Optical Engineering 48(4), 047203 (April 2009)

Scene categorization with multiscale category-specific visual words

Jianzhao Qin
Nelson H. C. Yung
The University of Hong Kong
Department of Electrical & Electronic Engineering
Laboratory for Intelligent Transportation
Systems Research
Pokfulam Road
Hong Kong SAR, China
E-mail: jzhqin@eee.hku.hk

Abstract. We propose a novel scene categorization method based on multiscale category-specific visual words. The novelty of the proposed method lies in two aspects: (1) visual words are quantized in a multiscale manner that combines the global-feature-based and local-feature-based scene categorization approaches into a uniform framework; (2) unlike traditional visual word creation methods, which quantize visual words from the entire set of training, we form visual words from the training images grouped in different categories and then collate visual words from different categories to form the final codebook. This generation strategy is capable of enhancing the discriminative ability of the visual words, which is useful for achieving better classification performance. The proposed method is evaluated over two scene classification data sets with 8 and 13 scene categories, respectively. The experimental results show that the classification performance is significantly improved by using the multiscale category-specific visual words over that achieved by using the traditional visual words. Moreover, the proposed method is comparable with the best methods reported in previous literature in terms of classification accuracy rate (88.81% and 85.05% accuracy rates for data sets 1 and 2, respectively) and has the advantage in simplicity.
 [DOI: 10.1117/1.3115471]

Subject terms: scene categorization; multiscale; category-specific; visual words.

Paper 080853R received Oct. 31, 2008; revised manuscript received Feb. 6, 2009; accepted for publication Feb. 8, 2009; published online Apr. 7, 2009.

1 Introduction

Automatic labeling or classification of an image to a specific scene category (e.g., *indoor, outdoor, forest, coast*) is a challenging problem, but finds wide-ranging applications in disciplines such as image retrieval[1–3] and intelligent vehicle and robot navigation.[4,5] Scene classification helps not only to organize image databases, but also to acquire knowledge of the environment in order for an intelligent agent to interact with it. Additionally, the category of a scene provides vital contextual information for object recognition,[6,7] visual surveillance, and other computer vision tasks. The challenge of scene labeling and classification comes from the ambiguity and variability in the content of scene images, which is further worsened by the variations in illumination and scale.

In previously published literature, a popular approach for scene classification is to employ global features to represent the scene. In principle, they consider the whole image as an entity and then rely on low-level features (color, edges intensity, texture, gradient, etc.) to represent the characteristics of the scene. Chang et al.[2] proposed color and texture features as descriptors of the scene. Vailaya et al.[3] used global color distributions, saturation values, and edge direction histograms to describe the scene instead. Siagian and Itti[4,8] proposed a global feature called the *gist* to represent the scene. The gist employs a visual attention model to combine global color, intensity, and orientation features.

Using global features to represent the scene may be sufficient for separating scenes with significant differences in these properties. For example, the colors of *forest* and *inner city* are not the same, and the edges of *tall building* and *mountain* are quite different. On the other hand, if scenes with similar global characteristics (e.g. bedroom versus sitting room) are to be differentiated, then global features may not be discriminative enough. Thus, features extracted from local regions in a scene have been proposed for classification.[9,10]

In the methods proposed by Luo and Savakis[9] and by Vogel and Schiele,[10] types of objects that exist in the scene, such as sky, grass, water, trunks, foliage, field, rocks, flowers, and sand, are identified by supervised learning. The types of the local regions are man-labeled or automatically labeled by a semantic concept classifier based on low-level image features. Theoretically, if the types of the local regions can be successfully identified (i.e., object recognition), the classification of the scene becomes trivial. In practice, however, accurate object recognition remains an unattainable goal at the moment. This is coupled with the fact that a large number of training images is needed to train the classifier for each object. Besides, manually labeling the training images for object recognition is a time-consuming, expensive, and tedious process.

Recently, representing an image by a collection of local image patches of certain size using unsupervised learning methods[11–13] has become very popular and achieved a certain success in visual recognition, image retrieval, scene modeling/categorization, etc., due to its robustness to oc-

clusions, geometric deformations, and illumination variations. In object recognition and image retrieval, the object is represented by a set of visual parts with specific geometric configurations. In scene categorization, the scene type is represented by the co-occurrence of a large number of visual components or the co-occurrence of a certain number of visual topics (intermediate representation).[14,15] In this type of methods, the local image features are quantized into a set of visual words (in analogy to the words in text) to form a codebook. Then an image is represented by the distribution of the visual words in the codebook with or without geometric configurations. Fei-Fei and Perona[16] and Quelhas et al.[17] independently proposed two different unsupervised learning methods to learn visual words from local regions of the scene images, from which the distributions of the visual words are used to represent the images. Additionally, a latent variable called the *theme*[16] is also learned and taken as the intermediate representation of the scene. A comparative study conducted by Bosch et al.[18] has pointed out that using visual-word representations jointly with different techniques, such as probabilistic latent semantic analysis (pLSA)[16,17,19] or latent Dirichlet allocation,[16] one obtains the best classification results for scene classification.

