Descriptive Visual Words and Visual Phrases for Image Applications

Shiliang Zhang[1], Qi Tian[2], Gang Hua[3], Qingming Huang[4], Shipeng Li[2]

[1]Key Lab of Intelli. Info. Process., Inst. of Comput. Tech., CAS, Beijing 100080, China

[2]Microsoft Research Asia, Beijing 100080, China

[3]Microsoft Live Labs Research, Redmond, WA 78052, U.S.A.

[4] Graduate University of Chinese Academy of Sciences, Beijing 100049, China

{slzhang, qmhuang}@jdl.ac.cn, {qitian, ganghua, spli}@microsoft.com

ABSTRACT

The Bag-of-visual Words (BoW) image 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 words in texts. However, massive experiments show that the commonly used visual words are not as expressive as the text words, which is not desirable because it hinders their effectiveness in various applications. 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, novel 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 from classic visual words for various applications. In a large-scale image database containing 1506 object and scene categories, the visual words and visual word pairs descriptive to certain scenes or objects are identified as the DVWs and DVPs. Experiments show that the DVWs and DVPs are compact and descriptive, thus are more comparable with the text words than the classic visual words. We apply the identified DVWs and DVPs in several applications including image retrieval, image re-ranking, and object recognition. The DVW and DVP combination outperforms the classic visual words by 19.5% and 80% in image retrieval and object recognition tasks, respectively. The DVW and DVP based image re-ranking algorithm: DWPRank outperforms the state-of-the-art VisualRank by 12.4% in accuracy and about 11 times faster in efficiency.

Categories and Subject Descriptors

I.2.10 [Vision and Scene Understanding]: VISION

General Terms

Algorithms, Experimentation, Performance

Keywords

Bag-of-visual Words, Object Recognition, Image Re-ranking

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MM'09, October 19–24, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-608-3/09/10...$10.00.

1. INTRODUCTION

Bag-of-visual Words (BoW) image representation has been utilized for many multimedia and vision problems, including video event detection [27, 30, 34], object recognition [11, 12, 15, 18, 20, 21, 26], image segmentation [28, 31], and large-scale image retrieval [17, 22, 29], *etc*. Representing an image as a visual document composed of repeatable and distinctive basic visual elements that are indexable is very desirable. With such a representation, lots of mature techniques in information retrieval can be leveraged for vision tasks, such as visual search or recognition. Recently, it has been demonstrated, BoW image representation is one of the most promising approaches for retrieval tasks in large-scale image and video databases [17, 22].

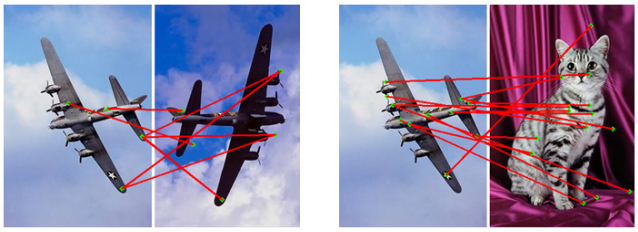


Figure 1. Matched visual words between the same and different objects

However, experimental results of reported work show that the commonly generated visual words [15, 17, 22, 28] are still not as expressive as the text words. Traditionally, a visual vocabulary is trained 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 [12, 15, 17, 18, 20, 22, 27, 28, 30, 31, 34], various numbers of visual words are generated for different tasks. There are two general observations: 1) using more visual words results in better performance [12, 17, 20]. 2) The performance will be saturated when the number of visual words reaches certain levels [12, 17, 20]. Intuitively, larger number of visual words indicates more fine-grained partitioning of the descriptor space. Hence the visual words become more descriptive 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 tasks. This strongly implies the limited descriptive ability of a single visual word. A toy example illustrating this observation is presented in Fig.1. In Fig. 1, SIFT descriptors are extracted on interest points detected by Difference of Gaussian (DoG) [14]. The three images are then represented as BoW with a visual vocabulary containing 32357 visual words, by replacing their SIFT descriptors with the indexes of the closest visual words. In the figure, two interest points are connected with red lines 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.

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 information, thus is not effective in presenting the characteristics of objects and scenes. This can be explained by an analogy between basic English alphabets and commonly used 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. 2) Previous K-means based visual vocabulary generation can not lead to very effective and compact visual word set [17, 22, 29]. This is because simply clustering the local descriptors generates lots of unnecessary and non-descriptive visual words in the cluttered background.

Aiming at the first problem, lots of works are conducted to combine multiple visual words with spatial information [1, 12, 15, 20, 27, 29, 31-33]. In general, this is achieved by identifying visual word combinations sharing stable spatial relationships. *E.g.*, in [12] the authors select the most discriminative visual word combinations with Adaboost [25] for object recognition; visual word correlogram and correlaton are utilized for object recognition in [20]; the spatial distribution of texton is modeled in [1] for scene classification. In a recent work [29], visual words are bundled for large-scale near-duplicate image retrieval. Proposed as grouped visual phrases in [33], Visual Synset presents better discrimination and invariance power than the traditional BoW representation in object categorization. As for the second problem, lots of works have proposed novel feature quantization algorithms [9, 10, 16, 19], targeting for more effective and discriminative visual vocabularies. *E.g.*, in [9] the shortcomings of K-means are analyzed, and a new acceptance-radius based clustering method is proposed to generate better visual codebooks. Another interesting work is reported by Lazebnik, *et al.* [10]. Using the results of K-means as initializations, the authors generate discriminative vocabularies according to the Information Loss Minimization theory. In [16], Extremely Randomized Clustering Tree is proposed for visual vocabulary generation, which shows promising performance in image classification. Although lots of approaches have been proposed for effective visual vocabularies and show impressive performances in many vision tasks, most of them are expensive to compute and are designed for small-scale applications. Moreover, most of the generated vocabularies are specific problem oriented (*i.e.*, for image classification, object recognition, *etc.*), thus they are still not comparable with the text words, which could be used as effective features and perform impressively in various information retrieval applications.

