Vocabulary hierarchy optimisation based on spatial context and category information

Zhiguo Yang, Yuxin Peng* and Jianguo Xiao

Institute of Computer Science and Technology,
Peking University,
Beijing, 100871, China
E-mail: yangzhiguo@pku.edu.cn
E-mail: pengyuxin@pku.edu.cn
E-mail: xjg@pku.edu.cn
*Corresponding author

Abstract: In this paper, we focus on the hierarchy and discriminating ability of visual vocabulary. We propose to use the category information of images and the spatial context of keypoints to select appropriate visual words from different hierarchical levels. Existing approaches, such as flat vocabulary and vocabulary tree, can change the hierarchy of all visual words at the same time, by setting different cluster numbers and tree height respectively. However, the most appropriate visual words may be at different hierarchical levels, and existing approaches could not adjust the hierarchy of different visual words separately. To address this problem, we propose an object function to describe the consistence of visual words, with category information of images and spatial context of keypoints, and then we adopt simulated annealing algorithm to search for a sub-optimal solution, which corresponds to a visual vocabulary selected from the vocabulary tree. Different from existing methods, the proposed approach can select the most appropriate visual words from different levels adaptively, which can improve the performances in image annotation and classification tasks. Experiments on widely-used 15-scenes dataset demonstrate the effectiveness of the proposed approach.

Keywords: bag-of-visual-words; BoVW; vocabulary tree; VT; spatial context; category information; hierarchy selection; simulated annealing.

Reference to this paper should be made as follows: Yang, Z., Peng, Y. and Xiao, J. (2013) ‘Vocabulary hierarchy optimisation based on spatial context and category information’, *Int. J. Multimedia Intelligence and Security*, Vol. 3, No. 1, pp.93–107.

Biographical notes: Zhiguo Yang received his BSc in Computer Science and Technology from Nanjing University, Nanjing, China. He is currently working toward his PhD degree in the Institute of Computer Science and Technology, Peking University, Beijing, China. His research interests include multimedia information retrieval, computer vision and machine learning.

Yuxin Peng is Professor and Director of Multimedia Information Processing Lab (MIPL) in the Institute of Computer Science and Technology (ICST), Peking University. He received his PhD in Computer Application from the School of Electronics Engineering and Computer Science (EECS), Peking University, in July 2003. After that, he worked as an Assistant Professor in ICST, Peking University. From August 2003 to November 2004, he was a Visiting Scholar with the Department of Computer Science, City University of Hong Kong. He has been promoted to an Associate Professor in August 2005

and Professor in August 2010. He has published over 50 papers in international journals and conference proceedings including *TCSVT*, *TIP*, *ACM-MM*, *ICCV*, *CVPR* and *AAAI*. In addition, he has obtained 12 patents. His current research interests include image and video understanding and retrieval, multimedia search and mining, computer vision and machine learning.

Jianguo Xiao is Head and Professor in the Institute of Computer Science and Technology (ICST), Peking University, Beijing, China. He received his MS in Computer Science and Technology from Peking University in 1988. His research interests include image and video processing, and text mining. For his work and contributions, he was the recipient of some famous awards in China, including the first prize of the National S&T Progress Award in 1995, the second prize of the National S&T Progress Award in 2006, 2007 and 2009. In 2008, he won the 7th Guanghua Award of Engineering. In 2010, he won the 10th Bisheng Award for Outstanding Achievement in Printing.

1 Introduction

In the bag-of-visual-words (BoVW) model, which has become very popular in multimedia research, visual vocabulary plays a key role. It is trivial to determine the words and vocabulary in the text domain, where the BoVW model derives from. However, in the image/video domain, the generation of visual vocabulary is a more difficult and error-prone task. The local image patches, which are usually detected or densely-sampled keypoints, are described by statistical features, and then clustered into visual words, which are region centres in the high-dimensional space. Much work has been proposed to generate more accurate visual vocabularies.

