Quadtree Decomposition based Extended Vector Space Model for Image Retrieval

Vignesh Ramanathan, Shaunak Mishra and Pabitra Mitra
Indian Institute of Technology, Kharagpur - 721302
vigneshram.iitkgp,shaunak.mishra.iitkgp,pabitra @gmail.com

Abstract

Bag of visual words approach for image retrieval does not exploit the spatial distribution of visual words in an image. Previous attempts to incorporate the spatial distribution include modification of visual vocabulary using visual phrases along with visual words and use of spatial pyramid matching (SPM) techniques for comparing two images. This paper proposes a novel extended vector space based image retrieval technique which takes into account the spatial occurrence (context) of a visual word in an image along with the co-occurrence of other visual words in a pre-defined region (block) of the image obtained by quadtree decomposition of the image up to a fixed level of resolution. Experiments show a 19.22% *increase in Mean Average Precision (MAP) over the BoW approach for the Caltech 101 database.*

1. Introduction

Image retrieval schemes based on Bag of Words (BoW) or the vector space model are inspired from the analogy between a text document which consists of a collection of words in a meaningful order and an image that can be represented as a collection of "visual words" obtained by the Scale Invariant Feature Transform (SIFT) [5]. Recent work on BoW based image retrieval have shown promising results for large scale video and image retrieval applications [1, 3, 6, 7, 8, 9, 13]. However, the analogy with text document, which is central to BoW model, breaks down beyond a point as textual data is discrete and semi-structured while an image is continuous and non-structured.

Scope for improvement of the BoW model still remains because of two major weaknesses of the BoW model e.g. failure to capture visual pattern specific information in frequently co-occurring visual words and lack of a framework to incorporate image specific information on spatial distribution of visual words in the image. The BoW model considers an image to be an order-less collection of local features. It considers a histogram of visual word occurrences in an image as the sole basis for comparison with other images. The BoW scheme does not take into account the co-occurrences of visual words characteristic in a visual pattern and thus ill represents an image. In practice, clustering of visual words results in over-representation of visual patterns (synonymy) as well as under representation of visual patterns due to the context specific meaning of a visual word (polysemy) [11, 12, 16]. An approach to take the co-occurrence aspect into account is to define pairs of frequently co-occurring visual words as visual phrases [11, 12, 14] analogous to text phrases as a refinement of the visual vocabulary.

Vocabulary refinement based on visual phrases does not entirely capture the image specific spatial information, leading researchers to the need for a framework to incorporate spatial distribution of visual words. Multi-resolution histogram based matching method [2] involves computing histogram of pixel values at different levels obtained by successively sub-sampling the image. Further improvement is observed in the spatial pyramid matching (SPM) technique [4] which is based on visual word occurrences over fixed sub-regions of an image. The SPM technique considers features computed at a fixed resolution but varies the spatial resolution at which the occurrences of visual words are considered. Results in [4] confirm the effectiveness of global non-invariant representations in categorizing scenes, even when images are affected by background clutter along with variation in superficial appearance of the object of interest.

The goal of the present work is to enrich the vector space model vocabulary by including co-occurrences of visual words and considering spatial distribution of visual words. We propose a novel extended vector space representation of an image inspired from extended vector space models for XML information retrieval [10]. The nature of XML documents lies between structured data and unstructured data, opening possibilities of analogous methods in image retrieval. The proposed scheme extends the vector space model used in BoW approach by jointly encapsulating co-

978-1-4244-9497-2/10/$26.00 ©2010 IEEE

occurrences of visual words and their global positions in the image captured by sub-dividing the images into blocks as discussed in the subsequent section of the paper.

The remainder of the paper is organized as follows. The proposed extended vector space model for image retrieval is introduced in Sec. 2. Experimental results are provided to demonstrate the performance of the proposed algorithm in Sec. 3. Finally, concluding remarks are given in Sec. 4.

2. Extended Vector Space Image Retrieval Approach

In this section we propose an extended vector space based image retrieval scheme. This method helps capture the information of spatial distribution of visual words which is ignored by the traditional BoW approach. Firstly, the image visual word spatial distribution is represented by a suitable structure to include the positional information. This is followed by the construction of an extended visual words vector which is consequently used for image comparison. A tf-idf comparison scheme is then used to evaluate the similarity between the query image and other images in the database.