The creation of the visual words is one of the key issues for the visual-word-representation-based methods. The quality of the generated visual words will significantly influence the performance of the classification. The traditional visual-word creation strategy divides an image into patches regularly or based on interest-point detection (e.g., the scale-invariant interest-point detector[20]); then the scale-invariant feature transform (SIFT) features[20] are extracted from the overlapped square patches to describe their gradient features (other features also can be used). The SIFT features then form a feature pool, which is subsequently quantized into N visual words. The visual words are represented by the centroids of the clusters. They describe the set of patches with similar features.

There are two disadvantages of the traditional visual-word-creating strategy. First, since each image is often divided into hundreds of local patches (about 500 in Fei-Fei and Perona's method[16]), the computational burden of performing quantization (clustering) on this large feature pool is heavy in terms of memory and computational time. Considering a 13-category scene classification task and 100 images per category, this will result in 650,000 128-dimensional features. Second, since the clustering algorithm is performed on the whole image set, it cannot guarantee to generate visual words with better discriminative ability for scene classification. For instance, the features extracted from the patches depicting the grass in an *open country* scene and the features extracted from the patches depicting the trees in a *forest* scene may be grouped in the same cluster due to the similarity of these two types of patches in texture and color (Fig. 1). The clustering algorithm will likely quantize the two different features into one visual word. Therefore, this visual word will lose its discriminative ability to separate these two scenes. Some algorithms [e.g., mutual information and the linear support vector machine (SVM) hyperplane normal coefficient[21]] in the field of object recognition have been proposed to select the visual words with better discriminative ability; but after getting the visual words, some useful ones may disappear during the traditional visual-word creation process, which cannot be compensated for by selection. Even if we use visual-word selection to select the most discriminative visual word after their creation, it still may be judged useless for classification, and thus be eliminated in the selection process.

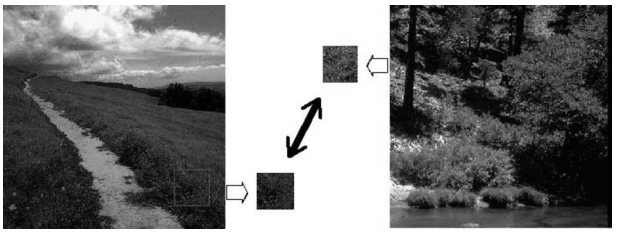


Fig. 1 Two similar patches from different scenes. (The left patch is from Open Country, which depicts trees, and the right patch is from Forest, which depicts grassland).

In this paper, we propose a scene categorization method based on multiscale category-specific visual words, which combines the global features and local features of an image into a uniform framework. Instead of creating the visual words from a single scale or randomly selected scales with limited range as in previous approaches,[16,19] we propose to quantize multiscale features into visual words that cover all scales from the coarsest (global) to the finest (local regions). The multiscale approach provides a richer description of the scene image, which effectively helps to separate scenes of different categories. Furthermore, we introduce a category-specific visual word creation strategy, which can generate more discriminative visual words than the traditional visual-word creation strategy and consume less memory in each clustering operation.

The paper is organized as follows. In Sec. 2, we first formulate the scene classification problem based on the visual-word representation. This is followed by an overview of the proposed method. It describes the method and various steps involved in generating the multiscale visual words using the category-specific creation strategy, as well as the feature extraction process and the classifier training. Section 3 describes how to choose the number of scales, the number of visual words, and the number of training images for visual-word creation. Section 4 presents the experimental results. This paper is concluded in Sec. 5.

2 Proposed Method

2.1 *Problem Formulation*

The scene classification problem based on visual words representation can be formulated in the following manner: Given an image $\mathbf{I} \in \mathbb{R}^{m \times n}$ and a set of scene categories $\mathbf{c}=\{c_1, c_2, \ldots, c_m\}$, we first represent the image $\mathbf{I}$ by a codebook $\mathbf{V}$ consisting a set of visual words $\mathbf{V}=\{v_1, v_2, \ldots, v_k\}$. We denote this representation by $R(\mathbf{I})$; it yields a vector $\mathbf{r}=R(\mathbf{I})$, $\mathbf{r} \in \mathbb{R}^k$, that indicates the distribution or the presence of the visual words. The problem then becomes one of finding a projection