Existing approaches can be roughly divided into two categories: flat vocabulary and tree vocabulary. *Flat vocabulary*: Early research such as the work by Sivic and Zisserman (2003) uses k-means to build flat visual vocabulary. In fact, k-means is still one of the most popular methods of vocabulary generation until now. To overcome the high computational complexity of k-means, Philbin et al. (2007, 2008) propose approximate k-means (AKM) algorithm, which adopts '*approximate* nearest neighbours' instead of 'nearest neighbours', and accelerate the quantisation of keypoints by indexing. The AKM method is faster than k-means, and can generate much larger visual vocabularies for large-scale real applications. However, the visual vocabulary generated by AKM algorithm is still single-level flat vocabulary. The problem with flat vocabulary lies in the inconvenience to adjust visual word hierarchy. It is only possible to adjust all the visual words at the same time, by changing the size of visual vocabulary: with smaller vocabulary size, the feature space is split into fewer but larger regions, so all visual words become more general and less discriminative; with larger vocabulary size, the feature space is split into more but smaller regions, so all visual words become less general and more discriminative. *Tree vocabulary*: Nister and Stewenius (2006) propose the vocabulary tree (VT) method, which builds a tree structure as vocabulary. The VT is generated by a top-down hierarchical clustering process, with pre-specified branch number and level number. Ji et al. (2009) also adopt tree vocabulary, where the clustering process may locally end before reaching maximum tree level, if the keypoint number of a leaf node is smaller than a threshold. For tree vocabulary, the hierarchy of leaf nodes can be adjusted by changing the maximum level, or changing the threshold of keypoint number per node. However, it is also very difficult for a tree vocabulary to adjust the hierarchy of different visual words differently. A compromise approach is to use all the nodes at all levels with different weights, which does not solve the problem at root. In one word, both flat vocabulary and tree vocabulary fail to handle the hierarchical diversity of visual words. Some visual words need to be at higher levels to be more general, while other visual words may need to be at lower levels to be more discriminative. The most appropriate visual words may be at different hierarchical levels. Ignoring these facts may lead to inappropriate visual vocabulary.

Recently much research has been done to exploit information other than the statistical information of keypoint. For example, Ji et al. (2010) propose to integrate image category information from *Flickr* labels for supervised vocabulary construction. Zhang et al. (2009) propose to use visual phrases as the visual correspondence to text phrases, where visual phrases refer to the frequently co-occurring visual word pairs. Zheng et al. (2009) construct visual phrases from frequently co-occurring visual word-set with similar spatial context, and further cluster visual phrases into visual synsets, based on class probability distribution. There are also works that use spatial information in kernels (Lazebnik et al., 2006; Lu and Ip, 2009), or to re-rank original retrieval results based on a measure of spatial consistency (Sivic and Zisserman, 2003; Philbin et al., 2007). Motivated by these works, in this paper, we will exploit image category information and spatial information to optimise vocabulary hierarchy.

Our proposed approach is related with but different from two recent approaches. Li et al. (2008) propose to learn optimal compact codebooks by selecting a subset of discriminative visual words from a large visual vocabulary. To do that, they adopt two discriminative criteria, namely likelihood ratio and Fisher, to evaluate the importance of each visual word. Our proposed approach is similar to Li et al. (2008) in that both approaches try to select a subset from an initial large vocabulary of candidate words. But our proposed approach differs from Li et al. (2008) in the following aspects:

1 Li et al. (2008) select the visual words from a flat vocabulary that is generated by k-means, while we select the visual words from an over-segmented hierarchical VT, which enables us to find the most appropriate visual words from different hierarchical levels.

2 The criteria adopted by Li et al. (2008) are only based on the category information of images.

Besides the category information, we also exploit the spatial context of keypoints to evaluate the selected subset of visual words. Another related work is by Li et al. (2011), which proposes to improve the bag-of-words representation by modelling the semantic conceptual relation and spatial neighbouring relation between local patches. Different from our proposed approach which selects an appropriate subset of visual words as visual vocabulary, Li et al. (2011) directly operate on the histogram of words. By hierarchically merging, the bins of related visual word pairs, Li et al. (2011) could generate a multi-resolution representation of the pyramid structure.

The main contribution of this paper can be summarised as follows: We propose to select the most appropriate visual words *from different levels* of the VT. We propose an object function, which is based on the category information of images and the spatial context of keypoints, to describe the hierarchical property of selected subset of visual words. Then simulated annealing algorithm is adopted to search for a sub-optimal solution to the object function, which corresponds to a sub-optimal subset of visual words. With the proposed method, the most appropriate visual words can be selected from different levels adaptively, and thus performance improvements can be achieved in real applications such as image annotation and classification.

The rest of this paper is organised as follows: We will present the detailed algorithm of the proposed approach in Section 2. In Section 3, experiment results on 15-scene categories dataset are reported. And we conclude this paper in Section 4.

2 Proposed approach

2.1 Overview

To generate a hierarchically more proper visual vocabulary, we propose to use the category information of images and the spatial context of keypoints to select the most appropriate visual words from an over-segmented hierarchical tree. The proposed approach consists of four steps, which is summarised as follows and also illustrated in Figure 5:

1 Initialisation:

Generate a *K*-branch *L*-level VT (denoted as *VT*), where *K* and *L* are the parameters that respectively specify the number of children at each non-leaf node and the maximum tree level. Like most approaches with VT, we would use a small *K* value (e.g., 2 or 3), to accelerate the hierarchical clustering process and to simplify the initialisation process. And we would use a fairly large *L* value to ensure that the leaf nodes are over-segmented regions in the feature space, and the most hierarchically proper visual words are the leaf nodes' ancestors that can be found somewhere in the middle of the VT.