In order to overcome the above two problems and generate visual element set comparable to the text words, 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, once established, DVWs and DVPs will lead to compact and effective BoW 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, visual elements and their combinations that are descriptive to certain objects or scenes could be selected as DVWs and DVPs, respectively. The DVWs and DVPs composed of these elements and combinations will function more similar to the text words than the classic visual words because: 1) they are compact to describe specific objects or scenes. 2) Only unique and effective visual elements and combinations are selected. This significantly reduces the negative effects of visual features from the background clutter. Therefore, the DVWs and DVPs would be more descriptive. 3) Based on the large-scale image training set containing various 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 objects and scenes.

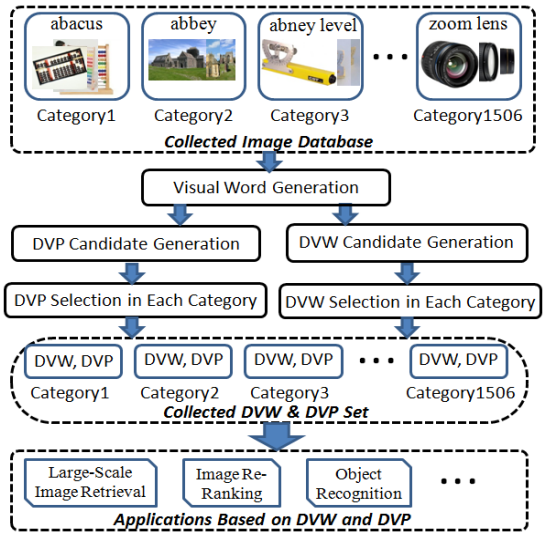


Figure 2. The proposed framework for DVW and DVP generation and application

To gather reliable statistics on large-scale image dataset, we collected about 376,500 images belonging to 1506 object and scene categories, by downloading and selecting images from Google Image. The details of our data collection will be presented in Section 4.1. Fig. 2 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 extracted from each category are considered as the DVW candidates for the corresponding category. DVP candidates in each category are generated by identifying the co-occurred visual word pairs within a certain spatial distance. A novel visual-word-level ranking algorithm: VisualWordRank which is similar to that of PageRank [2] and VisualRank [8] is proposed for identifying and selecting DVWs. 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 the selected candidates across different categories. Massive experiments on image retrieval tasks show that the DVW and DVP set presents stronger descriptive power than the classic visual words. Furthermore, in image re-ranking and object recognition, DVWs and DVPs show promising performances.

In summary, the contributions of our work are:

1) The drawbacks of classic visual words are discussed. A novel large-scale web image based solution is proposed for generating DVWs and DVPs.

2) The idea of PageRank [2] and VisualRank [8] is leveraged in VisualWordRank for DVW selection. Experiments show the promising effectiveness and efficiency of VisualWordRank.
3) The proposed DVWs and DVPs show impressive results in three applications: image retrieval, object recognition, and web image re-ranking with simple non-parametric algorithms. The DVW and DVP combination outperforms classic visual words by 19.5% and 80% in image retrieval and object recognition tasks, respectively. Moreover, the proposed image re-ranking algorithm: DWPRank outperforms the recently reported VisualRank [8] by 12.4% in accuracy and about 11 times faster in efficiency.

The rest of the paper is organized as follows. DVW and DVP candidate generation will be introduced in Section 2. The DVW and DVP selection algorithms are presented in Section 3. Section 4 discusses the applications and evaluations. Finally, Section 5 concludes the paper.

2. CANDIDATE GENERATATION

In our framework, the classic visual words appearing in each image category are taken as the DVW candidates for the corresponding category. Moreover, the semantic meaningful visual word pairs are identified as DVPs. Thus, generating the visual vocabulary and representing each training image as BoW are the first steps of our framework. In this section, we introduce how we generate the visual vocabulary, and how we generate the DVW and DVP candidates.

2.1 Visual Vocabulary Generation

Similar to existing work [17, 29], we train visual word vocabulary by clustering a large number of SIFT descriptors. Since millions of descriptors are used as the training set, we adopt hierarchical K-means to conduct the clustering for its high efficiency. Though some other clustering methods such as one-step K-means, Affinity Propagation [5] or some recent visual vocabulary generation methods [9, 10, 16, 19], could also be adopted, they are in general less efficient, 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. Thus, with the hierarchical structure, finding the closest visual word for a local feature descriptor can be performed very efficiently. More details about the vocabulary tree and its applications can be found in [17]. After clustering the local feature descriptors, a vocabulary tree is generated and the leaf nodes (cluster centers) are considered as the classic visual words. By searching hierarchically in the vocabulary tree, images in each training category are represented as BoW by replacing their SIFT descriptors with the indexes of the corresponding nearest visual words [17]. During this process, the scale of each interest point is kept for the corresponding visual word to achieve scale invariance when computing the DVP candidates.

2.2 DVW Candidate Generation

For each image category, we define the DVW candidates as the contained classic visual words. In our experiment, for a vocabulary tree with 32357 visual words, corresponding numbers of DVW candidate in 1506 image categories are sorted and shown in Fig. 3. In the figure, the numbers of DVW candidates in the 1506 categories are sorted in ascending order. Thus, the candidate number of each category can be intuitively compared with the total visual word number (32357). Obviously, the DVW candidates in each category are portions of the total visual word vocabulary. It can be inferred that only parts of the entire visual vocabulary are descriptive to the corresponding objects and scenes. Thus, selecting DVWs from their candidate set would be more efficient and reasonable than from the entire visual vocabulary.

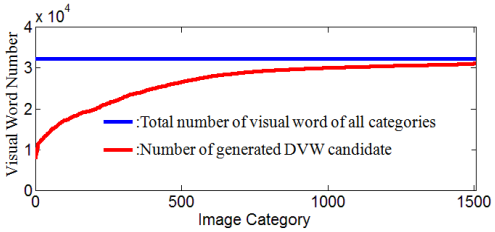


Figure 3. The number of generated DVW candidates in each image category

2.3 DVP Candidate Generation

DVPs are defined as the descriptive and commonly co-occurring visual word pairs within a constrained spatial distance in certain object or scene categories. In order to identify such visual word pairs and compute their frequency of co-occurrence, we utilize the rotation invariant spatial histogram illustrated in Fig. 4 for DVP candidate generation. Spatial histogram is commonly used for spatial relationship computation between interest points. More details can be found in [12]. In Fig. 4, each visual word pair co-occurring within the histogram is considered as a DVP candidate.