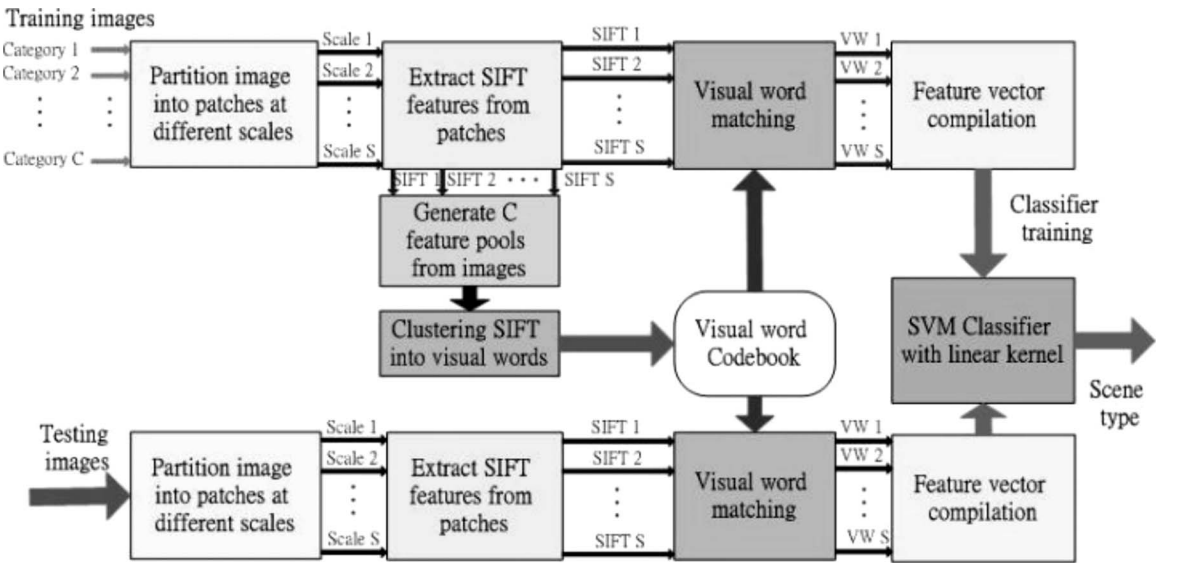


Fig. 2 Framework of the proposed method.

$$f{:}R(\mathbf{I}) \rightarrow \mathbf{c}, \tag{1}$$

that projects the visual-word representation of the image to the scene category c_i, $i=1,\ldots,m$, where it belongs.

2.2 *Overview*

Figure 2 depicts the overall framework of the proposed method. It contains a *training* part that creates a visual word codebook from various categories of images and trains the classifier. For visual-word creation, each training image is divided into regular patches at different scales, from which their SIFT features[20] are extracted. Given the SIFT features, clustering is performed according to different scales and scene categories to create representative visual words, which are represented by the centroids of the clusters. The visual words are then entered into a codebook. From the same training set, each image is evaluated against the visual-word codebook in order to determine a list of visual words that can best represent the image. This list is further compiled into a feature vector to be used for training the classifier.

In the classification of *testing* images, the unknown image (one not found in the training set) is partitioned into patches at different scales and its SIFT features extracted. As in training, a list of visual words that best represent the local features of the image is selected, and then a feature vector is compiled according to this list. Finally, the feature vector is classified by the SVM to obtain the scene type.

2.3 *Representation of the Scene Image Using Multiscale Visual Words*

In this subsection, we introduce the concept of using multiscale visual words to represent a scene image. The previous visual-word creation methods[16,19] creates the visual words from a single scale or randomly selected scales with limited range (from 10 to 30 pixels, with the aim of adapting the scale variation of the visual word). They may fail to describe the image regions at other scales (especially the global characteristic of the entire image). Thus, we propose to quantize visual words on all scales, from the coarsest (global) to the finest regions. The visual words at the scale of the whole image are capable of describing the global characteristic of the image scene, while the visual words with the consecutive smaller scales are able to represent local features with different scales. Therefore, a multiscale visual word combines the global features and local features into a framework that can give us a richer representation of the scene image.

Let us assume the codebook $\mathbf{V}$ consists of a set of multiscale visual words $\mathbf{V}=\{\mathbf{V}_s,\ s=1,2,\ldots,S\}$, where $\mathbf{V}_s=\{\mathbf{v}_{i(s)},\ i=1,2,\ldots n_s\}$ is a set of visual words at scale s. Based on this codebook, we employ a vector $\mathbf{r}=M(\mathbf{I})$, $\mathbf{r}=\{\mathbf{r}_s \in \mathbb{R}^{n_s},\ s=1,2,\ldots,S\} \in \mathbb{R}^{\sum_{s=1}^{S} n_s}$ to represent an given scene image $\mathbf{I} \in \mathbb{R}^{m \times n}$. The vector indicates the distribution or the presence of the visual words at different scales $s=1,2,\ldots,S$ (S is the number of scales, and n_s is the number of visual words at scale s).