2 Object function definition:

Since we are going to select a proper subset of visual words *V* from the *VT*, we need a reliable way to quickly evaluate the quality of the selected subset. In our views, a proper visual vocabulary should have good consistence with the spatial context of keypoints and category information of images, so we will define an object function *E(V)* to assess the quality of the selected word subset by evaluating its consistence with spatial and category information.

3 Search with simulated annealing:

A trivial solution is to find the optimal solution to the object function with brutal search. However, brutal search is too time-consuming and is not affordable. Instead, we adopt the simulated annealing algorithm to search for a sub-optimal solution to the object function, which corresponds to a sub-optimal visual word subset.

4 Soft selection of hierarchy:

After the subset *V* is selected, it is used as the visual vocabulary. We can use the tree structure of *VT* to accelerate keypoint quantisation. Beside the selected visual vocabulary *V*, we also adopt an auxiliary vocabulary *VA*, which consists of all the children of the nodes in selected vocabulary *V*. We conduct the keypoint quantisation with these two visual vocabularies separately, and combine their histogram features with different weights, which is similar to the soft-weighting approach in the quantisation step of flat-vocabulary.

We will explain each step of the proposed approach in detail in the following subsections (2.2 to 2.5).

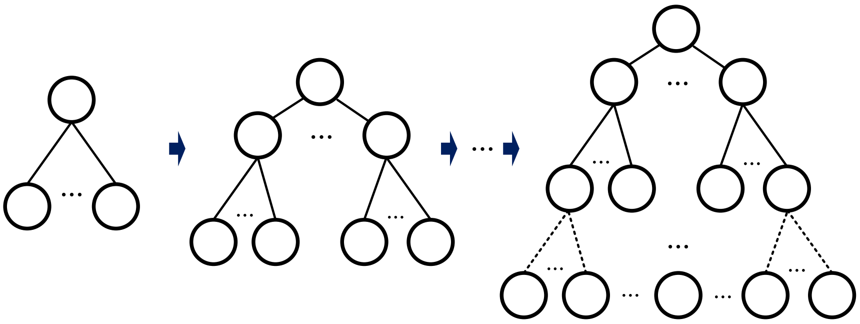
2.2 Initialisation

In the initialisation phase, we adopt the following steps to generate the over-segmented VT, which contains possible candidate visual words at all hierarchical levels (as illustrated in Figure 1):

1. All the keypoints belong to the root node, which is the top level visual word and is obviously not discriminative at all. Then we cluster all keypoints of the root node into *K* clusters using k-means algorithm, and use the *K* cluster centres as the root's children.
2. For each level-*l* node, we cluster its keypoints into *K* clusters, and use the *K* centroids as its level-(*l*+1) children nodes.
3. We continue this clustering process iteratively, from top to bottom, from coarse to fine, splitting the feature space hierarchically into a visual VT. This process ends when the pre-specified maximum level *L* is reached, or when the node contains too few keypoints (less than *minKP*, which is a pre-specified parameter).

In the initialisation step, *L* should be large enough and *minKP* should be small enough, ensuring that the leaf nodes of the tree are over-segmented, and thus the VT contains the proper visual words at some levels in the middle sections.

Figure 1 Initialisation of VT (see online version for colours)



2.3 Object function definition

As described above, *VT* contains visual words at different hierarchical levels: higher-level nodes are usually too coarse and thus not discriminative enough to distinguish keypoints from images of different categories, while lower-level nodes would be too fine and thus not general enough that they may assign keypoints with similar semantic meanings into different visual words. Our aim is to select the visual words at the most appropriate levels, which can tell images from different categories apart, while at the same time keep similar images in the same category. To achieve this goal, at first we must make it clear that we are going to select a subset of word nodes V that obey the following rules:

a Any two word nodes in V do not contain each other, that is, there do not exist two nodes v_i and v_j such that v_i is the ancestor of v_j.

b V is a complete subset of VT, that is, adding any new word node into V breaks rule (a).

With the above two constraints, we could greatly reduce the scope of search without losing important candidate subsets. Then we define an object function, which evaluate the consistence with the spatial context of keypoints and category information of images, enabling us to quickly assess the quality of the selected subset of word nodes.

