Intuitively, because DVWs and DVPs only keep the descriptive visual words and visual word pairs, they would be descriptive, compact, and clean. Once established, they will lead to compact and effective BoWs representation.

Generating DVW and DVP set seems to be a very difficult problem, but statistics in large-scale image datasets might provide us some help. Because images are carriers of different visual objects or visual scenes, classic visual words and their combinations that are descriptive to certain objects or scenes could be selected as DVWs and DVPs, respectively. The corresponding DVWs and DVPs will function more similar to the text words than the classic visual words because of the reasons given here.

1) Only unique and effective visual words and combinations are selected. Thus, the selected set would be compact to describe specific objects or scenes. In addition, this significantly reduces the negative effects of visual words generated from the cluttered background. Therefore, the DVWs and DVPs would be more descriptive.
2) Based on the large-scale image training set containing different scenes and objects, DVWs and DVPs might present better descriptive ability to the real word and could be scalable and capable for various applications. Consequently, our algorithms identify and collect DVWs and DVPs from a large number of object and scene categories.

To gather reliable statistics on the large-scale image dataset, we collect about 376 500 images, belonging to 1506 categories, by downloading and selecting images from Google Image. We will give the details of our data collection in Section V-A. Fig. 3 illustrates the framework of our algorithm. A classic visual word vocabulary is first generated based on the collected image database. Then, the classic visual words appear in each category are considered as the DVW candidates, from which we will identify the DVWs that are descriptive for the corresponding categories. DVP candidates in each category are generated by detecting the co-occurring visual words within a certain spatial distance threshold. A novel visual-word-level ranking algorithm: VisualWordRank which is similar to the PageRank [2] and VisualRank [13] is proposed for identifying and selecting DVWs efficiently. Based on the proposed ranking algorithms, DVWs and DVPs for different objects or scenes are discriminatively selected. The final DVW and DVP set is generated by combining all of the selected DVWs and DVPs across different categories. Extensive experiments on image retrieval tasks show that the DVW and DVP present stronger descriptive power than the classic visual words. Furthermore, DVW and DVP show promising performance in image search reranking and object recognition tasks.

In summary, the contributions of our work are given here.

- The drawbacks of classic visual words are discussed. A novel large-scale web image-based solution is proposed for generating DVWs and DVPs.
- The idea of PageRank [2] and VisualRank [13] is leveraged in VisualWordRank for DVW selection. Experiments validate the effectiveness and efficiency of VisualWordRank.

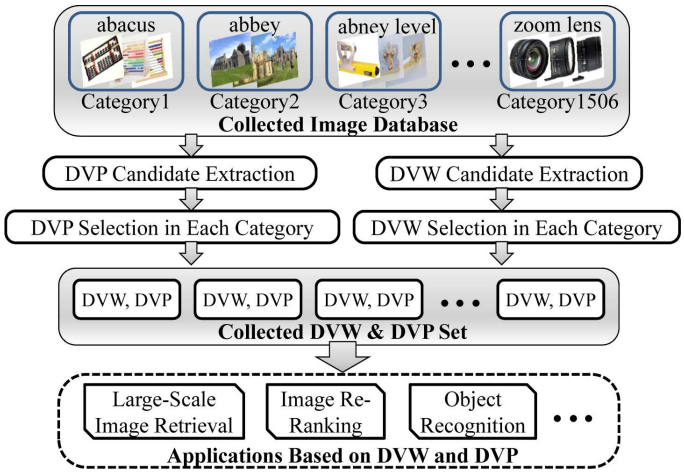


Fig. 3. Proposed framework for DVW and DVP generation.

- The proposed DVWs and DVPs are general and perform impressively in three applications: large-scale near-duplicated image retrieval, web image search reranking, and object recognition with simple nonparametric algorithms.

The remainder of this paper is organized as follows. Section II reviews and summarizes the related works on visual vocabulary. DVW and DVP candidate generation will be introduced in Section III. The DVW and DVP selection is presented in Section IV. Section V discusses the applications and evaluations. Finally, Section VI concludes the paper.

II. Related Work

To improve the descriptive power of visual vocabulary, many approaches have been proposed. These approaches can generally be divided into two categories, i.e., they either try to optimize the unsupervised clustering for feature quantization, or try to model more spatial information among the visual descriptors. In the following two paragraphs, we will review these algorithms in detail.

For visual vocabulary generated from unsupervised clustering, lots of noisy visual words can be generated from the local features in the cluttered background and large quantization error could be introduced. To overcome these shortcomings, many works have proposed novel feature quantization algorithms [15], [22], [26], targeting for more effective and discriminative visual vocabularies, e.g., an interesting work is reported by Lazebnik *et al.* [15]. Using the results of K-means as initializations, the authors generate discriminative vocabularies according to the Information Loss Minimization theory [15]. In [22], Extremely Randomized Clustering Tree is proposed for visual vocabulary generation, which shows promising performance in image classification. The visual word ambiguity and the influences of visual vocabulary size on quantization error and retrieval performance are studied in [7]. To reduce the quantization error introduced in feature space partition, soft-quantization [10], [27] quantizes a SIFT descriptor to multiple visual words.

In addition, to generate the visual vocabulary from single-image local descriptors, the K-means clustering commonly employs a general distance metric, such as Euclidean distance, to cluster or quantize the local features. This is unsatisfactory since it largely neglects the semantic contexts of the local features. With a general distance metric, local visual features with similar semantics may be far away from each other, while the features with different semantics may be close to each other. As a result, the local features with similar semantics can be clustered into different visual words, while the ones with different semantics can be assigned into the same visual words. This defection results in some incompact and noisy visual words, which are also closely related with the mismatches occurred between images.

There have been some works attempting to address this phenomenon by posing supervised distance metric learning [19], [38], [42], [45]. In [19], the classic visual vocabulary is used as the basis, and a semantic distance metric is learned to generate more effective high-level visual vocabulary. In a recent work [38], the authors capture the semantic contexts in each object category by learning a set of effective distance metrics between local features. Then, semantic-preserving visual vocabularies are generated for different object categories. Experiments on large-scale image database demonstrate the effectiveness of the proposed algorithm in image annotation. However, the codebooks in [38] are created for individual object categories, thus they are not universal and general enough, which limits their applications.

It has been illustrated that a single local feature cannot preserve enough spatial information in images, which has been to be proven important for visual matching and recognition [16], [17], [21], [27], [31], [39], [42], [45]. To combine BoWs with more spatial information, spatial pyramid matching is proposed to capture the hierarchical spatial clues of visual words in images [16]. Video Google utilizes structure-free spatial clues in neighboring visual words to remove the mismatched visual words between images [31].

Recently, many works have been conducted to seek visual word combinations to capture the spatial information among visual words [17], [21], [27], [39], [42], [45]. This may be achieved, for example, by using feature pursuit algorithms such as AdaBoosting [34], as demonstrated by Liu *et al.* [17]. Visual word correlogram and correlation [27], which are leveraged from the color correlogram [27], are utilized to model the spatial relationships among visual words for object recognition in [27]. In a recent work [39], visual words are bundled and the corresponding image indexing and visual word matching algorithms are proposed for large-scale near-duplicated image retrieval. Defined as descriptive visual word combination in [42], collocation pattern captures the spatial information among visual words and presents better discriminative ability than the traditional visual vocabulary in object categorization tasks. Generally, considering visual words in groups rather than single visual word could effectively capture the spatial configuration among them.

Although these approaches have shown impressive performance in many vision tasks, most of them are small-scale problem-oriented [15], [16], [19], [22], [26], [27], [42], [45] or do not take the spatial contexts into consideration [10], [15], [19], [22], [26], [27], [38]. Moreover, most of these generated visual vocabularies are specifically designed for one problem (i.e., for image or classification, image annotation), thus these proposed visual vocabularies are still not comparable with the text words, which could be used as effective features and perform impressively in various information retrieval tasks.

Our proposed algorithm is different from the previous ones in the following aspects.

1) We identify the DVWs and filter the noisy visual words, thus the shortcomings of unsupervised K-means clustering are depressed. Additionally, we extract DVPs to capture more spatial clues. Therefore, we integrate the two solutions in a joint framework. This is different from the previous works, which commonly only consider one of the two factors, i.e., optimizing unsupervised clustering, modeling more spatial contexts.
2) The DVWs and DVPs are capable to handle large-scale image datasets and show promising performance in three applications, i.e., large-scale image retrieval, objection recognition, and image search reranking. Therefore, our approach shows advantages in generalization ability and scalability than previous algorithms.