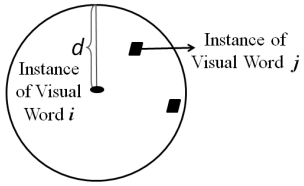


Figure 4. Spatial histogram for DVP candidate generation

As shown in Fig.4, the spatial histogram centered at an instance of visual word i is defined as:

$$SH_i[t_j, d]$$

where t_j is the number of instances of visual word j that fall in the histogram. Thus t_j would be 2 in the example illustrated in Fig. 4.

Note that the radius d is an important parameter related to the constraint of co-occurrence. Because objects may have various scales in different images, the d should be properly computed to achieve scale invariance. Based on the scale information [14] of the detected interest points, from which the SIFT descriptors are extracted, we compute d with Eq. (1).

$$d = Scale_i \bullet P_d \quad (1)$$

where, $Scale_i$ is the scale of the interest point [14] from which the instance of visual word i is computed, and P_d is a parameter controlling the constraint of co-occurrence. From our experiments, larger P_d is necessary for identifying reliable spatial relationships between two visual words and overcoming the sparseness of the generated DVW candidates. However, large P_d will also increase the computational cost and the occurrence of noise. In this paper, we experimentally set P_d as 4, a good trade-off between efficiency and performance. Intuitively, if t_j presents large values, strong co-occurrence can be indicated between the visual word i and j.

Suppose visual word i and j co-occur within the spatial histogram in an image category. Then, the DVP candidate containing the two visual words for this category can be defined as:

$$DVPCandidate^{(C)}[i, j, T_{i,j}^{(C)}]$$

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 spatial histogram, $T_{i,j}^{(C)}$ will present a large value. Hence, $T_{i,j}^{(C)}$ reflects the strength of their spatial relationship in category C. Algorithm 1 presents the detailed computation of $T_{i,j}^{(C)}$ in image category C.

Algorithm1: compute the co-occurrence frequency $T_{i,j}^{(C)}$

Input: Instances of visual word i and j in the P images contained in image category C.

Output: $T_{i,j}^{(C)}$

Suppose: in image p, the number of instances of visual word i is N_p, and the total number of visual word instances is M_p.

$T_{i,j}^{(C)} = 0$

For image p, $p = 1, \ldots, P$

Initialize N_p spatial histograms $SH_i^{(k)}[t_j^{(k)}, d^{(k)}]$, $k = 1, \ldots, N_p$, use each instance of visual word i as the histogram reference center.

For $k = 1, \ldots, N_p$ **do**

Compute $d^{(k)}$ with Eq. (1).

Suppose n_k instances of visual word j fall into $SH_i^{(k)}[t_j^{(k)}, d^{(k)}]$

Then, $t_j^{(k)} = n_k$

End

$$T_{i,j}^{(p)} = \sum_{k=1}^{N_p} t_j^{(k)} / M_p \quad and \quad T_{i,j}^{(C)} = T_{i,j}^{(C)} + T_{i,j}^{(p)}$$

End

$$T_{i,j}^{(C)} = T_{i,j}^{(C)} / P$$

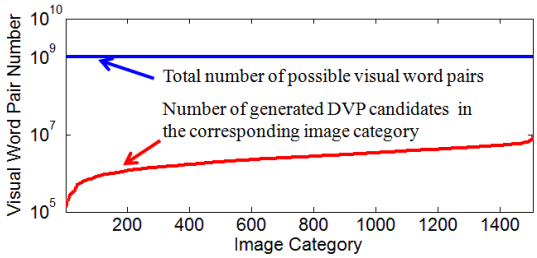


Figure 5. The number of generated DVP candidates in each image category

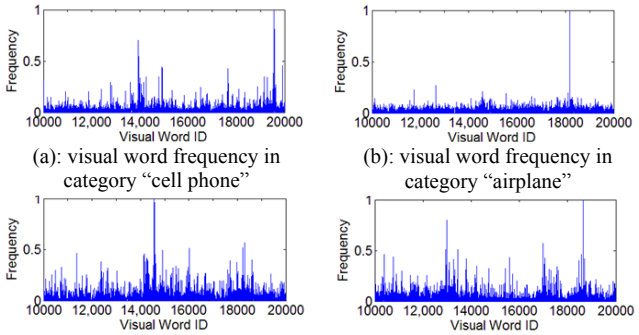
Because each visual word pair is considered as a possible DVP candidate, the result of DVP candidate generation in each image category can be intuitively seen as a $VWnum \times VWnum$ sized matrix ($VWnum$ is the size of the visual vocabulary). Each non-empty element in it is a DVP candidate carrying the frequency of co-occurrence information between the corresponding two visual words. The numbers of generated DVP candidates in each image category are sorted and presented in Fig. 5. It can be seen that, although the generated candidates are only small portions of the entire possible visual word pairs, their sizes are still very huge. Therefore, effective and compact DVP set is needed to be selected from the candidates.

3. DVW AND DVP SELECTION

3.1 Descriptive Visual Word Selection

DVWs are designed to capture certain objects or scenes, 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] and VisualRank [8], we design a novel visual-word-level ranking algorithm: VisualWordRank to effectively incorporate the two criteria for DVW selection.

According to the first criterion, the frequency-of-occurrence information of DVW candidates in the total image set and in each individual image category would be important for identifying DVWs. Fig. 6 (a-d) show the frequencies of occurrence of visual words with index number: *e.g.*, $1 \times 10^4 \sim 2 \times 10^4$ in four image categories. The frequencies shown are normalized between 0 and 1 by the maximum and minimum frequencies of the selected visual words. It is clear that, the same visual words (*e.g.,* visual words with ID 14000-16000) present different frequencies in different image categories. Thus, their different significances for each category can be indicated.