2.3.1 Consistence with category information

The selected subset of visual words (denoted as V) should be consistent with the category distribution of images. We define the category consistence (CC) of V as follows in equation (1):

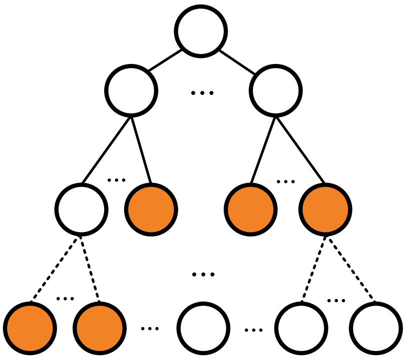
$$CC(V) = \frac{\sum_{v_i \in V} CC(v_i)}{\| V \|} \tag{1}$$

$$CC(v_i) = \max_k \frac{tf(v_i) \cdot l_k}{\|tf(v_i)\| \|l_k\|} \tag{2}$$

$$tf(v_i) = \langle tf_{i1}, tf_{i2}, \ldots, tf_{in} \rangle \tag{3}$$

$$l_k = \langle l_{k1}, l_{k2}, \ldots, l_{kn} \rangle \tag{4}$$

Figure 2 A sample subset (solid orange circles) that obeys the two rules (see online version for colours)



As shown in equation (1), the category consistence (CC) of subset V is the mean value of the category consistence of all visual words v_i in V. Equation (2) calculates the category consistence of single visual word v_i as the maximum cosine similarity between $tf(v_i)$ and l_k, both of which are defined in equations (3) and (4), respectively. As shown in equation

(3), $tf(v_i)$ is the image distribution vector of word node v_i, where tf_{ij} is the occurrence number of visual word v_i in the j^{th} image in training dataset, and n is the number of training images. In equation (4), l_k is the ground-truth image label vector for the k^{th} category, where $l_{kj} \in \{0, 1\}$ is the ground-truth label of the k^{th} category in the j^{th} image. Larger *CC* value of subset *V* means its better consistence with the category information of images.

2.3.2 Consistence with spatial information

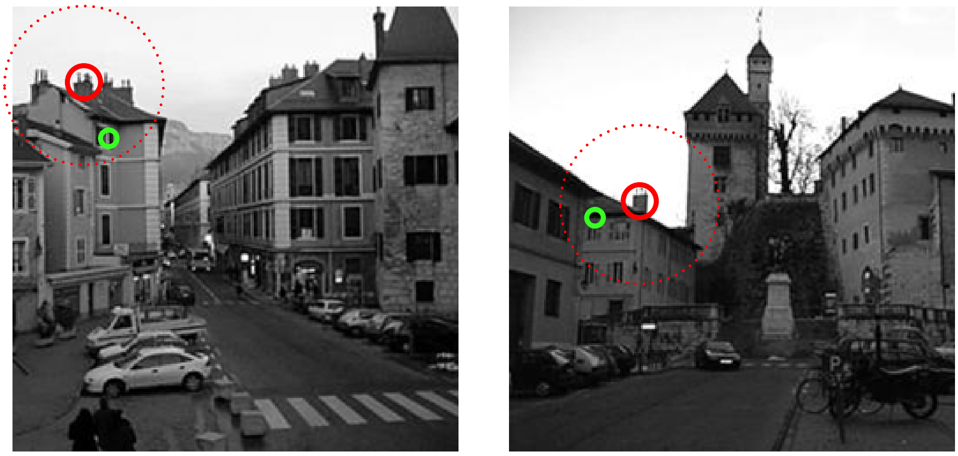
The selected subset of visual words *V* should be consistent with the spatial context of keypoints. We define the spatial consistence (*SC*) of *V* as follows in equation (5):

$$SC(V) = \frac{\sum_{v_i \in V} SC(v_i)}{\| V \|} \tag{5}$$

$$SC(v_i) = \frac{1}{-(K-1)} \sum_{pa}(v_i) = pa(v_i) \frac{nw(v_i) \cdot nw(v_j)}{\left\| nw(v_i) \cdot nw(v_j) \right\|} \tag{6}$$

$$nw(v_i) = \left\langle nw_{i1}, nw_{i2}, \ldots, nw_{im} \right\rangle \tag{7}$$

Figure 3 Co-occurrence with rotation-invariant spatial histogram (see online version for colours)



Notes: The red and green circles are instances of two visual words. The green circle is in the red keypoint's context in both images.

As shown in equation (5), the spatial consistence (*SC*) of subset *V* is the mean value of the spatial consistence of all visual words v_i in *V*. In equation (7), $nw(v_i)$ is the neighbouring-word vector of word node v_i, where nw_{ij} is the co-occurrence number of word node v_i with visual word v_j ($nw_{ij} = 0$ for $i = j$). Similar to Zhang et al.'s (2009) work, here we adopt rotation-invariant spatial histogram (Liu et al., 2008) to count the co-occurrence of visual words, which is illustrated in Figure 3. In equation (6), $SC(v_i)$ reflects the difference of spatial context between v_i and its sibling word nodes, where $pa(v_i)$ denotes the parent node of v_i (we have $pa(v_i) = pa(v_j)$ for sibling nodes). Notice that we introduce the minus sign in equation (6), to change 'similarity' into 'difference'.