III. Candidate Generation

The DVWs and DVPs are defined as the representative visual words and co-occurring visual word pairs that are descriptive to certain objects or scenes, respectively. According to our framework in Fig. 3, we select DVWs and DVPs from their candidates in each category. The DVW candidates for a certain category are defined as the classic visual words appear in this category. While the DVP candidates for a certain category are defined as the co-occurring classic visual word pairs within a certain spatial distance. Thus, generating the classic visual vocabulary and identifying the appeared classic visual words in each training category are the first steps of our framework. Here, we first introduce how we generate the classic visual vocabulary, and then proceed to induce the generation of DVW and DVP candidates.

A. Classic Visual Vocabulary Generation

Similar to existing works [23], [39], we train classic visual vocabulary by clustering a large number of SIFT descriptors [20]. We adopt hierarchical K-means to conduct the clustering for its high efficiency. Though some other clustering methods such as Affinity Propagation [6] or some recent visual vocabulary generation methods [14], [15], [19], [22], [26], [38], could also be adopted, they are expensive to compute, in terms of either time or space complexity. Another advantage of hierarchical K-means is that the generated visual words can be organized in the vocabulary tree and the leaf nodes are considered as the classic visual words [23]. Thus, with the hierarchical structure, searching the nearest visual word for a local feature descriptor can be performed efficiently. More details about the vocabulary tree can be found in [23]. By searching hierarchically in the vocabulary tree, images in each training category are represented as BoWs representation by replacing their SIFT descriptors with the indexes of the corresponding nearest visual words [23]. During this process, the scale of each local feature is kept for the corresponding visual word to achieve scale invariance when computing the DVP candidates.

B. DVW Candidate Generation

Recall that the DVW candidates for a certain category are defined as the classic visual words appearing in this category. In our experiment, for a vocabulary tree with 32357 visual words,

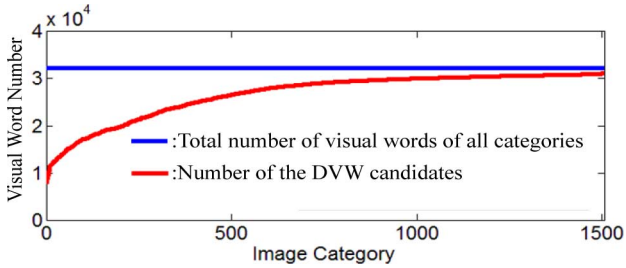


Fig. 4. Sorted number of DVW candidates in the 1506 categories.

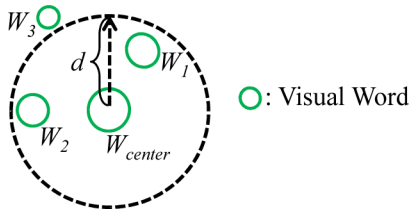


Fig. 5. Utilized DVP candidate detector.

the corresponding numbers of appeared visual words, i.e., the DVW candidates, in 1506 categories are sorted in ascending order and shown in Fig. 4. Obviously, the DVW candidates in each category are portions of the total visual vocabulary (i.e., the blue line, 32 357 classic visual words). It can be inferred that only parts of the entire visual vocabulary are descriptive to the corresponding categories. Thus, selecting DVWs from their candidates would be more efficient and reasonable than from the entire visual vocabulary.

C. Descriptive Visual Phrase Candidate Generation

In literature, different algorithms are proposed for capturing the spatial clues among visual words, e.g., the spatial histogram proposed in [17]. However, these algorithms are expensive to compute, additionally, capturing complicated spatial relationships commonly causes the sparseness of the generated visual word combinations [17] and accumulates the quantization error introduced in the visual vocabulary. Therefore, we capture the simple co-occurring clues between two visual words, and the corresponding DVP candidates for a certain category are defined as the co-occurring classic visual word pairs in this category.

Suppose visual word i and j co-occur in an image category C. Then, the DVP candidate containing the two visual words for this category can be denoted as

$$\mathrm{DVPCandidate}^{(C)}\left[i, j, T_{i,j}^{(C)}\right]$$

where $T_{i,j}^{(C)}$ is the overall average frequency of co-occurrence computed between the visual word i and j in image category C, e.g., if visual word i and j frequently co-occur in the category C, $T_{i,j}^{(C)}$ will present a large value. Hence, $T_{i,j}^{(C)}$ reflects the strength of their spatial relationship in category C.

In order to identify co-occurring visual word pairs, we define a spatial distance d which is related to the constraint of co-occurrence. As illustrated in Fig. 5, each visual word co-occurring with the visual word W_{center} within the distance d composes a DVP candidate with W_{center}.

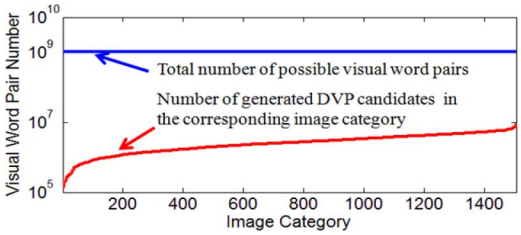


Fig. 6. Number of DVP candidates in each image category of our training set, which contains 1506 image categories.

As shown in Fig. 5, the distance d is an important parameter related to the constraint of co-occurrence. Because objects may have various scales, we compute the d in

$$d = \mathrm{Scale}_i \bullet P_d \tag{1}$$

to achieve scale invariance, where Scale_i is the scale of the interest point [20] from which the instance of visual word i is computed, and P_d controls the constraint of co-occurrence. Intuitively, if an image is magnified, the co-occurrence relationships among the visual words within it remain the same because of the magnified *Scale*. From our experiments, larger P_d is necessary for identifying reliable spatial co-occurrence between two visual words and overcoming the sparseness of the generated DVP candidates. However, large P_d also increases the computational cost and the occurrence of noise. In this paper, we experimentally set P_d as 4, which is a good tradeoff between efficiency and performance.

The DVP candidates can be identified by scanning the neighborhood of each visual word with the detector in Fig. 5. Meanwhile, the co-occurrence frequency $T_{i,j}^{(C)}$ can be computed by counting the time of co-occurrence within the spatial distance d between visual word i and j in category C.

The numbers of generated DVP candidates in each image category are sorted and presented in Fig. 6. We can observe that, although the generated candidates are only small portions of the entire possible visual word pairs (32357^2), their sizes are still very huge. Therefore, effective and compact DVP set needs to be selected from the candidates.

IV. DVW and DVP Selection

A. DVW Selection

DVWs are defined as the representative visual words that are descriptive to certain objects or scenes. It is designed to describe certain categories, thus several unique features are desired in them.

1) If one object or scene appears in some images, the DVWs descriptive to it should appear more frequently in these images. Also, they should be less frequent in images that do not contain such object or scene.
2) They should be frequently located on the object or scene, even though the scene or object is surrounded by cluttered background.

Inspired by PageRank [2], we design a novel visual-word-level ranking algorithm: VisualWordRank to combine the two clues for DVW selection.

According to the first criterion, the frequency of occurrence of DVW candidates in the total image set and in each individual image category would be an important clue for identifying DVWs. Fig. 7(a)–(d) shows the frequencies of occurrence of visual words with index number: 1×10^4–2×10^4 in four categories. The frequencies shown are normalized between 0 and 1. It is clear that, the same visual words (e.g., visual words with index number 14 000–16 000) present different frequencies in different image categories. Thus, their different significances for each category can be indicated.

Besides the frequency information of single visual word, if two visual words frequently co-occur within short spatial distance in images containing the same object or scene, strong spatial consistency could be inferred between them in such images. Considering that these images contain the same object but different backgrounds, the spatially consistent visual words are more likely to be located on the foreground and the object. Hence, the spatial co-occurrence frequency between two visual words, i.e., $T_{i,j}^{(C)}$ is adopted in DVW selection to depress the negative influences caused by the cluttered background. As a result, the second criterion can be met.

Therefore, we use two clues: 1) each DVW candidate's frequency information and 2) its co-occurrence with other candidates to identify DVWs. This can be formalized as a visual word ranking problem which is very similar to the one of webpage ranking. Thus, we propose the VisualWordRank algorithm which leverages the idea of well-known PageRank [2]. In PageRank, a matrix is built to record the inherent importance of different webpages and the relationships among them. Iterations are then carried out to update the weight of each webpage based on this matrix. After several iterations, the weights will stay stable and the final significance of each webpage is obtained combining both its inherent importance and the relationships with other webpages [2].