(a): visual word frequency in category "cell phone"
(b): visual word frequency in category "airplane"
(c): visual word frequency in category "ant"
(d): visual word frequency in category "bike"

Figure 6. The visual word frequencies in different categories

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 same object but different backgrounds, the spatially consistent visual words are more likely to be located on the object. Hence, the spatial co-occurrence information between visual word pairs in Algorithm 1 is adopted in DVW selection to depress the negative influences caused by the cluttered background.

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 between 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 relationships with other webpages [2].

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

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

i is a DVW candidate. 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 BoW.

The non-diagonal element $R_{i,j}^{(C)}$ is defined as the average co-occurrence frequency of visual word i and j in image category C:

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

where $T_{i,j}^{(C)}$ is from the DVP candidate computed in Algorithm 1.

After computing the $R^{(C)}$, we normalize the diagonal elements and non-diagonal elements, respectively and assign them with weights: W_{freq} and W_{cooc}. The two input weights control the influences of frequency factor and co-occurrence factor in VisualWordRank, respectively. From extensive experiments, we conclude that setting the two weights equal value results in best performance for most of the image categories. The detailed computation of $R^{(C)}$ is summarized in Algorithm 2.

Algorithm2: Compute matrix $R^{(C)}$ for image category C

Input: W_{freq}, W_{cooc}; F_i, $f_i^{(C)}$, $i = 1,...,VWnum^{(C)}$; DVP candidates in category C.

Output: The matrix $R^{(C)}$

For i and j=1,…, $VWnum^{(C)}$ **do**

Assign the value of $R_{i,j}^{(C)}$ based on Eq. (2) and Eq. (3)

Get the sum of diagonal elements: Sum_{diag}

Get the sum of non-diagonal elements: $Sum_{non-diag}$

End

For i, and j=1,…, $VWnum^{(C)}$ **do**

If ($i \neq j$) $R_{i,j}^{(C)} = W_{cooc} \bullet R_{i,j}^{(C)} / Sum_{non-diag}$

If ($i == j$) $R_{i,j}^{(C)} = W_{freq} \bullet R_{i,j}^{(C)} / Sum_{diag}$

End

Algorithm3: VisualWordRank

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

Output: The rank of each DVW candidate to the category C: $Rank_i^{(C)}$, $i = 1,...,VWnum^{(C)}$

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 $iter < maxiter$

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

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

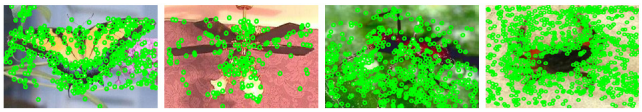
$OldRank^{(C)} = NewRank^{(C)}$

$iter$++

End

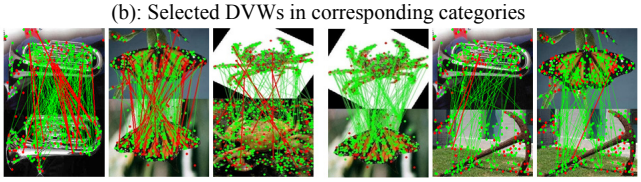
$Rank^{(C)} = NewRank^{(C)}$

With the matrix $R^{(C)}$, we set the initial rank value of each DVP candidate equal and then start the rank-updating iterations. The detailed descriptions of VisualWordRank are presented in Algorithm 3. 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 DVW set in object category C can be generated by selecting the top N ranked DVW candidates or choosing the ones with rank values larger than a threshold.



(a): DVW candidates before VisualWordRank

(b): Selected DVWs in corresponding categories



(c): Matched DVWs and visual words between same and different objects

Figure 7. (a) DVW candidates, (b) DVWs, and (c) matched DVWs (red lines) and matched visual words (green lines)

Fig. 7 (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. 7 (b). Obviously, although there are many candidates on the cluttered background, most of the selected DVWs appear on the object. In order to show the descriptiveness of the selected DVW set, the matched visual words and DVWs between same and different objects are compared in Fig. 7 (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 visual words between different objects, lots of 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 will be presented in Section 4.

3.2 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 (*i.e.*, spatial relationship) information of visual word pair has already been integrated in the generated DVP candidates, 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 an 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:

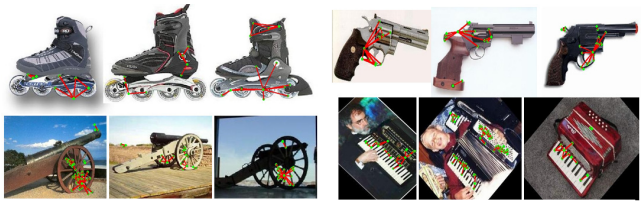
$$VPI_k^{(C)} = VPf_k^{(C)} / \ln(VPF_k) \quad (4)$$

where, $VPI_k^{(C)}$ is the importance of the DVP candidate k to the category C, $VPf_k^{(C)}$ and VPF_k stand for the frequencies of

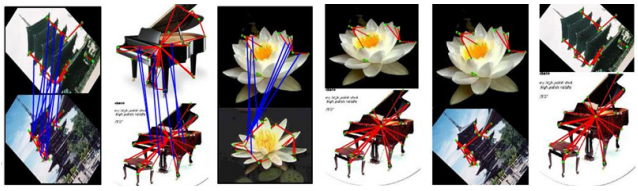
occurrence of candidate k in category C and all categories, respectively. Suppose there are M image categories and the two visual words contained in k are visual word i and visual word j, respectively, then $VPf_k^{(C)}$ and VPF_k can be computed with Eq. (5).

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

Consequently, after computing the importance of each DVP candidate, the DVPs for category C could be identified and selected from the top ranked $VPI^{(C)}$.



(a): Selected DVPs in: "inline skate", "revolver", "cannon", and "accordion"



(b): Matched DVPs between the same and different objects

Figure 8. The selected DVPs and the matched DVPs between the same and different objects

In Fig. 8 (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. 8 (b) shows the matched DVPs across same and different objects. All 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 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. It also can be inferred that DVPs are more effective in describing objects than single visual words. The performance of DVPs will be further evaluated in the Section 4.