Larger $SC(v_i)$ means greater diversity with sibling nodes, and that visual word v_i (and its sibling nodes) shall not be replaced with its parent. Larger $SC(V)$ value means better consistence with spatial information.

$$E(V) = \alpha \cdot CC(V) + 1(1-\alpha) \cdot SC(V) \tag{8}$$

We define the object function as a weighted combination of category consistence and spatial consistence, as shown in equation (8). By optimising the following object function, we could find a hierarchically proper subset of visual words that is consistent with both the spatial context of keypoints and category information of images.

2.4 Search with simulated annealing

Due to the huge computational complexity of brutal search, we adopt the simulated annealing algorithm to search for a sub-optimal solution to the object function, which corresponds to a sub-optimal visual word subset.

The simulated annealing algorithm is inspired by the annealing process in metallurgy, which involves the heating and controlled cooling of a material to reduce defects. By analogy with the physical cooling process, each step the simulated annealing algorithm attempts to replace the current solution by a new solution that is randomly generated near the current solution. The new solution may be accepted with a probability that depends not only on the value change of the object function, but also on the current temperature T, which is gradually decreased during the process.

$$V' = V - \{v_i\} + \{v_j \mid v_j \in VT \wedge pa(v_j) = v_i\} \tag{9}$$

$$V' = V - \{v_j \mid v_j \in VT \wedge pa(v_j) = pa(v_i)\} = pa(v_i) \tag{10}$$

In our proposed approach, each step we may apply one of the following two changes on the current subset V, as illustrated in Figure 4:

a *Top-down change*: replace a randomly-selected word node in the subset, with all its children nodes, as denoted in equation (9).

b *Bottom-up change*: replace a randomly-selected word node v_i and all its sibling nodes, with its parent node, as denoted in equation (10). Notice that this kind of change can apply on the current subset only when all the sibling nodes of v_i are the current subset.

The proposed simulated annealing algorithm for vocabulary hierarchy selection is as follows:

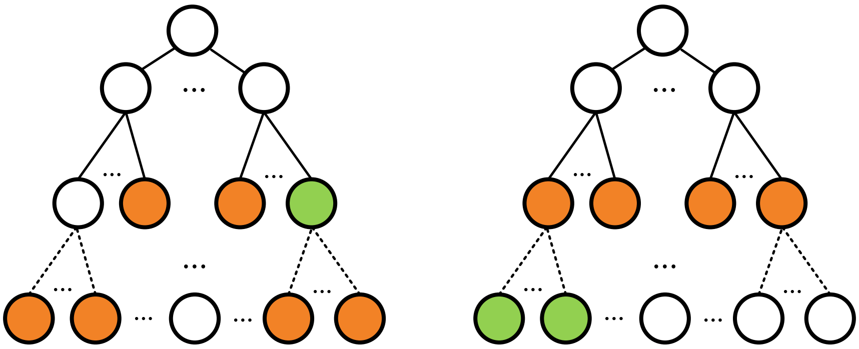
a Select all leaf nodes of VT as the initial subset V.

b Iteratively try to alter the current subset V by applying one of the two changes described above in equations (9) and (10). If the newly-generated subset V' is better than the current subset V (that is, V' has higher value of object function), we will accept the change, and use the newly-generated subset V' to replace V. Otherwise, we may still accept the newly-generated subset V' at a small probability, which is decided by the current temperature T. As the iteration goes, the temperature would gradually decrease, reducing the probability of accepting subsets slightly worse than the current subset.

c The simulated annealing algorithm stops searching when no change is accepted for successive steps at the current temperature, or when the temperature is lower than a pre-specified threshold.

In the above simulated annealing algorithm, we could also adopt a restarting strategy that would roll back to a previous subset that was significantly better than the current subset, rather than always moving from the current state.

Figure 4 Two possible changes based on Figure 2, (a) replace the green circle with its children (change a) (b) replace the green circles with their parent (change b) (see online version for colours)



(a) (b)

2.5 Soft selection of hierarchy

Similar to the soft-weighting approach in the keypoint quantisation of flat-vocabulary, which increases the robustness of quantisation by assigning each keypoint to multiple visual words with the shortest distances, in this paper we adopt the soft selection of visual word hierarchy.

$$VA = \left\{ v_j \mid v_j \in VT \wedge v_i \in V \wedge pa\left(v_j\right) = v_i \right\} \tag{11}$$

Besides the visual vocabulary V that is selected from the VT, we also adopt an auxiliary vocabulary VA that is close to the selected vocabulary V in word hierarchy. The aim of the auxiliary vocabulary is to increase the robustness of the selected visual vocabulary by redundancy of hierarchy. The auxiliary vocabulary VA consists of all the children of the nodes in selected vocabulary V, as denoted in equation (11). When generating the histogram feature for images, we would conduct the keypoint quantisation with V and VA separately, and combine VA-based histogram feature with V-based histogram feature, but with lower weights.