Based on the same idea, for an image category C, we build a $\mathrm{VWnum}^{(C)} \times \mathrm{VWnum}^{(C)}$ matrix $R^{(C)}$ to combine the frequency and co-occurrence cues for DVW selection. $\mathrm{VWnum}^{(C)}$ is the number of DVW candidates for category C. In matrix $R^{(C)}$, we define the diagonal element as

$$R_{i,i}^{(C)} = f_i^{(C)} \Big/ \ln(F_i) \tag{2}$$

where i is a DVW candidate and F_i and $f_i^{(C)}$ denote its average frequency in all categories and the *within-category* frequency in category C, respectively. $R_{i,i}^{(C)}$ stands for the *inherent-importance* of candidate i. Thus, i would be inherently more significant to category C if $R_{i,i}^{(C)}$ has larger values. $f_i^{(C)}$ and F_i are computed beforehand when transforming the images in training dataset into BoWs representations.

The nondiagonal element $R_{i,j}^{(C)}$ is defined as the average co-occurrence frequency of visual word i and j as

$$R_{i,j}^{(C)} = T_{i,j}^{(C)} \tag{3}$$

where $T_{i,j}^{(C)}$ is computed during DVP candidate generation.

After computing $R^{(C)}$, we normalize the diagonal elements and nondiagonal elements, respectively and assign them with weights W_{freq} and W_{cooc}, respectively. The two input weights control the influences of frequency factor and co-occurrence factor, respectively. From extensive experiments, we conclude

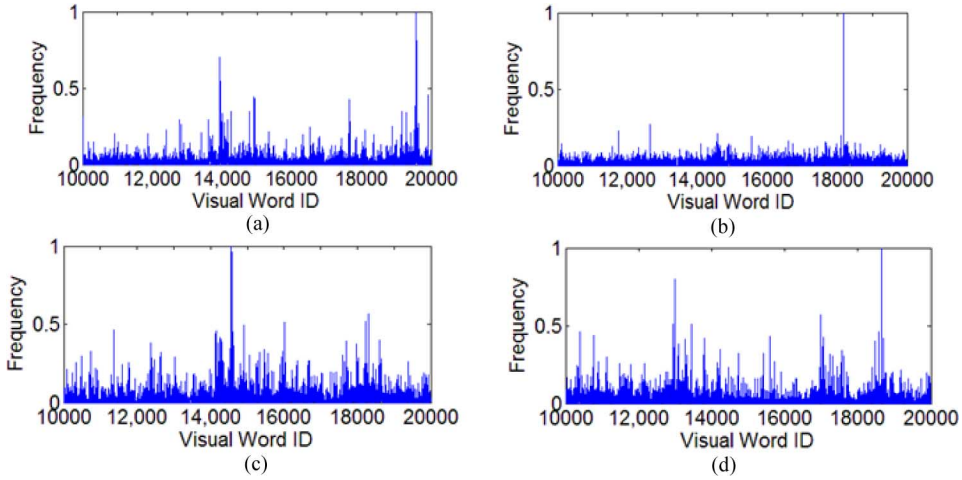


Fig. 7. Visual word frequencies in different categories. Frequency in: (a) "cell phone," (b) "airplane," (c) "ant," and (d) "bike."

that setting the two weights equal value results in good performance for most of the image categories.

Algorithm 1: VisualWordRank

Input: $R^{(C)}$; maximum iteration time: *maxiter*.

Output: The rank value of each DVW candidate to the category C :

$Rank_i^{(C)}, i = 1, \ldots, VWnum^{(C)}$

Initialize each element in the $VWnum^{(C)} \times 1$ sized rank vector: $OldRank^{(C)}$ as 1; **Normalize** the sum of each column of $R^{(C)}$ as 1 [2]; **Set** $iter = 0$

While $iter < maxiter$

$NewRank^{(C)} = R^{(C)} \cdot OldRank^{(C)}$

If $(|NewRank^{(C)} - OldRank^{(C)}| \leq \varepsilon)$ **break**

$OldRank^{(C)} = NewRank^{(C)}$

$iter{+}{+}$

End

$Rank^{(C)} = NewRank^{(C)}$

With the matrix $R^{(C)}$, we set the initial rank value of each DVW candidate equal and then start the rank-updating iterations. The detailed descriptions of VisualWordRank are presented in Algorithm 1. Intuitively during the iteration, the candidates having large *inherent-importance* and strong *co-occurrence* with large-weighted candidates will be highly ranked. After several iterations, the DVWs in object category C can be identified by selecting the top N ranked candidates or choosing the ones with rank values larger than a threshold.

Fig. 8(a) shows the DVW candidates in image categories: *butterfly, ceiling_fan, ant*, and *crab*. The selected DVWs in the corresponding categories are presented in Fig. 8(b). Obviously, although there are many candidates (i.e., classic visual words) on the cluttered background, most of the selected DVWs appear on the object. In order to show the descriptiveness of the selected

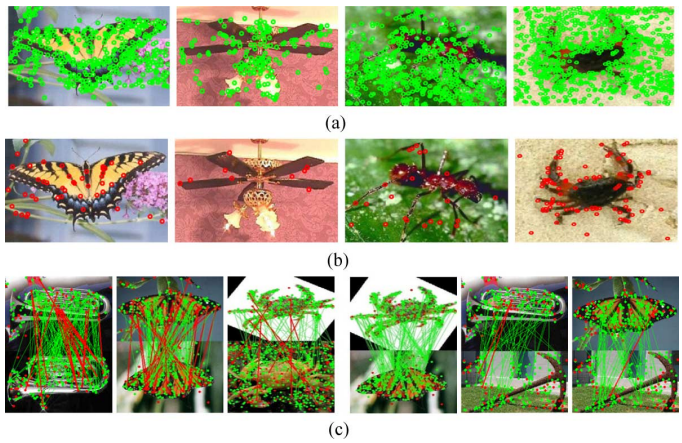


Fig. 8. DVW candidates, the selected DVWs, and the matched DVWs (red lines) and matched visual words (green lines) between the same and different objects. (a) DVW candidates before VisualWordRank. (b) Selected DVWs in corresponding categories. (c) Matched DVWs and visual words between same and different objects.

DVW set, the matched classic visual words and matched DVWs between same and different objects are compared in Fig. 8(c). In the figure, visual words and DVWs are denoted by green dots and red dots, respectively. The identical visual words and DVWs across images are connected by green lines and red lines, respectively. In the left three images, matches are conducted between same objects. It can be observed that, though some DVWs exist on the background, most of the matched ones locate on the object. In the right three figures, which show the matched DVWs and classic visual words between different objects, lots of classic visual words are wrongly matched. Nonetheless, there are very few mismatches occurred between DVWs. Thus, it can be observed that DVWs are more descriptive and more robust than classic visual words. The detailed evaluations of DVWs are presented in Section V.

B. Descriptive Visual Phrase Selection

Similar to the DVW selection, the DVP selection is desired to select the visual word pairs descriptive to certain objects or scenes. Since the co-occurrence information of visual word

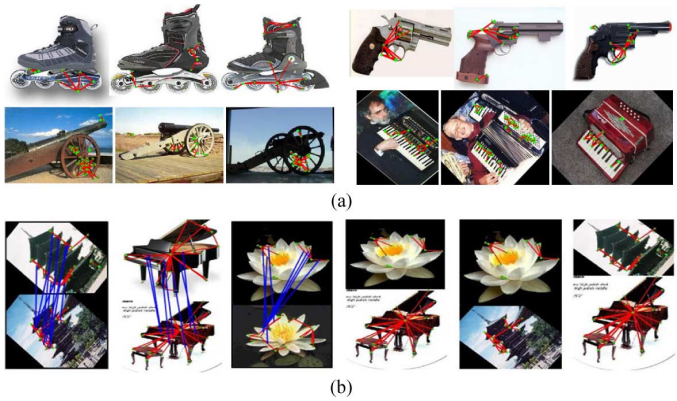


Fig. 9. Selected DVPs and the matched DVPs between the same and different objects. (a) Selected DVPs in: "inline skate," "revolver," and "cannon." (b) Matched DVPs between the same and different objects.

pair has already been integrated in the generated DVP candidates (i.e., the DVP candidates with high frequency of co-occurrence $T_{i,j}^{(C)}$ have high spatial consistency and strong spatial relationships in category C), we now compute the DVP candidate frequencies within a certain category and the overall categories. According to the TF-IDF weighting in information retrieval theory, a DVP candidate is considered important to a category if it appears more often in it and less often in others. Based on this strategy, the importance of a DVP candidate k to the category C is computed as

$$VPI_k^{(C)} = VPf_k^{(C)} \Big/ \ln(VPF_k) \tag{4}$$

where $VPI_k^{(C)}$ is the importance of the DVP candidate k to the category C, and $VPf_k^{(C)}$ and VPF_k stand for the frequencies of occurrence of DVP candidate k in category C and all categories, respectively. Suppose there are M image categories and visual word i and visual word j are contained in DVP candidate k, then the $VPf_k^{(C)}$ and VPF_k can be computed with

$$VPf_k^{(C)} = T_{i,j}^{(C)} \quad VPF_k = \sum_{m=1}^{M} T_{i,j}^{(m)} \Big/ M. \tag{5}$$

Consequently, after computing the importance of each DVP candidate, the DVPs for category C could be identified and selected by ranking the candidates based on $VPI^{(C)}$.