2.1. Visual Vocabulary formation

In this paper, we use SIFT[5] features to describe the key patches in an image. The SIFT feature describes local patches in an image as a 128 dimension vector. The local patches in the image are then represented by this descriptor and their position in the image. These descriptors are further vector quantized to form visual words. The entire set of features obtained from a database is clustered into K clusters through hierarchical K-mean clustering [14]. Vector quantization of a descriptor is achieved by suitably assigning it to any of the K clusters of features.

2.2. Tree representation of Spatial Distribution of Visual Words

The spatial location of visual words in an image should be encoded in the extended vector space in order to utilize the information of their spatial distribution in the image. This can be achieved by adopting a few techniques used in XML retrieval. A XML document has a tree like structure, where every term (word) called the leaf element is associated with a set of tags called nodes which form the structure of the XML document as shown in Fig.1. Thus starting from the root node or any other non leaf node, one can associate a path with every word ending at the leaf element. A similar approach can be followed to assign structure paths to visual words in images. We describe a tree based representation of an image below.

Quadtree decomposition is used to split an image into equal sized blocks. The decomposition begins with the

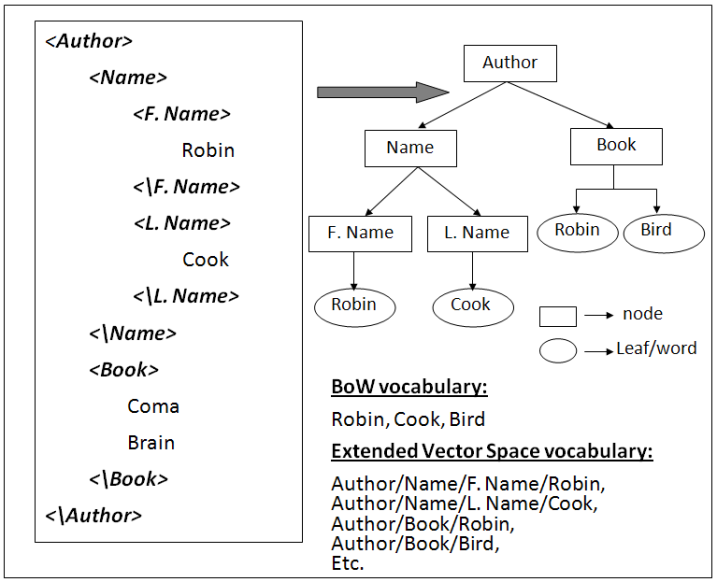
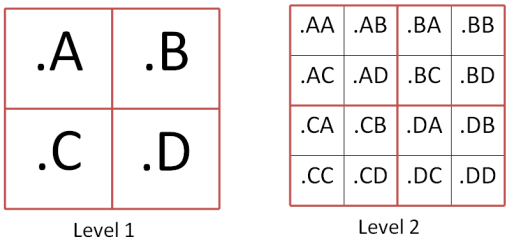


Figure 1. An XML document and its corresponding tree representation with nodes and leaves marked.

complete image which is called the root of the quadtree and is hereby represented by a dot $\{.\}$. This root is broken into four rectangular blocks each with quarter the area and half the perimeter of the root as shown in Fig. 2a. These blocks are called the children of the root and are labeled A, B, C, D according to their position in the root as shown in Fig. 2a. The complete path of these blocks is obtained by adding the path of their root as a prefix to the label. This is called the first level of decomposition. In this case, the complete path of the blocks would be given by $.A, .B, .C, .D$ as shown in Fig.2a. For the second level of decomposition each of these four blocks are further split into four rectangular blocks to get their respective children as shown in Fig. 2b. These children are again labeled A, B, C, D based on their position in their parent block and their complete path is obtained by prefixing their parent's complete pathname to the label as seen in Fig. 2b. We call this the second level of decomposition. Proceeding similarly the image is decomposed into L levels, where the block at the k^{th} level has a k letter long label preceded by a dot. The path of a visual word in the image is hence the complete path corresponding to the smallest block containing the visual word. This path along with other visual words present in the same block contribute to the positional and contextual information of the visual word respectively as explained in Sec. 2.3.

2.3. Extended Vector Space construction

In the extended vector space model of XML document, the vocabulary considers not just words but the context in which the words occur. For example, in Fig. 1, the BoW vocabulary makes no difference between the word "Robin"



(a) (b)

Figure 2. Example of quadtree decomposition: (a) Decomposition at Level 1, (b) Decomposition at Level 2

occurring under the tag "F.Name" and the one occurring under the tag "Book". However, the extended vector space vocabulary makes a clear distinction by adding the associated path with each word. Analogously, in our current approach we propose a method to extend the BoW vector to include the positional and contextual information of visual words in an image.