4. APPLICATIONS AND EVALUATIONS

4.1 Image Dataset Collection

The DVW and DVP generation is based on the statistics of their candidates in different image categories. Moreover, the DVW and DVP set is desired to be semantic meaningful and descriptive for various objects and scenes. Thus, representative image database with enough object and scene categories is an important basis for DVW and DVP set. Decades ago, it was very hard to collect such large-scale labeled image database because of the limited data source and hardware ability. However, the boom of web image search engines and the explosively increasing images on internet have already made it feasible to collect and store large-scale image databases. Representative start-of-the-art work of knowledge mining from web-scale images can be found in [3, 24]. In [24], the authors propose simple but robust methods for challenging tasks such as person detection, scene recognition, object classification, *etc.,* based on the large-scale loosely-labeled web images. Similarly, large-scale labeled web image database: ImageNet is collected and released in [3]. Therefore, collecting meaningful training dataset from Internet has been proved feasible.

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 [3, 24]. We first use WordNet V.2.1 [4] to get a comprehensive list of objects and scenes by extracting 117097 non-abstract 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 filtered. Then, from the remaining images, we carefully selected 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. To the best of our knowledge, our collected dataset is one of the most representative large-scale image training sets in literature. Thus, extracting and selecting DVWs and DVPs based on it would be statically meaningful.

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 will result in huge number of possible visual word pairs, and low repeatability of each DVP candidate. 2) Single visual word shows limited descriptive ability, no matter how fine-scaled it is [12, 17, 20]. 3) The training images are evenly selected from the representative database to obtain better description of the feature space as much as possible. Based on the generated visual words, the entire image dataset (376,500) is then used for candidate generation and final DVW and DVP selection.

4.2 Image Retrieval based on DVW and DVP

In recent work, BoW image representation has been proven promising in large-scale image retrieval [17]. Thus, experiments are carried out to compare classic visual words with the proposed DVWs and DVPs on image retrieval tasks. We choose Corel 5000 as the testset because it is a widely used benchmark dataset in CBIR community. In this dataset, 50 image categories are included and each contains 100 images. Each image in the database is first represented as BoW, then, indexed with the contained visual words using inverted file structure. Similarly, the images are also indexed with the DVWs and DVPs within them, respectively. In the retrieval process, TF-IDF weighting is applied to compute the similarities between images. The retrieval precision of the first k returned images is computed with Eq. (6):

$$Precision_k = Correct_k / k \tag{6}$$

where, $Correct_k$ is the number of relevant images among the first k returned images.

To make the performance comparisons between classic visual words and DVWs and DVPs more visible, we use *PrecisionRatio* computed with Eq. (7) as a measurement.

$$PrecisionRatio_k = Precision_k^{(a)} \Big/ Precision_k^{(b)} \tag{7}$$

where, $Precision_k^{(a)}$ and $Precision_k^{(b)}$ are the precision values based on two different image features (*e.g.* classic visual word, DVW, or DVP) in the first k returned images, respectively. Thus, if $PrecisionRatio_k$=1, these two image features show the same performance. The classic visual word [17] is used as $Precision_k^{(b)}$.

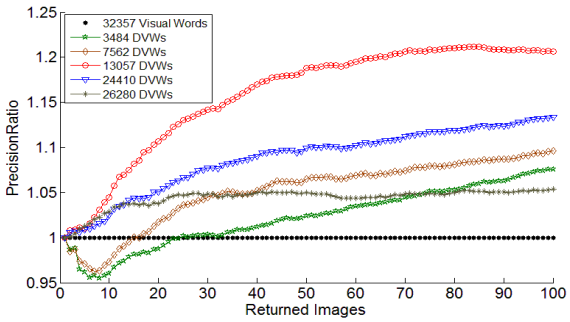


Figure 9. Performance comparisons between DVWs and classic visual words

Fig. 9 demonstrates the performance comparisons between classic visual words and DVWs. All the 32357 classic visual words are utilized. While, 3484, 7562, 13057, 24410, and 26280 DVWs are collected from the training image categories. All the 5000 images in the testset are indexed and used for retrieval. The ratio curves in the figure are computed based on the overall average precisions of the 5000 queries. From Fig. 9, it can be seen that DVW set with the size 13057 shows obvious 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. 9 that, DVW sets with the size 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), the correctly matched visual words between them and the query image are more likely to be disturbed by the negative effects of background clutter. Because the DVW set with small size keeps the most descriptive visual words and has filtered most of the noisy ones, the interferences of such background noise are depressed. As a result, the correct matches would be relatively stronger than the noisy ones. Consequently, DVWs perform better than the classic visual words in the case when more noises exist. Since DVWs are selected from the 32357 classic visual words, DVW sets with larger size will contain more noises and thus 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 26280), their performance will start to decrease. Therefore, we could conclude that DVWs with an appropriate size are more compact, descriptive and robust than classic visual words.

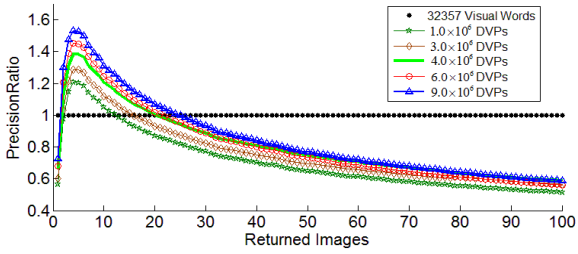


Figure 10. Performance comparisons between DVPs and classic visual words

To evaluate the performance of the DVPs, we adopt the classic visual words as the baseline. DVP sets with different sizes are collected based on different thresholds. Images are indexed and retrieved based on the DVPs they contain just like the previous experiment. The selected DVP numbers and the corresponding experimental results are presented in Fig. 10. 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 phrase space. Since DVP candidates contain both spatial and appearance information, 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. Since DVP set with the size 9.0×10^6 is still a very small portion of the possible visual word pairs (*i.e.*, 32357^2), we could conclude that the selected DVP set is compact. From the figure, 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 query images in categories such as "Beach" and "Wave" do not present consistent spatial characteristics and contain very few or even zero DVPs. Thus DVPs do not work well for these cases. Because each object commonly presents several different typical appearances (*e.g.,* photos of "car" taken from different viewpoints present different spatial characteristics), a DVP from one category may be only descriptive to a certain appearance of the corresponding object. Therefore, the DVPs can only effectively recognize the near-duplicate images showing similar appearances with the query one. This is the reason why DVPs show obvious advantages in the first several returned images but perform worse when the returned images exceed certain numbers. On the other hand, it can be observed that DVPs and DVWs show very distinct performances and can be complemented to each other. Thus, the performance of DVW+DVP is further evaluated.