In Fig. 9(a), the visual words are denoted as green dots and the dots connected by red lines denote the selected DVPs. Because there are dense visual words on the background in each image, it can be inferred that there would be a lot of DVP candidates generated on the object and background. As we can clearly observe, most of the selected DVPs appear on the object and maintain obvious spatial characteristics of the corresponding object. Fig. 9(b) shows the matched DVPs across same and different objects. All of the DVPs in the example images are denoted as red lines and the matched ones are connected by blue lines. It can be seen that, many DVPs are correctly matched between the same objects, while between images containing different objects, none of the DVPs is matched. Therefore, it can be concluded that the selected DVPs are valid and descriptive.

After selecting DVWs and DVPs in each category, the final DVW and DVP set can be created by combining all of the selected candidates across different categories. Since the DVWs and DVPs are descriptive for certain objects or scenes, the final DVW and DVP sets are desired to be descriptive and general. Further tests on DVWs and DVPs are carried out in Section V.

C. Discussion About the Computational Complexity

The generation of DVWs and DVPs mainly consists of three steps: classic visual word generation, candidate extraction, and DVW, DVP selection. The classic visual word generation is finished efficiently with hierarchical K-means clustering [23]. The DVW candidate extraction is finished by simply counting the frequency of the visual words appeared in each category. As for the DVP candidate generation, because of the limited number of local features in images (typically 500 for a 480×320 sized image), and the properly selected distance d in (1), this operation is also efficient by linearly scanning the images with the detector illustrated in Fig. 5 in each category.

The most time consuming operation in our algorithm should be the DVW and DVP selection. Suppose the number of candidates in a category is N, the complexity of the VisualWordRank would be $\mathrm{O}(N^2)$. Because of the limited number of DVW candidates in each category (the average number is about 25 000 in Fig. 4), and the fast convergence of the random walk algorithm [2], the efficiency of this process is still acceptable. The extraction of DVWs for 1506 categories can be finished within one day on a server with 2.9-GHz CPU, 8-GB memory. The DVP selection is efficient by computing the (5) and sorting the DVP candidates by their importance. Thus, the complexity would be $\mathrm{O}(N \log N)$.

V. Applications and Evaluations

A. Image Dataset Collection

1) Image Category Collection for DVW and DVP Generation: The DVW and DVP generation is based on the statistics of their candidates in different image categories. Moreover, the DVW and DVP sets are desired to be semantically meaningful, descriptive, and general for different objects and scenes. Thus, we spend a huge amount of time and energy to systematically select our training dataset. The raw image dataset is collected with the method similar to [4] and [33]. We first use WordNet [5] to get a comprehensive list of objects and scenes by extracting 117 097 nonabstract nouns. The extracted list is then used for searching and downloading image categories from Google Image. The top 250 returned images of each query are saved. The downloading task is finished within one month by 13 servers and 65 downloading processes. In the collected raw database, categories with images less than 100 are removed. Then, from the remaining images, we carefully select 1506 categories with visually consistent single objects or scenes, by viewing the thumbnails in each category. Finally, we form a dataset composed of about 376 500 images. The final dataset sufficiently covers the common visual objects and scenes. Thus, extracting DVWs and DVPs based on it would be statistically reasonable.

Based on the collected dataset, a vocabulary tree containing 32357 visual words is generated. We do not generate larger numbers of visual words because of the following three considerations: 1) large visual vocabulary results in huge number of possible visual word pairs and low repeatability of the DVP candidate; 2) single visual word shows limited descriptive ability, no

TABLE I
QUERY WORDS OF THE SELECTED TRAINING CATEGORIES AND CORRESPONDING TEST CATEGORIES FOR OBJECT RECOGNITION

Query word	*Piano Accordion*	*Pocket Calculator*	*Dueler*	*Euphonium*	*Golden Gate Bridge*
Test category	Accordion	Calculator	Car-tire	Euphonium	Golden-Gate-Bridge
Query word	*Headphone*	*Semiautomatic Pistol*	*Panda*	*Lotus*	*Scissors*
Test category	Headphone	Revolver	Panda	Lotus	Scissors
Query word	*Adjustable Wrench*	*Motorbike*	*Hockey Skate*	*Spinet*	*Lander-Back Chair*
Test category	Wrench	Motorbike	Inline-skate	Grand-piano	Windsor-chair

matter how fine-grained it is [17], [23], [27]; and 3) we evenly select the training images from the representative database to get a better description of the feature space as much as possible. Based on the generated visual words, the entire image dataset (376 500 images) is then used for candidate generation and final DVW and DVP selection.

2) Dataset Collection for Large-Scale Image Retrieval: In order to test the DVW and DVP in large-scale image retrieval, we first build a one-million basic image dataset by crawling images from the Internet. To finish this, we build a web-image crawler which recursively downloads webpages and extracts the URLs of images on them. Then we download images according to these URLs. This is a similar process of the one in bundled feature [39]. Then, we manually download 315 images belonging to ten categories, including "Abbey Road," "American Gothic," "Pisa Tower," as the image set with ground-truth labels. The images in each category are partial duplicates of each other. Similar to [39], we add these labeled images into the basic dataset to construct an evaluation dataset for large-scale near-duplicated image retrieval.

3) Dataset Construction for Image Search Re-Ranking: An image re-ranking dataset is created by first selecting 40 image categories from the image database collect by Google Image. Each selected category contains 250 images and presents single visual concept (i.e., same objects or scenes). Hence, we assume all of the 250 images are relevant to the concept. Then, 100 randomly selected images are added to each of these categories. Finally, we construct a dataset containing 40 categories and 14 000 images as our evaluation dataset.

4) Training Set and Test Set Collection for Object Recognition: We select 15 commonly used object categories from the Caltech 101 and Caltech 256 datasets as the test set. For each test category, the training category containing the same object is selected from the image database collected from Google Image. The query words of training categories and the corresponding test categories are listed in Table I. Note that each training category contains 250 images returned from Google Image, and each category contains some noisy images.

B. Large-Scale Image Retrieval Based on DVW and DVP

In recent work, BoWs image representation has been proven promising in large-scale image retrieval [23], [39] by leveraging the classic information retrieval algorithms such as inverted file indexing and TF-IDF weighting. In this part, experiments are carried out to compare the state-of-the-art algorithms with the proposed DVWs and DVPs on large-scale near-duplicated image retrieval tasks. Near-duplicated image retrieval differs with common image retrieval in that the target images are usually obtained by editing the original image with changes

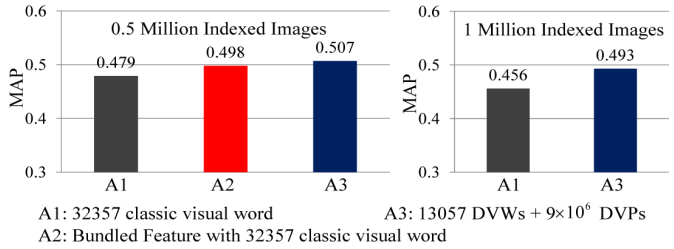


Fig. 10. Comparison of MAP among three features.

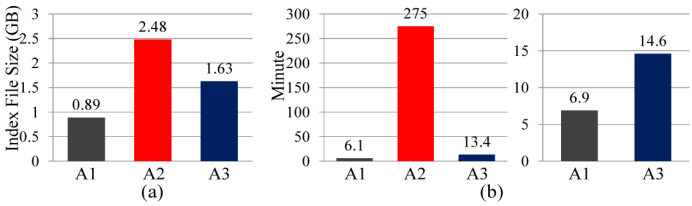


Fig. 11. Comparisons of memory consumption and efficiency. (a) Size of the index file when 0.5 million images are indexed. (b) Total time needed by the three features to retrieve 315 images.

in color, scale, or partial occlusion. In near-duplicated images, different parts are often cropped from the original image and pasted in the target image with modifications. The result is a partial-duplicated version of the original image with different appearances.