Prior to defining the extended vector space for an image, a few terms are introduced. The pathname of a block at the k^{th} level of decomposition is given by $\mathcal{P}_0\mathcal{P}_1\mathcal{P}_2\ldots\mathcal{P}_k$, where $\mathcal{P}_0 \in \{.\}$ represents the root image. $\mathcal{P}_i$, where $i > 0$ is defined in Eq. 1.

$$\mathcal{P}_i \in \{A, B, C, D\}, \forall i \in \{1, 2, \ldots, k\} \tag{1}$$

Each visual word in the visual vocabulary is denoted as v_i, where $i \in \{1, 2, \ldots, K\}$. Then, we define a set W^j as follows,

$$W^j = \{v_{i_1}v_{i_2}\ldots v_{i_j} | i_1 \leq i_2 \leq \cdots \leq i_j \in \{1, 2, \ldots K\}\} \tag{2}$$

Here, W^j represents all possible combination of j elements in the visual vocabulary, where $j \in \{1, 2, \ldots, K\}$. Note that the inequality condition in Eq. 2 is necessary to ensure that no visual word combination is repeated. We then proceed to define W, the set of all visual word combinations in the visual vocabulary as shown in Eq. 3.

$$W = W^1 \cup W^2 \cup \cdots \cup W^K \tag{3}$$

The extended vector set can now be defined as the set of all elements of the form $(\mathcal{P}_0\mathcal{P}_1\mathcal{P}_2\ldots\mathcal{P}_k; a)$, where $k \in \{0, 1, 2, \ldots, L\}$ and $a \in W$.

The vector elements present in an image are identified by navigating through each block at the finest level ($L^{th}level$) of decomposition. Now we consider a block with a pathname $P_0P_1\ldots P_L$. Let V be the set of all visual words present in this block. Note that the visual words in set V may not be distinct. For example a particular word may occur twice in the block and would be listed twice in the set. We then define a set w^j as shown in Eq. 4.

$$w^j = \{v_{i_1}v_{i_2}\ldots v_{i_j} | i_1 \leq i_2 \leq \cdots \leq i_j \; and \; v_{i_1}, \ldots v_{i_j} \in V\} \tag{4}$$

It is to be noted that, an element belonging to V should be used atmost once while forming an element of w^j. The definition of w^j is similar to that of W^j in Eq. 2, however w^j corresponds to all possible combination of j elements in the block under consideration. Also, $j \in \{1, 2, \ldots, m\}$, where m is the number of visual words in the block. As an illustrative example let us consider the shaded block in Fig. 3. The pathname of the block is $.AB$ and $m = 2$ in the block. Hence, for the shaded block $w^1 = \{v_1, v_2\}$ and $w^2 = \{v_1v_2\}$.

We also define the set w representing the set of all visual word combinations in a block in Eq. 5. For example, the w corresponding to the shaded block in Fig. 3 is given by $\{v_1, v_2, v_1v_2\}$.

$$w = w^1 \cup w^2 \cup \cdots \cup w^m \tag{5}$$

The extended vector elements present in a block are then given by the set E as shown in Eq. 6.

$$E = \{(P_0P_1\ldots P_k; a) \,|\, k \in \{0, 1, \ldots, L\}, a \in w\} \tag{6}$$

Proceeding similarly, we identify the extended vector elements present in all blocks of the image at the L^{th} decomposition level and calculate their respective frequency of occurrence in the image. A complete example of vector set elements for an image is shown in Fig. 3.

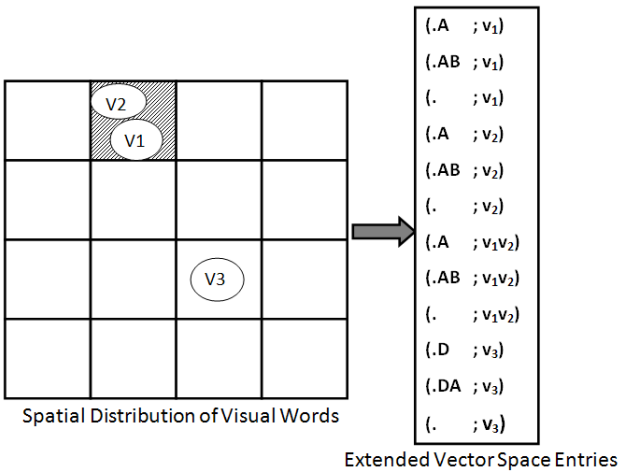


Figure 3. An example of vector set elements present in an image

As seen from Eq. 6, for a particular visual word combination a^c occurring in a block at the L^{th} level, the extended vector elements include the pathname of blocks which contain a^c at all decomposition levels. The need for such a scheme arises from the fact that two similar images can