To create the multiscale visual words, parts of the images are randomly selected from the training set. These images are regularly divided into overlapped patches at different scales. For scale s, the width and height of the overlapped square patches are $W/2^{s-1}$ and $H/2^{s-1}$, respectively where W is the width of the image and H is its height. Figure 3 depicts the sampling strategy for scales 1, 2, and 3. In scale 1, the whole image is taken as a patch. The features extracted from this patch represent the characteristics of the whole image. In scale 2, the image is divided into nine overlapped patches. The features extracted from the patches represent the characteristics of the regions in the scene image with scale $W/2$ horizontally and $H/2$ vertically. [The number of patches for scale s is $(2^s-1)^2$.] Similarly, the patches from subsequent scales represent the characteristics of regions in the scene with consecutively smaller scales. Figure 4 illustrates the patch samples of a *Coast* scene at different scales.

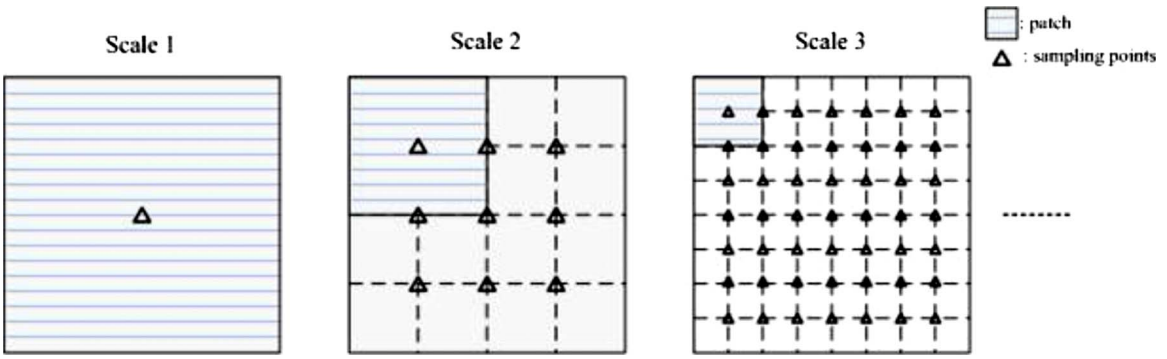


Fig. 3 Overlapped patches at different scales (triangles denote the sampling points).

2.4 *Category-Specific Visual-Word Creation Strategy*

In this subsection, we introduce the category-specific visual-word creation strategy. Figure 5(a) depicts the traditional visual-word creation process. The proposed category-specific visual word creation process is depicted in Fig. 5(b). Instead of quantizing the features from the whole feature pool, we firstly generate C feature pools from C-category scene images separately. Then we quantize the features to create visual words independently from each feature pool. Finally, the visual words are collated to form the final visual-word codebook.

The steps of the category-specific visual-word creation strategy are as follows:

Step 1. Divide the scene images into patches at different scales as described in Sec. 2.3.
Step 2. Extract SIFT features from the patches.
Step 3. Generate C (the number of categories) feature pools at scale s,

$$\mathbf{P}_1^s=\{\mathbf{f}_1^1,\mathbf{f}_2^1,\mathbf{f}_3^1,\ldots\},\quad \mathbf{P}_2^s=\{\mathbf{f}_1^2,\mathbf{f}_2^2,\mathbf{f}_3^2,\ldots\},\ldots,$$
$$\mathbf{P}_C^s=\{\mathbf{f}_1^C,\mathbf{f}_2^C,\mathbf{f}_3^C,\ldots\}.$$

The features in each pool are patch features belonging to the same category at scale s.
Step 4. Quantize the features in each pool separately, using k-means clustering, to create the visual words belonging to category c at scales: $\mathbf{v}_{1(s)}^c,\mathbf{v}_{2(s)}^c,\mathbf{v}_{3(s)}^c,\ldots,\mathbf{v}_{n(s)}^c,\ c=1,\ldots,C$.
Step 5. Group the visual words together to form the final codebook, $\mathbf{V}=\{\mathbf{v}_{1(s)}^c,\mathbf{v}_{2(s)}^c,\mathbf{v}_{3(s)}^c,\ldots,\mathbf{v}_{n(s)}^c\}$.

Figure 6 depicts the patch samples from a *highway* scene, whose features are clustered to create the visual words at scales 2 and 4 using the preceding method. As illustrated, the patch samples at scale 2 depict a road with sky at the top and plants or buildings at the sides. The visual-word created by clustering the features from these patch samples represent a theme. The patch samples at scale 4 depict the sky and the plants beside the roads, i.e., the visual words created from the features extracted from them represent *sky* and *plant*. Comparing these with the visual words at scale 2, we see that the visual words at scale 4 describe the local features in a more detailed manner and represent the components (*sky* and *plant*) more meaningfully. The visual words for the *highway* scene at scale 4 represent the kind of objects that may exist in this scene, while the visual words at scale 2 describe how these objects are organized.