Figure 5 Simulated annealing algorithm for vocabulary hierarchy optimisation

```
Algorithm: Vocabulary hierarchy optimisation with simulated annealing
Input:
    Keypoint set: $P$
    Parameters: $s$, $T$, $\delta$
Initialisation:
    Generate tree structure $VT$ from $P$
    Select leaf nodes as initial word subset $V_0$
Simulated annealing:
Repeat the following steps:
    Repeat for $s$ times:
        Generate new subset $V'$
        Calculate energy change $\Delta E = E(V') - E(V)$
        if $\Delta E > 0$, accept $V'$: $V \leftarrow V'$
        else, accept with probability $(-\Delta E / kT)$
    Update temperature:
        if no change accepted at $T$, stop
        else, update temperature $T \leftarrow \delta T$
Output:
    Final subset of visual words $V^*$
```

3 Experiments

3.1 Evaluation tasks

To evaluate the effectiveness of our proposed vocabulary hierarchy optimisation approach, we conduct our experiments on the following two tasks: image annotation and image classification.

The image annotation task is to automatically annotate descriptive labels on images. Instead of annotating binary decisions (having or not having the concept), the image annotation task usually outputs 0-1 probabilities to indicate the existence of concept. And when evaluating the performance of image annotation, we would rank the images according to their probabilities (that contain the concept) in descending order, and calculate *average precision* (*AP*; Yilmaz and Aslam, 2006) which is the average of precisions at indices across the ranked list where recall changes (i.e., at indices of images containing the concept). For image annotation task with multiple concepts, we would calculate the average of *AP* (*MAP*) as the overall performance metric.

The image classification task is to decide (classify) which one (and only one) of pre-specified categories an image belongs to. For the image classification task, we adopt *mean accuracy* (*MAC*) as evaluation metric, which is the average of diagonal elements in the confusion matrix.

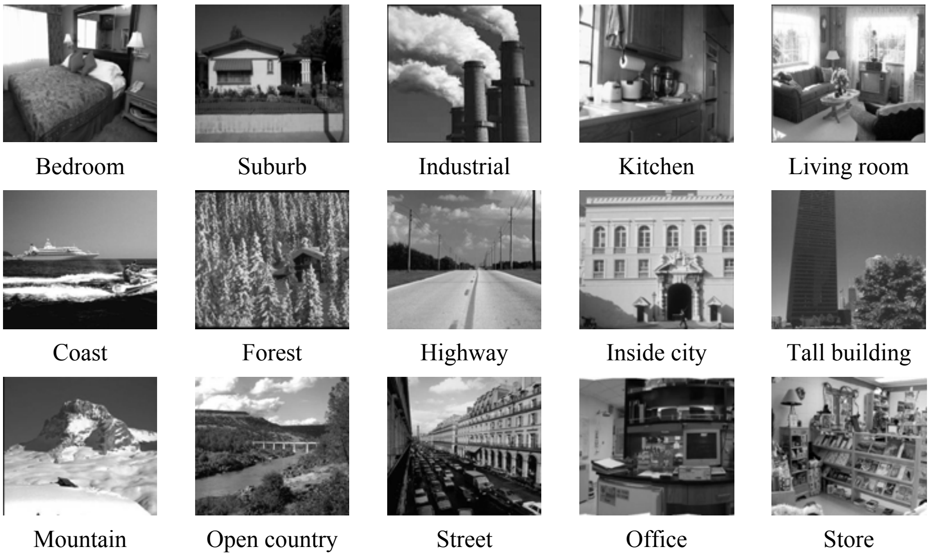
For both image annotation and image classification tasks, higher values of *MAP* and *MAC* imply better performance respectively.

In our experiments, we carry out the image annotation task first, and the image classification task afterwards. In the image annotation task, we adopt the *one-against-all* strategy and annotate the concepts one by one, that is, each time we select one category as the concept to annotate: in the training phase, we adopt images in the selected category as positive samples and the rest categories as negatives samples, based on which we can train classifiers; in the testing phase, for each image we predict the probability of containing the selected concept (category). In the image classification task, we classify the images based on the probability produced in image annotation task: we classify each image into the category with the highest probability. Both the image annotation and image classification tasks rely on the predicted probability of concept existence; the difference between them lies in: in image annotation task, for each concept, we compare the probability between images and sort a ranking list; in image classification task, for each image, we compare the probability between the categories and select only the highest category.