have a^c located in paths which are not exactly equal but similar upto a certain level of decomposition. For instance, let us consider two images $image_1$ and $image_2$ belonging to the same category. A visual word combination a^c could occur in path $P_0^1 P_1^1 \dots P_L^1$ in $image_1$ and $P_0^2 P_1^2 \dots P_L^2$ in $image_2$, such that $P_i^1 = P_i^2, \forall i < L$. The paths are similar except at the L^{th} level. Hence, including the pathname of blocks at lower resolutions which contain the same term a^c helps in improving the closeness measure between the two images. This approach is similar to the scheme presented in [4] where the histogram matching between two images is performed at different levels of decomposition.

The use of combination of closely occurring visual words in the form of pair of visual words and visual phrases has been shown to provide good performance in [11, 14, 15]. Again, the element becomes rarer with the increase in number of visual words combined to form the element and subsequently its occurrence carries more information. The combination of words help in maintaining the structural details between neighboring patches and give better representation of structural concepts in the image.

The importance of considering positional information is illustrated in Fig. 2.3, where the position of the same visual word combinations are highlighted in three different images. As seen clearly, the position of the visual words plays a major role in differentiating the car image from the image of Buddha. In the absence of positional information the Buddha image is similar to the car image on the basis of visual word frequency.

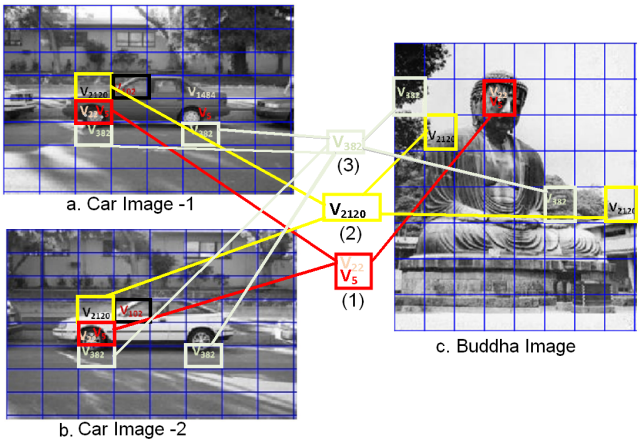


Figure 4. An illustration demonstrating the importance of positional information of visual words in an image. For instance, the visual word combination highlighted by box (1) is present in all three images, however its position is vastly different in Buddha's image.

2.4. Vector Space similarity

A tf-idf weighting approach is used in the proposed scheme to evaluate the similarity between two images. If the extended vector space constructed in Sec. 2.3 is considered to have $\mathcal{N}$ elements then each image in the database can be represented by a $\mathcal{N}$-vector of tf-idf weighted frequencies of the extended vector space elements in the image. The weighted frequency u_n of n^{th} vector space element in an image is given by Eq. 7

$$u_n = \frac{\mathcal{N}_{nd}}{\mathcal{N}_d} log\left(\frac{D}{\mathcal{N}_n}\right) \tag{7}$$

where, $\mathcal{N}_{nd}$ is the number of occurrences of the n^{th} element in the image, $\mathcal{N}_d$ is the total number of vector space elements in the image. D is the total number of images in the database and $\mathcal{N}_n$ is the total number of occurrences of the n^{th} term in the entire image database. $\frac{\mathcal{N}_{nd}}{\mathcal{N}_d}$ is referred to as the term frequency while $log(\frac{D}{\mathcal{N}_n})$ is known as the inverse document frequency.

The similarity measure between two images is provided by the cosine of angle between their respective weighted frequency vectors.