Our large-scale image dataset is introduced in Section V-A. Each image in the database is first represented as BoWs, with the classic visual word [23], DVW, and DVP, respectively. Then, the images are indexed using inverted file structure. In the retrieval process, TF-IDF weighting [23] is applied for similarity computation. All of the images with ground truth i.e., the 315 images, are used as queries. For each query, we compute the MAP, which takes the average precision across all different recall levels in the first 30 returned images. The DVW and DVP combination, classic visual word [23], and bundled feature [39] are compared. Fig. 10 shows their overall MAPs in image datasets with different image numbers.

From Fig. 10, it is clear that the A2 (i.e., bundled feature) and A3 (i.e., DVW and DVP combination) perform better than the classic visual word. This is because they capture more spatial cues by combining several visual words. It is also obvious that A3 outperforms A2. The reason why we do not test the bundled feature in larger image databases (i.e., 1 million images) is because the index size of bundled feature is large, and 0.5 million is the maximum image number that the 4.0-GB memory of our computer could handle. The sizes of index files of the three features are compared in Fig. 11(a).

Intuitively from Fig. 11(a), the bundled feature needs larger memory to load the index for image retrieval. This is because for each visual word, it needs to store certain numbers of 19-b "bundled bits [39]," which records the spatial contexts of visual words in each image. The bundled bit number equals to the number of bundled features where this visual word appears. Thus, in addition to 32-b image ID and the 16-b visual word frequency, extra space is needed, resulting in the large index file. Differently, for DVP and DVW based image index, we only need to store the image ID and the frequency for each DVP and DVW. Thus, the DVP and DVW based image index captures spatial contexts with more compact index size. It should be noted that the size of inverted file index is largely decided by two factors: the total number of images and the average number of classic visual words/DVWs/DVPs contained in each image. Although the DVP set size is significantly large, the average DVP number in each image is limited. This is because the DVP set only contains the descriptive and stable visual word pairs and discards most of the unstable ones in images. Therefore, the index size based on DVP is limited. As shown in Fig. 11(a), the index size based on DVP + DVW is acceptable, i.e., 1.63 GB for 0.5 million images. We will further discuss the possible solutions to make the DVP set more compact in Section V-E.

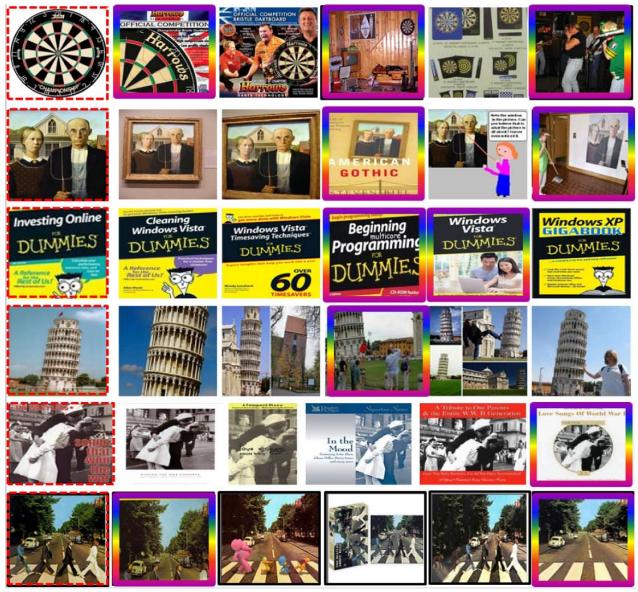
Besides the comparisons of precision and memory consumption, the efficiency is compared in Fig. 11(b). From the figure, it can be observed that bundled feature is time consuming. This is because the spatial verification between bundled features is carried out during the retrieval process [39], and large memory is needed to store the spatial configuration of the retrieved images for the spatial verification. Consequently, we can conclude that, the DVP and DVW show better performance than the bundled feature [39] and classic visual word in large-scale near-duplicated image retrieval. In addition, the DVP and DVW are proven better than the bundled feature in efficiency and memory consumption.

Fig. 12 shows some examples of DVP and DVW based near-duplicated image retrieval before the return of first false positive images, and the matched DVPs between queries and retrieved images. Obviously, although the images are edited by affine transformations, cropping, and cutting, they still can be retrieved with DVW and DVP. It is also obvious that DVPs between two near-duplicated images can be correctly matched. The images which cannot be retrieved by classic visual word are highlighted by the color boxes. We can infer that the classic visual word is not effective in retrieving the near-duplicated images with large cropping and cutting, which introduce more cutter background, and noisy visual words.

In order to show the difference between DVW and DVP, and compare their performances, we carry out further experiments on image retrieval. We choose Corel 5000 as the testset because it is a widely used benchmark dataset in CBIR community. In addition, it contains both rigid and nonrigid objects, thus is more general and fair for image retrieval tasks. In this dataset, 50 image categories are included and each contains 100 images. All of the 5000 images are indexed and used for retrieval.

To make the performance comparisons between classic visual words and DVWs, DVPs more visible, we use *PrecisionRatio* computed with

$$\text{PrecisionRatio}_k = \text{Precision}_k^{(a)} \Big/ \text{Precision}_k^{(b)} \qquad (6)$$



Some results of near-duplicated image retrieval based on DVW and DVP

Some examples of matched DVPs between two images

Fig. 12. Results of near-duplicated image retrieval and matched DVPs.

as a measurement, where $\text{Precision}_k^{(a)}$ and $\text{Precision}_k^{(b)}$ are the retrieval precision based on two different image features a and b (i.e., DVW, DVP, or classic visual word) in the first k returned images, respectively. Thus, if $\text{PrecisionRatio}_k = 1$, these two image features show the same performance.

As shown in Fig. 12, although the dartboards are different in scales and surrounding backgrounds, they still share stable spatial contexts, and thus their DVPs can be correctly matched. Therefore, we can conclude that the DVP captures more spatial information and is descriptive to the images containing rigid objects or stable spatial contexts. To further illustrate this conclusion, we first carry out some experiments showing the cases where the DVPs work or may fail. The 9×10^6 DVPs are used as feature a, and the 32 357 classic visual words are used as feature b. The $\text{PrecisionRatio}_{20}$ for several image categories are computed with (6) and are shown in Fig. 13. Obviously, DVPs work well for the image categories in Fig. 13(a), which contain stable spatial contexts. As for the nonrigid scene images in Fig. 13(b), because they lack stable spatial contexts, the DVPs cannot describe them effectively. As a result, the classic visual word outperforms the DVP.

Fig. 14 demonstrates the performance comparisons between classic visual words and DVWs in the entire dataset. The classic visual word [23] is used as feature b. Different numbers of DVWs are collected from the training image categories. The ratio curves in the figure are computed based on the overall average precisions of the 5000 queries. From Fig. 14, it can be seen that DVW set with the size 13 057 shows obvious

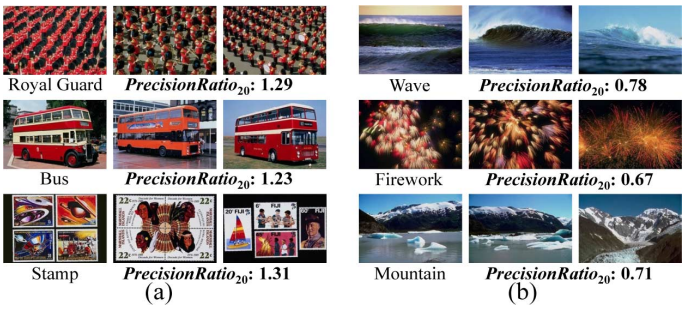


Fig. 13. Cases where DVP (a) outperforms the classic visual word and (b) fails.

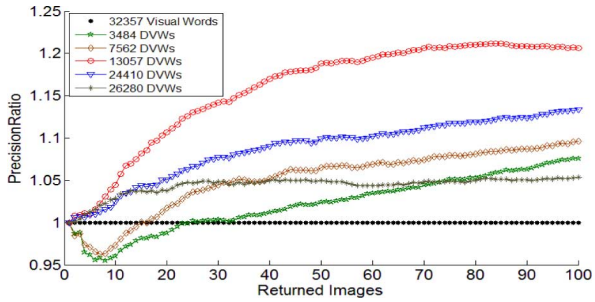


Fig. 14. Performance comparison between DVW and classic visual word.

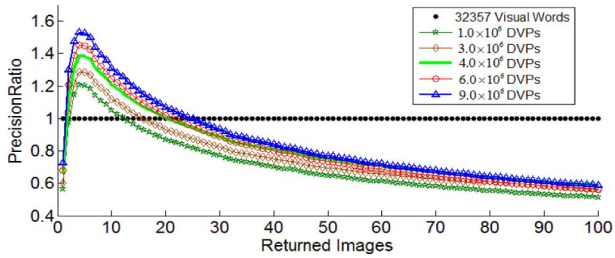


Fig. 15. Performance comparison between DVP and classic visual word.