As seen from the category-specific visual-word creation process, the visual words are quantized from features belonging to the same category. The process avoids mixing features that may provide important differentiating cues for classification from different categories. Take the example mentioned previously: The traditional visual-word creation strategy mixes the features of trees and grasses to form a visual word, which reduces its discriminative ability in separating the *forest* scene from the *open country* scene, whereas the category-specific creation strategy is able to separate these two similar features by forcing these patches of similar features from different scene categories to cluster separately. Moreover, for the category-specific strategy, since the feature pool of the traditional strategy is divided

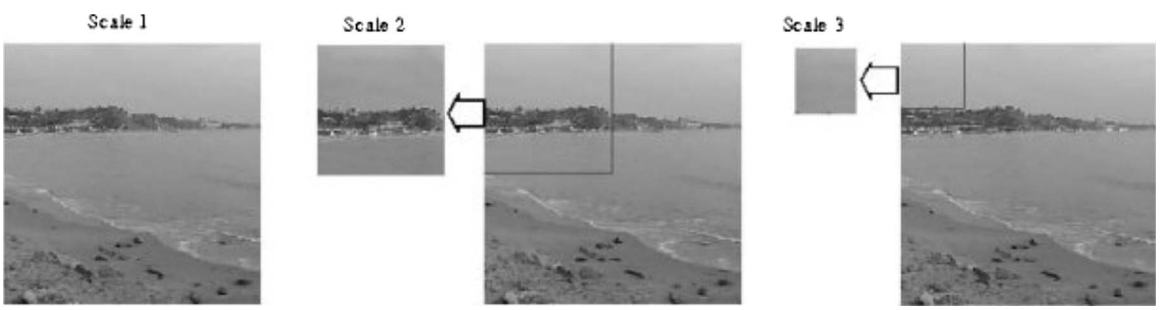


Fig. 4 Patch samples at different scales.