To test the overall performance of the obtained DVWs and DVPs, we compare different combinations of them with classic visual words. The compared combinations corresponding experimental results are presented in Fig. 11.

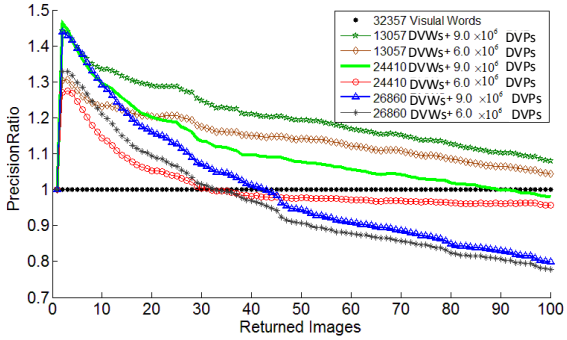


Figure 11. Performance comparisons between classic visual words and the combinations of DVW and DVP

From Fig.11, it can be observed that medium number of DVWs plus a large number of DVPs show the best performance. The combination containing 13057 DVWs and 9×10^6 DVPs shows the best performance in the figure and it outperforms the classic visual words by 19.5% in term of MAP (Mean Average Precision) computed in the top 100 returned images. Accordingly, we can come to the conclusion that, our proposed DVWs and DVPs are more effective for BoW image representation and more suitable for image retrieval than the widely used classic visual words. In the next sections, DVWs and DVPs are further evaluated on object recognition and image re-ranking applications.

4.3 Object Recognition

Object recognition has been a popular research topic for many years. Based on the well-designed features and classifiers, lots of recently reported works show promising performance in

challenging recognition tasks [11, 12, 20, 21, 26, 28]. Since DVWs and DVPs are designed to effectively describe certain objects or scenes. It is straightforward that the selected DVWs and DVPs in each image category should be discriminative for the corresponding object. Consequently, we utilize the object recognition task to illustrate the discriminative ability of DVWs and DVPs. Besides that, this experiment is also carried out to test the effectiveness of our algorithm in improving the discriminative power of original visual words, form which DVWs and DVPs are generated. Thus, classic visual word is utilized as the baseline feature. From Caltech101 and Caltech256 dataset, we select 15 commonly used object categories as the testset. For each test category, the training image 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 1. Note that each training category contains 250 images, some of which are irrelative ones.

Table 1. The query words of selected training categories and corresponding test categories

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

In the experiment, we first identify and collect 150 DVWs and 6000 DVPs from each training category with the algorithms introduced in previous sections. Then, for each object, we establish three discriminative feature pools containing DVWs, DVPs and both of them, respectively. In the testing phase, all the DVW and DVP candidates in each test image are first extracted. Then, 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". In similar way, another two recognition results based on DVP and DVW+DVP can also be obtained. In the baseline algorithm, each test image is recognized by computing its 10 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 non-parametric classifiers are used, the discriminative abilities of different features can be clearly illustrated. Experimental results are presented in Fig.12.

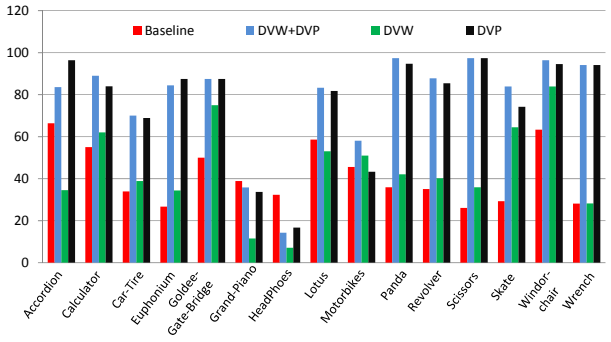


Figure 12. The comparisons of object recognition precision between DVWs, DVPs and classic visual words (baseline)

Obviously from Fig.12, the DVWs and DVPs outperform the baseline feature by a large margin for most of the categories. It can also be observed that the extracted DVPs are more discriminative than the DVWs. This is mainly because DVPs contain more spatial information, which significantly improves their descriptive power for different objects. The DVWs perform better than the classic visual words, from which they are selected. This shows the validity of our VisualWordRank algorithm. From the figure, it can be concluded that the combination of DVWs and DVPs presents the best performance and achieves significant improvement over the baseline by 80% in average. Thus, the discriminative ability of the selected DVWs and DVPs can be illustrated. The object recognition confusion matrix obtained by DVW+DVP is presented in Fig. 13.

	Accordion	Calculator	Car-tire	Euphonium	Golden-Gate-Bridge	Grand-Piano	Headphone	Lotus	Motorbike	Panda	Revolver	Scissors	Inline-skate	Windsor-Chair	Wrench
Accordion	83.6	9.1	3.6	0	0	0	0	0	0	0	0	0	0	3.6	0
Calculator	1.0	89.0	2.0	0	2.0	1.0	0	0	2.0	0	0	0	0	3.0	0
Car-tire	1.1	2.2	70.0	2.2	2.2	0	1.1	5.6	2.2	0	0	6.7	3.3	3.3	1.1
Euphonium	1.6	1.6	3.1	84.4	3.1	1.6	0	0	0	0	0	0	3.1	0	1.6
Golden-Gate-Bridge	2.5	0	0	1.3	87.5	0	0	2.5	1.3	1.3	1.3	1.3	0	0	1.3
Grand-Piano	9.4	6.3	2.1	2.1	4.2	36.8	1.1	8.4	5.3	1.1	8.4	0	4.2	4.2	7.4
Headphone	4.8	4.8	12.0	0	4.8	2.4	16.7	9.5	12.0	0	9.5	14.3	0	4.8	4.8
Lotus	1.5	0	1.5	0	1.5	3.0	0	83.3	3.0	1.5	1.5	1.5	0	0	1.5
Motorbike	3.6	8.1	4.5	2.1	1.2	4.5	2.6	1.4	58.1	1.2	3.8	2.6	1.2	1.0	4.0
Panda	0	0	0	0	0	0	0	0	5.3	94.7	0	0	0	0	0
Revolver	2.4	0	0	4.8	0	0	0	1.2	0	0	87.8	0	0	2.4	1.2
Scissors	4.9	0	0	0	0	0	0	0	0	0	0	91.0	0	0	4.9
Inline-skate	0	0	6.5	0	0	0	3.2	0	0	3.2	0	3.2	83.9	3.2	0
Windsor-Chair	3.6	0	3.6	0	0	0	0	0	0	0	0	0	0	92.8	0
Wrench	2.6	2.6	0	0	0	0	0	0	0	0	0	0	0	0	94.8