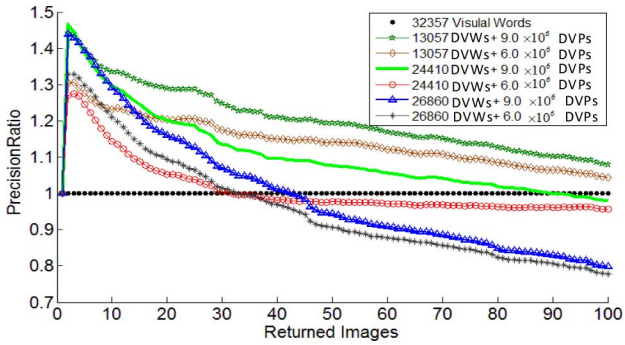


Fig. 16. Comparison among classic visual word, DVW, and DVP.

improvements over the classic visual words. This result proves that DVW set has stronger descriptive ability with more compact size. It is also interesting in Fig. 14 that DVW sets with the sizes 3484 and 7562 show worse performance in the first 25 returned images, but outperform classic visual words when more images are returned. This can be explained by the fact that, for the relevant images presenting weak visual similarities to the query image (e.g., the relevant images ranked after 25 in the returned image list), their similarities with the query image are more likely to be disturbed by the negative effects of cluttered background. Because the DVW set with small size keeps the most descriptive visual words and has removed most of the noisy ones, the background noise is depressed. Consequently, DVWs perform better than the classic visual words in the case where more noises exist. Since DVWs are selected from classic visual words, DVW sets with larger sizes will contain more noises and will function more similar to the classic visual words. This could explain why if more DVWs are selected (e.g., DVW set with the size 26 280), the performance will start to decrease. Therefore, we could conclude that DVW is more compact and descriptive than the classic visual word.

To evaluate the performance of the DVPs, we adopt the classic visual words as the baseline. The DVP numbers and the corresponding experimental results are presented in Fig. 15. From the figure, it can be observed that the DVP set with larger number shows better performance. This indicates valid DVPs are selected by our algorithm from the huge possible visual word pair space. Since DVP candidates contain both spatial and appearance cues, they are assumed to be more informative than the classic visual words. This might be the reason why the performance of DVPs remains increasing even with large size. It can also be observed that image retrieval based on DVPs cannot guarantee that the first returned image is the query one. This is because some nonrigid query images in categories such as "Beach" and "Wave" do not present consistent spatial contexts and contain very few or even zero DVPs. Thus, DVPs do not work well for these cases. As we discussed before, the DVPs are more effective in recognizing the near-duplicated images of the query one. This could be the reason why DVPs show obvious advantages in the first several returned images but perform worse when the returned images exceed certain numbers. From Figs. 14 and 15, it can be observed that DVPs and DVWs can be complemented to each other. Thus, the performance of DVW + DVP is further evaluated in Fig. 16.

Obviously in Fig. 16, medium number of DVWs plus a large number of DVPs show the best performance. The combination containing 13 057 DVWs and 9×10^6 DVPs shows the best performance and outperforms the classic visual words by 19.5% in term of MAP computed in the top 100 returned images. Accordingly, we can conclude that DVWs and DVPs are more descriptive for image retrieval than the widely used classic visual words.

C. Image Re-Ranking

Image search re-ranking is a research topic catching more and more attentions in recent years [9], [10], [18], [32]. The goal is to resort the images returned by text-based search engines according to their visual appearances to make the top-ranked images more relevant to the query. Generally, image re-ranking can be considered as identifying the common visual concept (i.e., scene, object, etc.), in the returned images and re-ranking the images based on how well each one fits the identified concept. DVWs and DVPs are effective in describing the objects and scenes where they are selected. Therefore, they can be utilized to measure the relevance between images and the concept. Based on this idea we proposed the DWPRank, which is detailed in Algorithm 2. We first carry out DWPRank on our database where each category contains the top 250 images returned from Google Image. Fig. 17 presents an example.

Extensive tests of DWPRank are carried out by comparing it with VisualRank on the image re-ranking testset introduce in

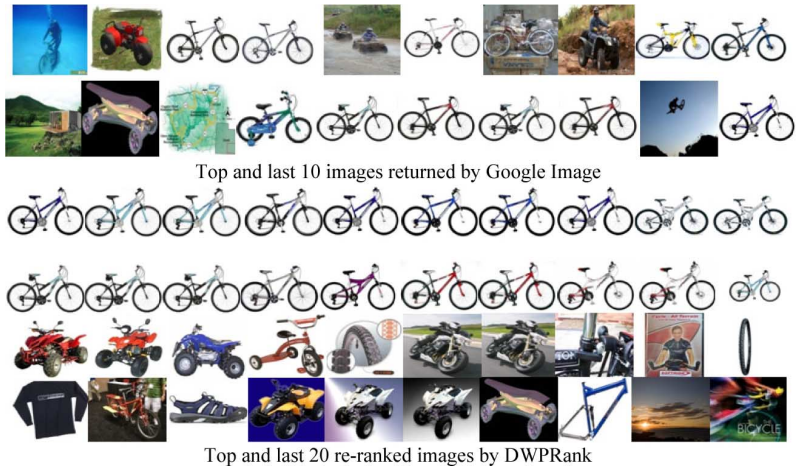


Fig. 17. Re-ranked images with query "**all-terrain bike**".

Section V-A. AP (Average Precision) computed in (7) is adopted to measure the effectiveness of the two algorithms.

$$\mathrm{AP} = \left(\sum_{i=1}^{250} \mathrm{correct}_i / i\right) \Big/ 250 \tag{7}$$

where, $\mathrm{correct}_i$ is the number of relevant images in the top i re-ranked images. Thus, if $\mathrm{AP} = 1$, it can be inferred that all of the 250 relevant images are in the top re-ranked image list, which is the most ideal case in image re-ranking.

Algorithm 2: DWPRank

Input: Images returned from the image search engine: $I_i, (i = 1, \ldots, N)$; weight of DVW and DVP: W_{DVW}, W_{DVP}.

Output: Re-ranked image list: $IReRanked_i, (i = 1, \ldots, N)$

Suppose: $Rel_i, (i = 1, \ldots, N)$ describes the relevance between image I_i and the query concept.

In $I_i, (i = 1, \ldots, N)$, generate the DVW and DVP candidates.

In $I_i, (i = 1, \ldots, N)$, select DVWs and DVPs.

For $i = 1 : N$ **do** $Rel_i = 0$

  For each DVW or DVP candidate D in image i **do**

    if (D is a DVW) $Rel_i = Rel_i + W_{DVW}$

    if (D is a DVP) $Rel_i = Rel_i + W_{DVP}$

  End

End

For $i = 1 : N$ **do**

  Find I_m which has the *i-th* largest Rel value.

  $IReRanked_i = I_m$

End

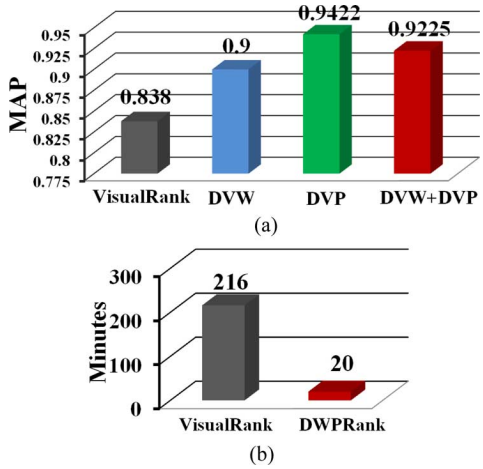


Fig. 18. The comparisons of MAP and efficiency (a) The MAP obtained by VisualRank and DWPRank, (b) Average time needed by VisualRank and DWPRank.

In our experiment, we run the standard VisualRank algorithm and DWPRank on the collected image database. 150 DVWs and 6000 DVPs are selected from each category. Three groups of DWPRank based on DVW, DVP and DVW + DVP are carried out by setting $W_{\mathrm{DVW}}, W_{\mathrm{DVP}}$ in Algorithm 2 as (1, 0), (0, 1) and (1, 1) respectively. Fig. 18 presents the results.