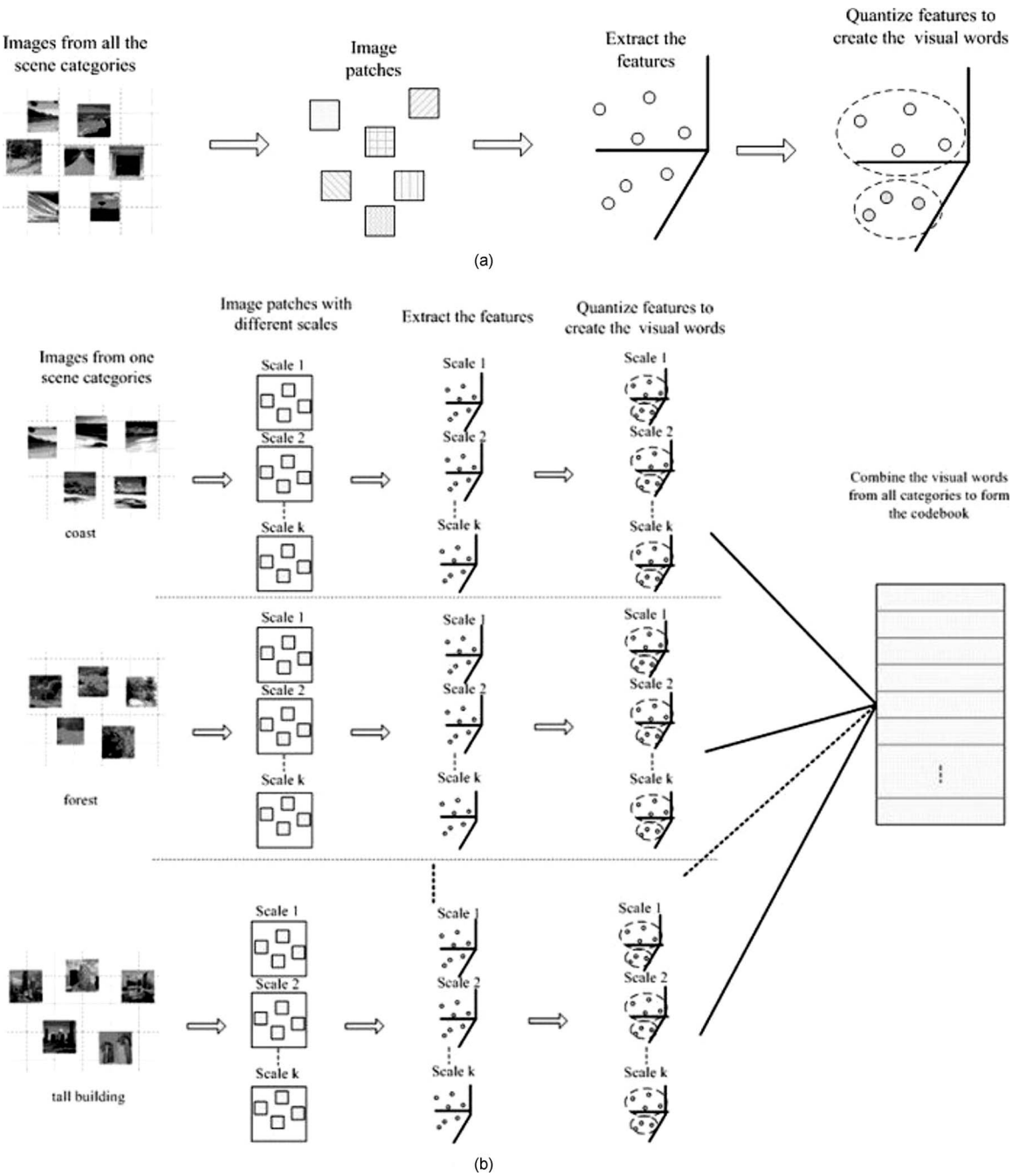


Fig. 5 (a) Traditional visual-word creation procedure; (b) multiscale category-specific visual-word creation procedure.

into C separate feature pools, the clustering operation is performed on a feature pool with size $1/C$ times that in the traditional strategy each time. Thus, the usage of memory is more efficient for the proposed strategy. In other words, if W bytes of memory space is needed for performing clustering in the traditional strategy, only W/C bytes of memory space is consumed by the category-specific strategy at each running of the clustering algorithm.

Inevitably, this visual word creation method will generate redundant visual words. These redundant words can be reduced using the visual-word selection method proposed by Nowak and Juries.[21] In our method, instead of introducing a separate visual-word selection process, we choose to combine the selection process with the classifier training process, which is be discussed in Sec. 2.6. Before that, Sec. 2.5 presents a theoretical analysis of how the category-specific strategy can generate more discriminative visual words.

2.5 *Theoretical Analysis of the Category-Specific Visual-Word Creation Strategy*

In order to measure the quality of the visual words created by different strategies, we employ the following equation to

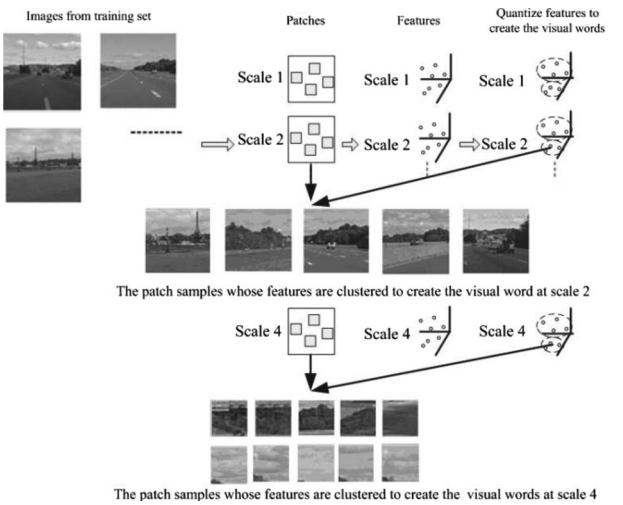


Fig. 6 Patch samples whose features are clustered to create the visual words at scales 2 and 4.

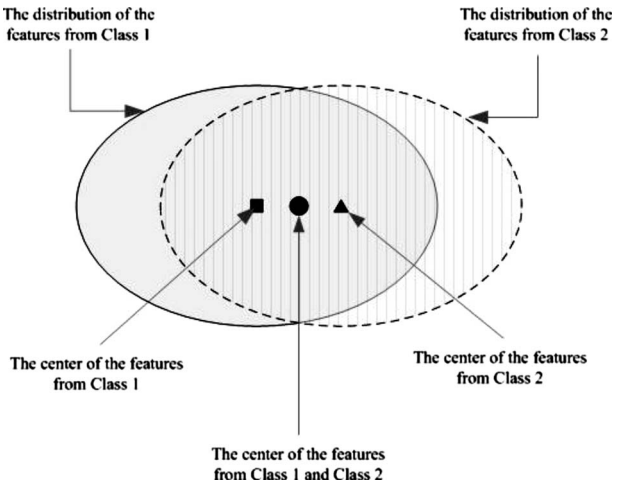


Fig. 7 Distributions of patch features from two different categories.

measure the discriminative ability of a visual word. Denote by c_i the i'th scene category and by C the number of scene categories. Let $\mathbf{V}_j=1$ mean that visual word j is detected in a scene image. The measurement of discriminative ability of visual word j corresponding to class c_i is calculated as

$$I(\mathbf{V}_j,c_i)=-\left\{\log\frac{p(\mathbf{V}_j=1|c_i)p(c_i)}{\Sigma_{k=1}^{C}p(\mathbf{V}_j=1|c_k)p(c_k)}+\log\left[1-\frac{p(\mathbf{V}_j=1|c_i)p(c_i)}{\Sigma_{k=1}^{C}p(\mathbf{V}_j=1|c_k)p(c_k)}\right]\right\}. \quad (2)$$

A larger value of $I(\mathbf{V}_j,c_i)$ means that the visual word has better discriminative ability. If $p(\mathbf{V}_j=1|c_i)p(c_i)/\Sigma_{k=1}^{C}p(\mathbf{V}_j=1|c_k)p(c_k)=0.5$, which means that the probability that visual word j is detected in scene category c_i is the same as the probability that it is detected in other scene categories, it indicates that the presence of $\mathbf{V}_j$ is not able to separate c_i from other scene categories. In this case, $I(\mathbf{V}_j,c_i)$ has its minimum value.