Figure 13. The confusion matrix obtained with DVW + DVP

From the confusion matrix in Fig. 13, it can be observed that, DVW and DVP combination shows recognition accuracy over 80% for most of the objects, even with a simple classifier. 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 collected from Google Image is representative of these objects, thus meaningful DVWs and DVPs can be obtained from the training set. 2) The selected objects present relatively constant appearances and obvious spatial characteristics, thus they can be effectively described by the more discriminative 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 images used to generate DVWs and DVPs are hard to cover all the possible appearances of some objects (*e.g. Grand-piano* and *Headphone*). For this consideration, we are collecting more training categories from other image search engines. Moreover, more classifiers will be designed and tested based on DVWs and DVPs in our future work. It is expected, with an enlarged training dataset, and well-designed classifiers, the object recognition accuracy would be improved and comparable with the state-of-the-art [11, 12, 20, 21, 26, 28] in more challenging recognition tasks.

4.4 Image Re-ranking

Image re-ranking is a research problem catching more and more attentions in recent years [7, 8, 13, 23]. The goal is to resort the images returned by text-based search engines according to their

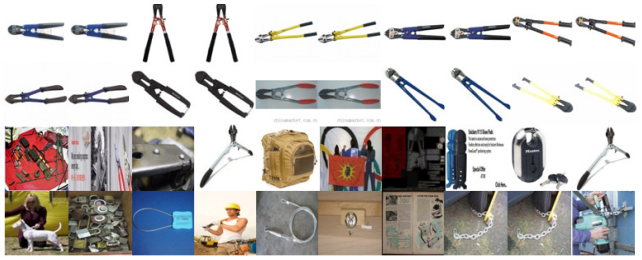
visual appearances to make the top-ranked images more relevant to the query. As a state-of-the-art work, VisualRank [8] computes the visual similarities between images and leverages the algorithm similar to PageRank [2] to re-rank the we images. Based on the DVW and DVP set, we propose a novel image re-ranking algorithm: DWPRank. Experiments and comparisons between VisualRank show that our DWPRank presents better performance in terms of both accuracy and efficiency.

The problem of image re-ranking can be seen as identifying the common visual concept (*i.e.,* scene, object, *etc.*) contained in the returned images from search engines 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, from which they are selected. Therefore, they can be utilized to measure the relevance between images and the concept. Based on this idea we proposed a novel image re-ranking algorithm: DWPRank, which is detailedly presented in Algorithm 4.

Top and last 10 returned images by Google Image with the query "**all-terrain bike**"

Top and last 20 re-ranked images by DWPRank in category "**all-terrain bike**"

Top and last 10 returned images by Google Image with the query "**bolt cutter**"



Top and last 20 re-ranked images by DWPRank in category "**bolt cutter**"

Figure 14. Examples of the re-ranked images by DWPRank

To illustrate the validity of DWPRank, we first carry out DWPRank on our collected database which contains the top 250 images returned from Google Image. Some examples of the re-ranking results are presented in Fig. 14. Obviously from Fig. 14, the relevant images are highly ranked, while the irrelevant ones from Google Image are ranked in the end of the list.

Extensive tests of DWPRank are carried out by comparing it with VisualRank on the same testset. In our experiment, an image re-ranking testset is collected by selecting 40 image categories from the image database introduced in Section 4.1. Each selected category contains 250 images and presents obvious visual concept (*i.e.*, same objects or scenes). Hence, we assume all the 250 images are relevant to the concept. After that, 100 randomly selected images are added (random mixture) to each of these categories. Finally, we construct a dataset containing 40 categories and 14000 images to compare our DWPRank with VisualRank. AP (Average Precision) computed in Eq. (8) is adopted to measure the effectiveness of the re-ranking algorithm.

$$AP = \left(\sum_{i=1}^{250} correct_i / i \right) \Big/ 250 \qquad (8)$$

where, $correct_i$ is the number of relevant images in the top i re-ranked images. Thus, if AP=1, it can be inferred that all of the irrelevant images are in the end of the re-ranked image list, which is the most ideal case in image re-ranking.

Algorithm4: DWPRank

Input: Images returned from the image search engine: $I_i, (i=1,...,N)$; weight of DVW and DVP: W_{DVW}, W_{DVP}.
Output: Re-ranked image list: $IReRanked_i$, (i=1,…,N)
Suppose: Rel_i, (i=1,…,N) describes the relevance between image I_i and the query concept.
In I_i, (i=1,..., N), generate the DVW and DVP candidates.
In I_i, (i=1,..., 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

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_{DVW}, W_{DVP} in Algorithm 4 as (1, 0), (0, 1) and (1, 1) respectively. The results of the DWPRank and VisualRank are presented in Fig. 15.