Obviously, from Fig. 18, DWPRank outperforms VisualRank. This is mainly because of two aspects: 1) more information and constrains (i.e., spatial and frequency clues) are considered in DVW and DVP, thus DVWs and DVPs are more effective in identifying and describing the visual concepts in returned images and 2) VisualRank computes the image-pair similarities based on all of the SIFT descriptors in each image, thus the cluttered background might disturb its performance. Differently, such influences are depressed in DWPRank through DVW and DVP selection. From Fig. 18, it can be also seen that compared with DVWs, DVPs are more effective in image re-ranking. Again, this can be explained by the fact that DVPs are more descriptive with more spatial information. We conclude that improvements of 7.4%, 12.4%, and 10.1% over the VisualRank are achieved by DWPRank with DVW, DVP and DVW+DVP, respectively.

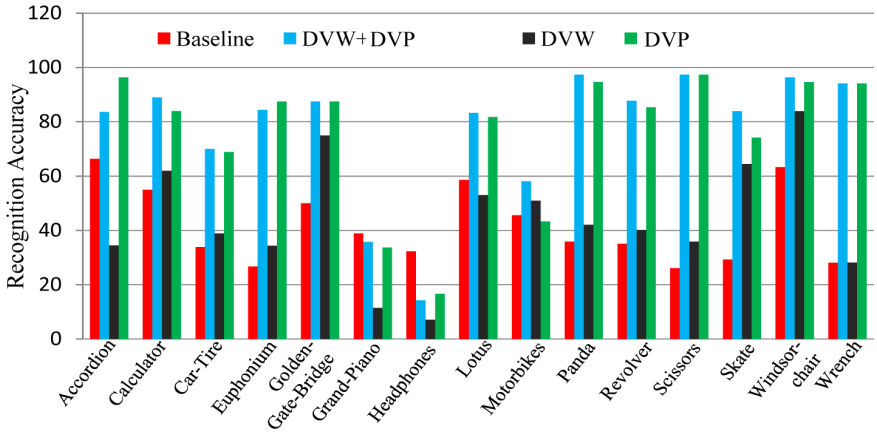


Fig. 19. Comparisons of object recognition among DVWs, DVPs and classic visual words (baseline).

Besides the improvements on accuracy, it is necessary to point out that, DWPRank is more efficient than VisualRank. The average time needed by VisualRank and DWPRank for re-ranking 350 images is compared in Fig. 18(b). Obviously, about $11\times$ improvement is achieved by DWPRank. The low efficiency of VisualRank is mainly rooted in the expensive image similarity computation based on SIFT and LSH [8]. VisualRank first maps each SIFT feature into L, i.e., 40, hash tables, each table with K, i.e., three hash functions. Then, to check if two features are matched across two images, VisualRank checks if they share three identical hash tables. This step is time consuming. For instance, if two images have M_1 and M_2 features. Then the total checking operation is $M_1 \cdot M_2 \cdot (K \cdot L)^2$. However, in DWPRank, DVP candidate generation and DVW selection, which are the most time-consuming operations, can be finished efficiently.

D. Object Recognition

Since DVWs and DVPs are designed to effectively describe certain objects or scenes. It is straightforward that the DVWs and DVPs of each image category should be discriminative for the corresponding object. Consequently, we utilize the object recognition task to test their discriminative ability. Moreover, this experiment is also carried out to test the validity of our algorithm in improving the discriminative power of original visual words, form which DVWs and DVPs are generated.

In the experiment, we first identify and collect 150 DVWs and 6000 DVPs from each training category. Then, for each object, we establish three discriminative feature pools containing DVWs, DVPs and both of them, respectively. In the testing phase, a naïve vote-based classifier is utilized, e.g., if most of the DVW candidates of an image appear in the DVW feature pool of "Accordion," then this image will be recognized as "Accordion." Similarly, another two recognition results based on DVP and $\mathrm{DVW} + \mathrm{DVP}$ can also be obtained. In the baseline algorithm, each test image is recognized by computing its ten nearest neighbors in the training dataset. Classic visual word histogram is computed in each image, and histogram intersection is used as the distance metric. Note that, since simple nonparametric classifiers are used, the discriminative abilities of these features can be clearly illustrated. Fig. 19 presents the experimental results.

Obviously from Fig. 19, the DVWs and DVPs outperform the baseline algorithm by a large margin for most of the categories, and the DVPs are more discriminative than the DVWs. The DVWs perform better than the classic visual words, from which they are selected. This shows the validity of our VisualWordRank. From the figure, it can be concluded that the combination of DVW and DVP shows the best performance and achieves improvement over the baseline by 80% in average. Especially for the category: *Panda*, *Scissors*, *Windsor-Chair* and *Wrench*, recognition accuracies over 90% are achieved. The good performance comes from two aspects: 1) our training set is representative of these objects, thus meaningful DVWs and DVPs can be obtained and 2) these objects present relatively constant appearances and spatial configurations, thus they can be effectively described by the DVPs. The bad performances for the two categories: *Grand-piano* and *Headphone*, show the weakness of our selected training dataset for these two objects. This is because the 250 training images are hard to cover all of the possible appearances of some objects (e.g., *Grand-piano* and *Headphone*). This issue will be discussed in detail in the next part. From this experiment, the discriminative ability of the selected DVWs and DVPs can be clearly illustrated. It also can be concluded that our algorithm is effective in improving the discriminative power of the original visual words, from which the DVW and DVPs are selected.

E. Discussions About Limitations and Solutions

In addition to the advantages, here we shall discuss the limitations of our schemes, as well as provide feasible directions for solutions in our future work.

The first limitation is the incompactness of the DVPs. From our experiments, millions of DVPs are needed. This limitation is mainly due to the quantization error introduced during the visual word generation. With the quantization error, local features should be matched in the feature space may fail to match, and this error can be accumulated in the visual word combination, i.e., DVPs with similar semantics may fail to match each other, and huge amount of DVPs are needed to capture certain semantics. To overcome this defect, two strategies might be effective: 1) pattern summarization can be utilized to summarize DVPs sharing similar semantics together to generate high-level visual phrase vocabulary and 2) spatial-appearance preserving visual vocabulary can be generated by treating local features combinations, rather than visual word combinations. Meaningful local feature pairs can be detected and quantized into visual vocabulary. Because rich spatial and appearance cues are included

in these pairs, the corresponding generated visual vocabulary could be more informative. In addition, the parameter P_d in DVP candidate generation, i.e., (1), plays an important role in capturing meaningful visual word pairs. The correlations of a salient point to the other points may depend on both its scale and the specific properties of the object in the category. For example, the larger objects may need larger P_d values to capture their spatial configurations than the ones for smaller objects. Therefore, a single P_d may not work well for all categories and some category-wise optimization may be beneficial.

The second limitation is that DVWs and DVPs are generated based on the classic visual vocabulary, which is generated in unsupervised way. This is not ideal since the lassic visual vocabulary largely ignores the semantic contexts exist between local features i.e., local features with similar semantics may be far from each other in the feature space, while the ones with different semantics may be near to each other. This defect limits the performance of classic visual vocabulary and the corresponding DVWs and DVPs. Thus, more semantic contexts should be introduced in the visual vocabulary generation process to make the generated DVWs and DVPs semantically more meaningful.

The third issue should be discussed is the influence of the training set. Since the proposed framework is data-driven, the completeness and diversity of training data would influence the descriptive power and generalization ability of the corresponding DVWs and DVPs. For instance, if all of the images in a category are near-duplicated images (i.e., low diversity), then the extracted DVWs and DVPs would be focused on a certain appearance of the object, which would largely decrease their descriptive ability for this object. In addition, if the number of images in a category is not enough to show the common visual patterns (i.e., low completeness), valid DVPs and DVWs will cannot be identified. This is why we spend a great deal of time carefully selecting our training set. In order to utilize the publically available large-scale image dataset such as ImageNet [4] and LabelMe [28], it would be necessary to study the strategy to automatically evaluate the quality of each image category, i.e., the completeness and the diversity, and then decide the number of DVWs and DVPs should be selected.

VI. Conclusion

In this paper, we propose the DVW and DVP, which are designed to be the visual correspondences to text words. A novel framework is proposed to generate DVWs and DVPs for various applications utilizing a representative training set collected from web images. Comprehensive tests on large-scale near-duplicated image retrieval, image search re-ranking, and object recognition show that our selected DVWs and DVPs are more informative and descriptive than the classic visual words.

Future work will be carried out focusing on the following three aspects: 1) multimillion-scale training database will be utilized; 2) more effective visual vocabularies (e.g., the ones in [14], [15], [22], and [26]) will be tested for DVW and DVP generation; and 3) the incompactness of the DVPs will be further studied.

References

[1] S. Battiato, G. Farinella, G. Gallo, and D. Ravi, "Spatial hierarchy of textons distribution for scene classification.," in *Proc. Eurocom Multimedia Modeling*, 2009, pp. 333–342.