Let us consider a two-class scene classification problem. Figure 7 depicts the distributions of the patch features coming from two different scene categories. The left ellipse, drawn with a solid line, represents the distribution of the patch features from class 1, and the right ellipse, drawn with a dashed line, represents the distribution of the patch features from class 2. If the traditional visual-word creation strategy has been used to generate the visual words, these overlapped patch features may cause the clustering algorithm to consider them as one cluster. The centroid of the patch features will then be used to represent the visual word $\mathbf{V}$. In this case, given a patch feature coming from one of the two classes, the probability of $\mathbf{V}$ given class c is the same, i.e., $p(\mathbf{V}=1|c_1)=p(\mathbf{V}=1|c_2)$. Assuming that $p(c_1)=p(c_2)=0.5$, the discriminative ability value of the visual word $\mathbf{V}$ is $I(\mathbf{V},c_1)=-2\log 0.5$, which is the minimum discriminative value. In other words, the visual word $\mathbf{V}$ cannot provide any useful information for separating the two scene categories. If we employ the category-specific strategy, the clustering is conducted in two different feature spaces separately, and then the centroids of classes 1 and 2 represent the visual words $\mathbf{V}_1$ and $\mathbf{V}_2$, respectively. Given a patch feature $\mathbf{f}$ from class 1, we have $p(\mathbf{V}_1=1|c_1)\geq p(\mathbf{V}_1=1|c_2)$ for $\mathbf{V}_1$ and $p(\mathbf{V}_2=1|c_1)\leq p(\mathbf{V}_2=1|c_2)$ for $\mathbf{V}_2$. Thus, $I(\mathbf{V}_1,C)=I(\mathbf{V}_2,C)\geq -2\log 0.5$, which indicates that $\mathbf{V}_1$ and $\mathbf{V}_2$ are better adapted to the classification problem. The results can be easily extended to the multiclass problem if we take the n-class problem as n separate two-class problems, which separate one class from another class each time. Therefore, it can be concluded that the category-specific visual-word creation strategy generates more discriminative visual words than the traditional strategy.

2.6 *Feature Extraction and Classifier Training*

This subsection presents the steps involved in extracting features from the scene image based on the visual words, and in training the classifier.

Given the visual words, a codebook is used to represent the scene image by calculating the presence of the visual words in the image. Assume that the codebook has n visual words, and a scene image is represented by an n-dimensional vector $\mathbf{x}$. The i'th element in the vector corresponds to the i'th visual word. If the i'th visual word exists in the current image, the corresponding i'th element of the feature vector $\mathbf{x}$ is set to 1, otherwise 0. The feature extraction steps are as follows:

Step 1. Given an image $\mathbf{I}$, divide it into m_s patches at scale s.
Step 2. Extract m_s SIFT features at scale s from the patches.
Step 3. Set $k=1$.
Step 4. For the k'th SIFT feature $\mathbf{f}_k$ that is at scale s, calculate its distance, $d_{kj}=\|\mathbf{f}_k-\mathbf{V}_j\|_2$, $j=s_1,\ldots,s_n$ ($s_1,\ldots,s_n$ are the indices of visual words in the codebook at scale s), to each visual word in the codebook at the same scale s. The k'th patch can be represented by the l'th visual word with the minimum distance to the feature of the patch, $l=\min_j\|\mathbf{f}_i-\mathbf{V}_j\|2$.
Step 5: Set the l'th element of $\mathbf{x}$ to 1.

Step 6. If k equals N ($N=\Sigma_{s=1}^{S} m_s$, the number of visual words), terminate the process; otherwise set $k \leftarrow k+1$ and go back to step 4.

In the training process, images in the training set can be represented as a set of n-dimensional features, $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \cdots, \mathbf{x}_L\}$, where L denotes the number of training images. To formulate it into a SVM classifier, given a training set of labeled data $\{(\mathbf{x}_i, y_i) | (\mathbf{x}_i, y_i) \in R^n \times \{\pm 1\}, \ i = 1, \ldots, L\}$, where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ are the n-dimensional features that have been labeled as $y_1, y_2, \ldots y_L$, the training of the SVM classifier with a linear kernel can be formulated as the following optimization problem for a two-class classification problem:

$$\min_{\mathbf{w},\eta} P\sum_{i=1}^{L} \eta_i + \frac{1}{2}\|\mathbf{w}\|_2$$

$$\text{s.t.} \quad y_i \mathbf{w}^T [\mathbf{x}_i^T 1]^T + \eta_i \geqslant 1, \quad \eta_i \geqslant 0, i = 1, \ldots, L, \tag{3}$$

where $P > 0$ is a penalty parameter ($P=1$ in this paper), and η_i is the slack variable that represents the classification error of $\mathbf{x}_i$. In the classification stage, given a feature vector $\mathbf{x}_t$, the sign of $\mathbf{w}^T[\mathbf{x}_i^T\ 1]^T$ determines the class of this vector. The two-class SVM classifier can be extended to a multiclass situation using the one-against-all method.[22] That is, we take the labels of samples from one class as 1 and the labels of samples from other classes as -1, then solve the preceding optimization problem C (the number of classes) times.