3.2 Experiment setup

We conduct our experiments on *fifteen scene categories* (*15-scenes*) dataset (Lazebnik et al., 2006), which is a common dataset adopted extensively in the research area of image classification. The images are mainly collected from the COREL image collection, personal photographs, and Google image search, by several researchers including Oliva and Torraba (2001), Fei-Fei and Perona (2005), and Lazebnik et al. (2006). The *15-scenes* dataset contains 4,485 images in total, which are divided into 15-scene categories, such as *bedroom*, *kitchen*, *coast*, *forest*, and *highway*. The number of images in each category ranges from 200 to 400. A common practice we follow in this paper is to randomly select 100 images from each category as training images, while using the left images for testing. The average image size of *15-scenes* dataset is about 300 × 250 pixels. Figure 6 shows some sample images, which are resized due to space limit.

Figure 6 Sample images of *15-scenes* dataset, which are resized due to space limit



Bedroom
Suburb
Industrial
Kitchen
Living room
Coast
Forest
Highway
Inside city
Tall building
Mountain
Open country
Street
Office
Store

Since the experimental procedure adopted in Lazebnik et al. (2006) has became the common practice in research works that use the *15-scenes* dataset, to make fair comparison, we strictly follow the same experimental procedures as Lazebnik et al. (2006), which makes our experimental results directly comparable to other papers. We randomly split the dataset into training set and testing set: the training set contains 100 images from each category, while the testing set contains the rest images. For more reliable results, the experiments are carried out on ten different random splits, and the mean value and standard deviation of *MAP* and *MAC* for each split are reported.

In the experiments, we adopt SVM (LibSVM implementation) as classifier, with default parameters and pre-computed histogram intersection kernel.

As for the parameters in the proposed algorithm, although better parameters could be selected by cross-validation, we simply adopt the following settings heuristically and already achieve good results: $K = 2$, $L = 20$, $\min Kp = 50$, $\alpha = 0.5$, and $\delta = 0.9$.

3.3 Experimental results

The experimental results are shown in Table 1, where ‘VocTree’ stands for the VT method adopted from Nister and Stewenius’s (2006) work, while ‘Our’ refers to the proposed approach to optimise vocabulary hierarchy.

For more comprehensive comparison, the same experiments are done for the following three keypoint detectors separately: Difference-of-Gaussian (DoG) (Lowe, 2004), Harris Laplace (Mikolajczyk and Schmid, 2005) and Dense Sampling (Oliva and Torraba, 2001). From Table 1, we can see that the proposed vocabulary hierarchy optimisation approach out-performs the VocTree method, in both image annotation and image classification tasks, and for all three keypoint detectors, which shows its effectiveness and robustness. In Table 2, we further compare our proposed approach with some state-of-the-art methods on the *15-scenes* dataset, where we can see our proposed approach achieves comparable result.

Table 1 Experimental results on 15-scenes dataset

	Annotation (MAP)		*Classification (MAC)*	
	VocTree	*Our*	*VocTree*	*Our*
DoG	0.749 ± 0.007	*0.762* ± 0.006	0.723 ± 0.005	*0.735* ± 0.004
Harris Laplace	0.778 ± 0.004	*0.794* ± 0.004	0.759 ± 0.006	*0.774* ± 0.005
Dense Sampling	0.821 ± 0.005	*0.835* ± 0.007	0.801 ± 0.006	*0.817* ± 0.007

Table 2 Experimental comparison with some state-of-the-art approaches

Approach	*MAC*
van Gemert et al. (2008), ECCV 2008	0.767 ± 0.004
Yang et al. (2009), CVPR 2009	0.803 ± 0.009
Our approach	0.817 ± 0.007

The proposed algorithm can improve the performance of image annotation and image classification, mainly because:

1 the proposed object function can correctly reflect the consistence of selected visual words with image category information and spatial context

2 the simulated annealing search algorithm can find a sub-optimal solution for the object function, which corresponds to a sub-optimal visual vocabulary.

In the experiments, we mainly compare our approach with Nister and Stewenius's (2006) work, which also uses hierarchical vocabulary but does not utilise spatial context and image category information to optimise the vocabulary. Lazebnik et al. (2006) is another research work that is related to our proposed approach, which partitions the image into increasingly fine sub-regions and computes the histograms of keypoints found inside each sub-region. But for each sub-region, Lazebnik et al. (2006) adopts a flat vocabulary generated by k-means algorithm. In fact, our approach and Lazebnik et al. (2006) are complementary to each other and could be combined together to further improve performance.