3. Experimental Results

The performance of the proposed scheme is tested on the benchmark Caltech 101 Database. The database has 8707 images belonging to 101 different categories. The number of images in an image category ranges from 30 to 800. Most images in the Caltech 101 database have a centered object which covers the entire image. Hence, the problem of geometrical variance of object representation is minimal as in the case of ideal object based image retrieval. The experiments were carried out by using a set of ten different, randomly sampled images from each image category as query images. Thus, the recall experiments were performed for 1010 query images. All experiments are performed on grayscale images. The effectiveness of our scheme is demonstrated by comparing the results with the traditional BoW approach [7] as well as the SPM based scheme [4]. The SPM based approach also uses positional information of visual words to enhance the BoW scheme and has been shown to be very effective for image classification [4]. Hence, comparative results with SPM based scheme are provided to evaluate our extended vector space approach for image retrieval based on spatial distribution and co-occurrences of visual words.

The visual words in an image are constructed from SIFT features. In the retrieval experiments, a visual vocabulary of 2500 ($K = 2500$) words is used. The maximum quadtree decomposition level L is set to 5 in the experiments. The same maximum resolution is used for the SPM based scheme as well. The Mean Average Precision (MAP) values for the three schemes are calculated over all the query images. Average precision of a single query is defined as the mean of precision scores after each relevant im-

Scheme	MAP
Bag of Words Scheme	0.0820
SPM Scheme	0.0888
Proposed Scheme	0.0978

Table 1. The Mean Average Precision values for different schemes over the 1010 query images for Caltech 101 database.

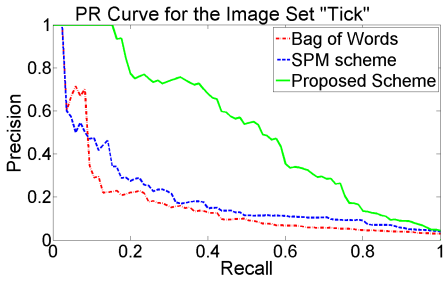
age retrieved. Mean Average Precision (MAP) is defined as the mean of the individual average precision scores for all query images. The proposed scheme is seen to produce an improvement of 19.22% over the traditional BoW scheme, while the SPM based scheme produces an improvement of 8.35%. The MAP values obtained for the BoW scheme, SPM based scheme and proposed scheme are listed in Table 1.

The Precision-Recall (PR) curves for a few image categories are provided in Fig. 5 and Fig. 6. Fig. 5 displays few examples of results where the proposed scheme performed better than BoW and SPM based scheme, while Fig. 6 shows examples of bad results. A few sample query images corresponding to datasets providing good and bad results have been shown in Fig. 5a,5b,5c and Fig. 6a,6b,6c respectively.In general, image sets which had distinct visual patterns and less background clutter responded well to the proposed scheme like the 'dollar bill' image set in the Caltech 101 database. Image sets where the position or orientation of the central object of interest varies significantly within the images in the set, are seen to show relatively poor results when used as query images for the proposed scheme. This could be attributed to the fact that the proposed scheme does not focus on geometrically invariant representation of the object structure. The scheme is also adversely affected by the presence of background clutter in a few image sets like 'Brontosaurus'.

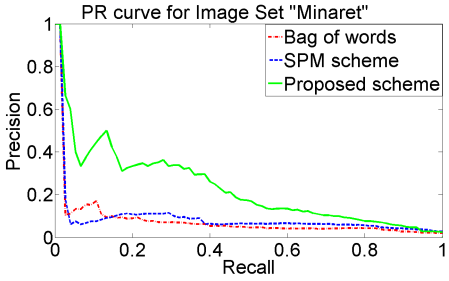
4. Conclusion

The proposed image retrieval scheme based on an extended vector space model, takes into account the context of a visual word in an image along with the co-occurrence of visual words in pre-defined blocks of the image obtained by sub-dividing the image into blocks using quadtree decomposition up to a fixed level of resolution. A significant improvement over existing visual vocabulary based image retrieval schemes in terms of MAP is observed. The execution time is more than the BoW and SPM scheme due to the increase in number of elements in the extended vector space. However, the frequency vector of an image, corresponding to the extended vector space is highly sparse. Hence the increase in computation time is not in proportion with the increase in the size of the vector space. The execution time as well as the MAP can be further improved by eliminating noisy elements in the vector space. An additional scope for

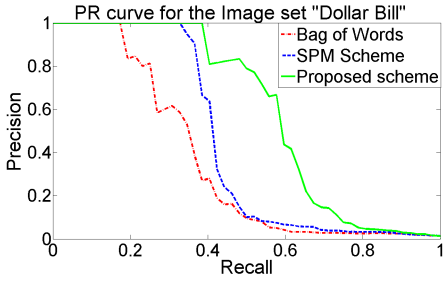
(a) (b) (c)