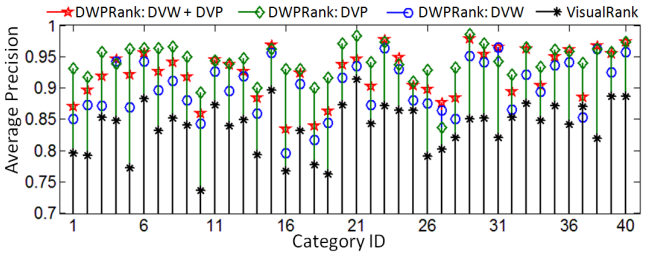
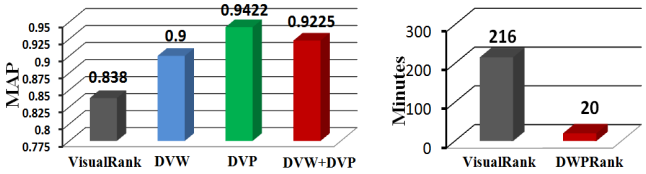


Figure 15. The comparisons between DWPRank and VisualRank

Obviously from Fig. 15, DWPRank outperforms VisualRank for most of the cases. The main reasons for the improvement are from two parts. 1) More information and constrains (*i.e.*, spatial and frequency clues) are considered in DVW and DVP selection, thus DVWs and DVPs are more effective in identifying and describing the visual concepts in returned images; 2) VisualRank computes the image-pair similarities based on all the SIFT descriptors in each image, thus the SIFT features on the background might disturb its performance during the re-ranking. Differently, such influences are mush depressed in DWPRank through DVW and

DVP selection. From Fig. 15, 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 capture more spatial information and are more descriptive. The Mean Average Precision (MAP) values obtained by DWPRank and VisualRank are presented in Figure 16 (a). From the figure, we conclude that improvements of 7.4%, 12.4%, and 10.1% over the VisualRank are achieved by DWPRank with DVWs, DVPs and their combination, respectively.



(a): The Mean Average Precision obtained by VisualRank and DWPRank

(b): Average time needed by VisualRank and DWPRank for re-ranking 350 images

Figure 16. The comparisons of accuracy and efficiency between VisualRank and DWPRank

Besides the obvious improvements on accuracy, it is necessary to point out that, DWPRank is more efficient than VisualRank. The above experiments are carried out on 12 workstations with 8 GB memory and 4-core 2.83 Ghz processor. The average time needed by VisualRank and DWPRank for re-ranking 350 images are compared in Fig. 16(b). Obviously, about 11 times of improvement is achieved by DWPRank. The low efficiency of VisualRank is mainly rooted in the expensive image pair similarity computation based on SIFT and LSH [6]. However, in DWPRank, DVP candidate generation and DVW selection, which are the most time-consuming operations, can be finished very efficiently with simple spatial histogram and efficient VisualWordRank. In short, DWPRank shows significant advantages on both accuracy and efficiency over the VisualRank.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose the DVWs and DVPs, which are designed to be the visual correspondences to text words. A novel framework is proposed for generating DVWs and DVPs for various applications based on a representative training set collected from web images. Comprehensive tests show that our selected DVWs and DVPs are compact and descriptive. Moreover, DVWs and DVPs show promising performances in tasks of image retrieval, object recognition and image re-ranking.

Future work will be carried out focusing on the following three aspects. 1) Multi-million-scale training database will be collected. 2) More effective visual vocabularies (*e.g.*, the ones in [9, 10, 16, 19]) will be tested for DVW and DVP generation; DVP candidate generation and selection algorithms will be further studied. 3) DVW and DVP based applications will be further explored.

6. REFERENCES

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

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *International World-Wide Web Conference*, pp. 107-117, 1998.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. *Proc. CVPR*, pp. 710-719, 2009.

[4] C. Fellbaum. *Wordnet: an electronic lexical database*. Bradford Books, 1998.

[5] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 16(315): 972-976, Jan. 2007.

[6] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. *Proc. VLDB*, pp. 518-529, 1999.

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

[8] Y. Jing and S. Baluja. VisualRank: applying PageRank to large-scale image search. *PAMI*, 30(11): 1877-1890, Nov. 2008.

[9] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. *Proc. ICCV*, pp. 17-21, 2005.

[10] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebook by information loss minimization. *PAMI*, 31(7): 1294-1309, July 2009.

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. *Proc. CVPR*, pp. 2169-2178, 2006.

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

[13] 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.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2): 91-110, Nov. 2004.

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

[16] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *PAMI*, 30(9): 1632-1646, Sep. 2008.

[17] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *Proc. CVPR*, pp. 2161-2168, 2006.

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

[19] F. Perronnin. Universal and adapted vocabularies for generic visual categorization. *PAMI*, 30(7): 1243-1256, July 2008.

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

[21] Z. Si, H. Gong, Y. N. Wu, and S. C. Zhu. Learning mixed templates for object recognition. *Proc. CVPR*, 2009.

[22] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. *Proc. ICCV*, pp. 1470-1477, 2003.

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

[24] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *PAMI,* 30(11): 1958-1970, Nov. 2008.

[25] P. Viola and M. Jones. Robust real-time face detection. *Proc. ICCV*, pp. 7-14, 2001.

[26] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. *Proc. CVPR*, 2009.

[27] F. Wang, Y. G. Jiang, and C. W. Ngo. Video event detection using motion relativity and visual relatedness. *Proc. ACM Multimedia*, pp. 239-248, 2008.

[28] J. Winn, A. Criminisi, and T. Minka. Object categorization by learning universal visual dictionary. *Proc. ICCV*, pp. 17-21, 2005.

[29] Z. Wu, Q. F. Ke, and J. Sun. Bundling features for large-scale partial-duplicate web image search. *Proc. CVPR*, 2009.

[30] D. Xu and S. F. Chang. Video event recognition using kernel methods with multilevel temporal alignment. *PAMI*, 30(11): 1985-1997, Nov. 2008.

[31] L. Yang, P. Meer, and D. J. Foran. Multiple class segmentation using a unified framework over mean-shift patches. *Proc. CVPR*, pp. 1-8, 2007.

[32] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. *Proc. CVPR*, pp.1-8, 2007.

[33] Y. T. Zheng, M. Zhao, S. Y. Neo, T. S. Chua, and Q. Tian. Visual synset: a higher-level visual representation. *CVPR*, pp. 1-8, 2008.

[34] X. Zhou, X. D. Zhuang, S. C. Yan, S. F. Chang, M.H. Johnson, and T.S. Huang. SIFT-bag kernel for video event analysis. *Proc. ACM Multimedia*, pp. 229-238, 2008.