As for the computational efficiency of the proposed approach, the initialisation step of the proposed algorithm is relatively efficient, since we adopt VT instead of flat vocabulary, and the hierarchical vocabulary has been reported to be more efficient (Nister and Stewenius, 2006). However, the optimisation step could be quite time-consuming, since each step we need to calculate the energy function in equation (8) for all the nodes in the current subset. Fortunately, the computational efficiency of the optimisation step could be significant improved by taking advantage of the fact that in each step only a small part of the nodes could be changed in the current visual word subset. In the proposed approach, each step we may apply one of the following two possible types of change:

a top-down change, which replaces one node with all its K children nodes

b bottom-up change, which replaces one node and all its (K – 1) sibling nodes with its parent node.

Notice that both types of change involve K+1 nodes only, so in each step we only need to update the energy change of K+1 nodes, instead of all the nodes in the current subset.

4 Conclusions

In this paper, we have proposed to use the category information of images and the spatial context of keypoints to optimise vocabulary hierarchy by selecting the most appropriate visual words from different levels of a hierarchical tree, which improves performance in real applications such as image annotation and classification.

Future work will be carried out focusing on the following aspects:

1 We will try to find better object function to more precisely describe the quality of visual words.

2 We will use other information, such as the image segmentation results, to help further optimise the visual vocabulary.

3 We will find further methods to improve the computation efficiency of the proposed optimisation algorithm.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant 61073084, Beijing Natural Science Foundation of China under Grant 4122035, National Hi-Tech Research and Development Program (863 Program) of China under Grant 2012AA012503, National Development and Reform Commission High-tech Program of China under Grant No. [2010]3044, and National Key Technology Research and Development Program of China under Grants 2012BAH07B01 and 2012BAH18B03.

References

Fei-Fei, L. and Perona, P. (2005) 'A Bayesian hierarchical model for learning natural scene categories', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ji, R., Xie, X., Yao, H. and Ma, W-Y. (2009) 'Vocabulary hierarchy optimization for effective and transferable retrieval', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ji, R., Yao, H. and Sun, X. (2010) 'Towards semantic embedding in visual vocabulary', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lazebnik, S., Schmid, C. and Ponce, J. (2006) 'Beyond bags of features: spatial pyramid matching for recognizing natural scene categories', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, T., Mei, T. et al. (2008) 'Learning optimal compact codebook for efficient object categorization', *IEEE Workshop on Applications of Computer Vision (WACV)*.

Li, T., Mei, T. et al. (2011) 'Contextual bag-of-words for visual categorization', *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, Vol. 21, No. 4, pp.381–392.

Liu, D., Hua, G., Viola, P. and Chen, T. (2008) 'Integrated feature selection and higher-order spatial feature extraction for object categorization', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lowe, D.G. (2004) 'Distinctive image features from scale-invariant keypoints', *International Journal of Computer Vision (IJCV)*,Vol. 60, No. 2, pp.91–110.

Lu, Z. and Ip, H.H.S. (2009) 'Image categorization with spatial mismatch Kernels', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mikolajczyk, K. and Schmid, C. (2005) 'A performance evaluation of local descriptors', *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol. 27, No. 10, pp.1615–1630.

Nister, D. and Stewenius, H. (2006) 'Scalable recognition with a vocabulary tree', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2161–2168.

Oliva, A. and Torraba, A. (2001) 'Modeling the shape of the scene: a holistic representation of the spatial envelop', *International Journal of Computer Vision (IJCV)*, Vol. 42, No. 3, pp.145–175.

Philbin, J., Chum, O., Isard, M., Sivic, J. and Zisserman, A. (2007) 'Object retrieval with large vocabulary and fast spatial matching', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Philbin, J., Chum, O., Isard, M., Sivic, J. and Zisserman, A. (2008) 'Lost in quantization: improving particular object retrieval in large scale image databases', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1–8.

Sivic, J. and Zisserman, A. (2003) 'Video Google: a text retrieval approach to object matching in videos', *International Conference on Computer Vision (ICCV)*.

van Gemert, J.C., Geusebroek, J-M., Veenman, C.J. et al. (2008) 'Kernel codebooks for scene categorization', *European Conference on Computer Vision (ECCV)*.

Yang, J., Yu, K., Gong, Y. and Huang, T. (2009) 'Linear spatial pyramid matching using sparse coding for image classification', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yilmaz, E. and Aslam, J.A. (2006) 'Estimating average precision with incomplete and imperfect judgments', *ACM Conference on Information and Knowledge Management (CIKM)*.

Zhang, S., Tian, Q., Hua, G., Huang, Q. and Li, S. (2009) 'Descriptive visual words and visual phrases for image applications', *ACM Multimedia Conference*.

Zheng, Y-T., Neo, S-Y., Chua, T-S. and Tian, Q. (2009) 'Visual Synset: a higher-level visual representation for object-based image retrieval', *The Visual Computer*, pp.1–8.