(d)



(e)



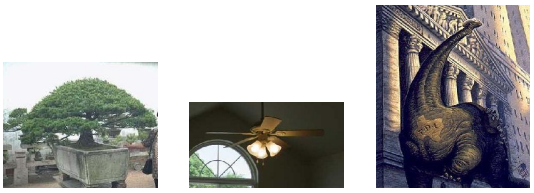
(f)

Figure 5. Sample query images from image sets where the proposed scheme outperformed SPM and BoW Scheme (a) Tick, (b) Minaret and (c) Dollar Bill. The PR curves corresponding to these sets are shown in (d),(e),(f) respectively.

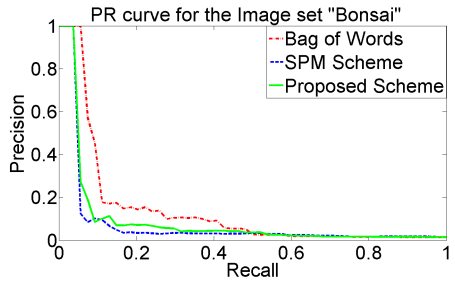
improvement lies in formulation of a geometrically invariant object structure representation based on the proposed extended vector space model for image retrieval.

References

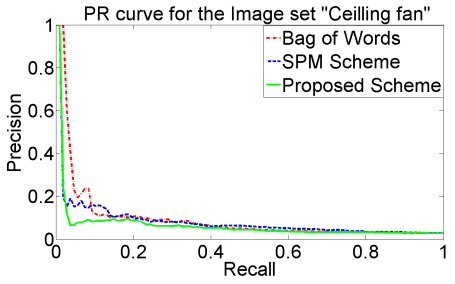
[1] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In



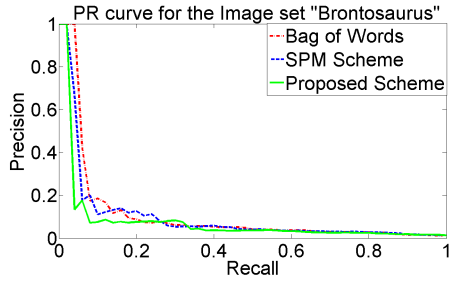
(a) (b) (c)



(d)



(e)



(f)

Figure 6. Sample query images from image sets where the proposed scheme performed bad (a)Bonsai, (b) Ceilling fan and (c) Brontosaurus. The PR curves corresponding to these sets are shown in (d),(e),(f) respectively.

Proc. IEEE International Conference on Computer Vision (ICCV), pages 1458–1465, 2005. 1

[2] E. Hadjidemetriou, M. Grossberg, and S. Nayar. Multiresolution histograms and their use in recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(7):831–847, Jul 2004. 1

[3] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2005. 1

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178, 2006. 1, 4

[5] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004. 1, 2

[6] M. Marszalek and C. Schmid. Spatial weighting for bag-of-features. In *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 2118–2125, 2006. 1

[7] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1470–1477, 2003. 1, 4

[8] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, page 257, 2003. 1

[9] J. Willamowski, D. Arregui, G. Csurka, C. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *Proc. International Conference on Pattern Recognition (ICPR) Workshop on Learning for Adaptable Visual Systems*, 2004. 1

[10] G. Yongming, C. Dehua, and L. Jiajin. An extended vector space model for xml information retrieval. In *Int. Workshop on Knowledge Discovery and Data Mining (WKDD)*, 2009. 1

[11] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, Jun 2007. 1, 4

[12] J. Yuan, Y. Wu, and M. Yang. From frequent itemsets to semantically meaningful visual patterns. In *Proc. ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pages 864–873, 2007. 1

[13] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, Jun 2007. 1

[14] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *Proc. ACM Int. Conf. on Multimedia*, pages 75–84, 2009. 1, 2, 4

[15] Q. F. Zheng, W. Q. Wang, and W. Gao. Effective and efficient object-based image retrieval using visual phrases. In *Proc. ACM Int. Conf. on Multimedia*, pages 77–80, 2006. 4

[16] Y. Zheng, M. Zhao, S. Neo, T. Chua, and Q. Tian. Visual synset:towards a higher level visual representation. In *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, Jun 2008. 1