[2] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *WWW*, pp. 107–117, 1998.

[3] O. Chum, M. Perdoch, and J. Matas, "Geometric min-hashing: Finding a (thick) needle in a haystack," in *Proc. CVPR*, 2009, pp. 17–24.

[4] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," *Proc. CVPR*, pp. 248–255, 2009.

[5] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT, 1998.

[6] B. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 16, no. 315, pp. 972–976, Jan. 2007.

[7] J. Gemert, C. Veenman, A. Smeulders, and J. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[8] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. Conf. Very Large Data Bases*, 1999, pp. 518–529.

[9] W. Hsu, L. Kennedy, and S. Chang, "Video search reranking through random walk over document-level context graph," *ACM Multimedia*, pp. 971–980, 2007.

[10] H. Jegou, H. Harzallah, and C. Schmid, "A contextual dissimilarity measure for accurate and efficient image search," in *Proc. CVPR*, 2007, pp. 1–8.

[11] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. ECCV*, 2008, pp. 304–317.

[12] H. Jegou, M. Douze, C. Schmid, and P. Petrez, "Aggregating local descriptors into s compact image representation," in *Proc. CVPR*, 2010, pp. 3304–3311.

[13] Y. Jing and S. Baluja, "VisualRank: Applying PageRank to large-scale image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1877–1890, Nov. 2008.

[14] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. ICCV*, 2005, pp. 17–21.

[15] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebook by information loss minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1294–1309, Jul. 2009.

[16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006, pp. 2169–2178.

[17] D. Liu, G. Hua, P. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *Proc. CVPR*, 2008, pp. 1–8.

[18] J. Liu, W. Lai, X. Hua, Y. Huang, and S. Li, "Video search re-ranking via multi-graph propagation," *ACM Multimedia*, pp. 208–217, 2007.

[19] J. Liu, Y. Yang, and M. Shah, "Learning semantic visual vocabularies using diffusion distance," in *Proc. CVPR*, 2009, pp. 461–468.

[20] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[21] M. Marszalek and C. Schmid, "Spatial weighting for bag-of-features," in *Proc. CVPR*, 2006, pp. 2118–2125.

[22] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1632–1646, Sep. 2008.

[23] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. CVPR*, 2006, pp. 2161–2168.

[24] M. Perdoch, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *Proc. CVPR*, 2009, pp. 9–16.

[25] F. Perronnin and C. Dance, "Fisher kernels on visual vocabulary for image categorization," in *Proc. CVPR*, 2007, pp. 1–8.

[26] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1243–1256, Jul. 2008.

[27] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. CVPR*, 2008, pp. 1–8.

[28] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 157–173, May 2008.

[29] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *Proc. CVPR*, 2006, pp. 2033–2040.

[30] Z. Si, H. Gong, Y. Wu, and S. Zhu, "Learning mixed templates for object recognition," in *Proc. CVPR*, 2009, pp. 272–279.

[31] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, 2003, pp. 1470–1477.

[32] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X. Hua, "Bayesian video search reranking," *ACM Multimedia*, pp. 131–140, 2008.

[33] A. Torralba, R. Fergus, and W. Freeman, "80 Million tiny images: A large dataset for non-parametric object and scene recognition," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.
[34] P. Viola and M. Jones, "Robust real-time face detection," in *Proc. ICCV*, 2001, pp. 747–754.
[35] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *Proc. CVPR*, 2009, pp. 1903–1910.
[36] F. Wang, Y. Jiang, and C. Ngo, "Video event detection using motion relativity and visual relatedness," *ACM Multimedia*, pp. 239–248, 2008.
[37] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learning universal visual dictionary," in *Proc. ICCV*, 2005, pp. 17–21.
[38] L. Wu, S. Hoi, and N. Yu, "Semantic-preserving bag-of-words models for efficient image annotation," in *Proc. ACM Workshop on LSMRM*, 2009, pp. 19–26.
[39] Z. Wu, Q. Ke, and J. Sun, "Bundling features for large-scale partial-duplicate web image search," in *Proc. CVPR*, 2009, pp. 25–32.
[40] D. Xu and S. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1985–1997, 2008.
[41] L. Yang, P. Meer, and D. Foran, "Multiple class segmentation using a unified framework over mean-shift patches," in *Proc. CVPR*, 2007, pp. 1–8.
[42] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: From visual words to visual phrases," in *Proc. CVPR*, 2007, pp. 1–8.
[43] S. Zhang, Q. Huang, Y. Lu, and Q. Tian, "Building pair-wise visual word tree for efficient image re-ranking," in *Proc. ICASSP*, 2010, pp. 794–797.
[44] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," *ACM Multimedia*, pp. 75–84, 2009.
[45] Y. Zheng, M. Zhao, S. Y. Neo, T. S. Chua, and Q. Tian, "Visual synset: A higher-level visual representation," in *Proc. CVPR*, 2008, pp. 1–8.
[46] X. Zhou, X. Zhuang, S. Yan, S. Chang, M. Johnson, and T. Huang, "SIFT-bag kernel for video event analysis," *ACM Multimedia*, pp. 229–238, 2008.

Shiliang Zhang is currently working toward the Ph.D. degree at the Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

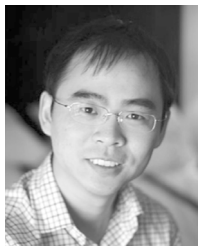
He was with Microsoft Research Asia, Beijing, China, as a Research Intern from 2008 to 2009. He returned to Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2009, and currently is a Graduate Research Assistant. His research interests include large-scale image and video retrieval, image/video processing, and multimedia content affective analysis.

Mr. Zhang was the recipient of the ACM Multimedia Student Travel Grants and the Microsoft Research Asia Fellowship in 2010.

Qi Tian (SM'04) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, in 2002.

He is currently an Associate Professor with the Department of Computer Science at the University of Texas at San Antonio (UTSA), San Antonio. His research interests include multimedia information retrieval and computer vision. He has authored or coauthored over 100 refereed journal and conference papers. He has served in various capacities for over 120 IEEE and ACM conferences. He has been a guest co-editor of the *Journal of Computer Vision and Image Understanding*, *ACM Transactions on Intelligent Systems and Technology,* and *EURASIP Journal on Advances in Signal Processing*. He is a member of the editorial board of the *Journal of Multimedia.*

Prof. Tian is a member of the Association for Computing Machinery. He is an associate editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and has served as a guest co-editor for the IEEE TRANSACTIONS ON MULTIMEDIA.



Gang Hua (M'03) received the B.S. degree in automatic control engineering and M.S. degree in pattern recognition and intelligence system from Xi'an Jiaotong University (XJTU), Xi'an, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical and computer engineering from Northwestern University, Evanston, IL, in 2006.

He is a Research Staff Member with the IBM Research T. J. Watson Center, Yorktown Heights, NY. Prior to that, he was a Senior Researcher with Nokia Research Center, Hollywood, CA, from 2009 to 2010, and a Scientist with Microsoft Live Labs Research from 2006 to 2009. He was enrolled in the Special Class for the Gifted Young of XJTU in 1994. He holds two U.S. patents and has 17 patents pending.

Dr. Hua is a member of the Association for Computing Machinery. He was the recipient of the Richter Fellowship and the Walter P. Murphy Fellowship from Northwestern University in 2005 and 2002, respectively.

Qingming Huang (M'04–SM'08) received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1994.

He was a Postdoctoral Fellow with the National University of Singapore from 1995 to 1996 and was with the Institute for Infocomm Research, Singapore, as a Member of Research Staff from 1996 to 2002. He joined the Chinese Academy of Sciences, Beijing, China, under Science100 Talent Plan in 2003, and is currently a Professor with the Graduate University, Chinese Academy of Sciences. His current research areas are image and video analysis, video coding, pattern recognition, and computer vision.

Wen Gao (M'92–SM'05–F'08) received the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1985 and 1988, respectively, and the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991.

He was a Research Fellow with the Institute of Medical Electronics Engineering, University of Tokyo, Tokyo, Japan, in 1992, and a Visiting Professor with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, in 1993. From 1994 to 1995, he was a Visiting Professor with the AI Lab, Massachusetts Institute of Technology, Cambridge. Currently, he is a Professor with the School of Electronic Engineering and Computer Science, Peking University, Peking, China, and a Professor of computer science with the Harbin Institute of Technology. He is also the Honor Professor in computer science with the City University of Hong Kong and the External Fellow with the International Computer Science Institute, University of California, Berkeley. His research interests are signal processing, image and video communication, computer vision, and artificial intelligence.