From the formulation of the SVM classifier, we can see that the absolute value of an element of $\mathbf{w}$ determines the importance of the corresponding visual word for classification. In this way, the SVM classifier with linear kernel simultaneously performs feature selection.

3 Parameter Setting Through *k*-Fold Cross-Validation

For such a multiscale and visual-word-based classification method, it is critical to select the right scale, the right number of visual words, and the right number of images used to create the visual words from the training set. In this section, we describe how to set these parameters using k-fold cross-validation.

The method of k-fold cross-validation is very popular for model selection and accuracy estimation.[23] This method first divides the samples into k folds. Then k experiments are performed. In each experiment, $k-1$ folds are selected for training and the remaining one for testing. The true accuracy rate is estimated as the average accuracy rates of the k experiments. Figure 8 depicts a four-fold cross-validation. The advantage of k-fold cross-validation is that all the examples in the data set are eventually used for both training and testing. It overcomes the proneness to misleading estimation of the holdout method, which simply performs a single train-and-test experiment. Of course, it is not as computationally demanding as the cross-validation method. A common choice for k is 10, which is adopted in this paper.

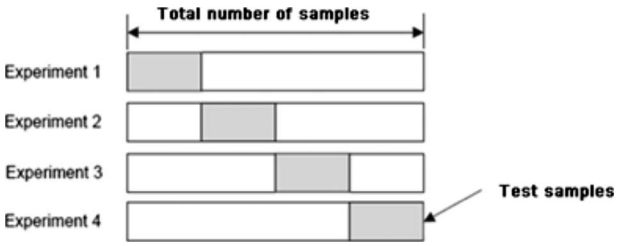


Fig. 8 Four-fold cross-validation.

For setting the number of visual words, we search the parameter space exponentially. That is, we search a number of visual words in the space equal to $2^b, 2^{b+1}, \ldots$ where b denotes the lower bound. We first set the lower bound of the number of visual words at scale s as $N_s = 2^b$. Then, we use the 10-fold cross-validation to estimate the accuracy rate in the training set for $N_s = 2^b, 2^{b+1}, \ldots$ sequentially until the accuracy rate improvement between two consecutive experiments is smaller than a threshold T_1 (we set $T_1 = 0.001$ in this paper) or the accuracy rate is lower than in the previous experiments. Figure 9(a) shows the change of accuracy rates with the increase in the number of visual words at scale 3. It can be seen that the accuracy rate appears to be asymptotic to 0.81, being 0.8175 for 2^9 visual words and 0.8172 for 2^{10}. Given that, 2^{10} is chosen as the best number of visual words for scale 3.

For setting the number of scales, we adopt the same strategy as in setting of the number of visual words. After setting the best number of visual words for scale 1, we obtain the best accuracy rate, denoted as R_1, for it. After setting the best number of visual words for scale 2, we obtain the best accuracy rate when including scale 2. Then, we combine scales 1 and 2 to obtain an accuracy rate denoted by R_2. Following the same procedure, we obtain $R_3, R_4, \ldots$. If $R_i - R_{i-1}$ is smaller than a threshold T_2 (we set $T_2 = 0.005$ in this paper), we choose i as the best number of scales. Figure 9(b) shows the accuracy rate versus the increment of scales (i.e., for scale 1, scale 1+2, scale 1+2+3, scale 1+2+3+4, and scale 1+2+3+4+5), which is obtained using the 10-fold cross-validation in the training set. We can see that the accuracy rate begins to level after including scale 5. Therefore, we choose scale 1+2+3+4+5 as the scale setting for this paper.

The number of images used to create the visual words from the training set may affect the performance and efficiency of the algorithm. On one hand, if too few images are used to create the visual words, the created visual words cannot describe the characteristics of the scene images sufficiently in certain categories, which may reduce the performance of the algorithm. On the other hand, using too many images to create the visual words may result in a large number of patch features for clustering, which is a time-consuming process. Therefore, we also employ the k-fold cross-validation to select the best number of images in the training set for creating the visual words. Figure 9(c) shows the change of accuracy rates with the increase in the number of visual words. [The numbers of visual words are 48 (6 in each category), 80 (10 in each category), 320 (40 in each category), 800 (100 in each category), and 1280 (160 in each category)]. We can see that the effect of the number of

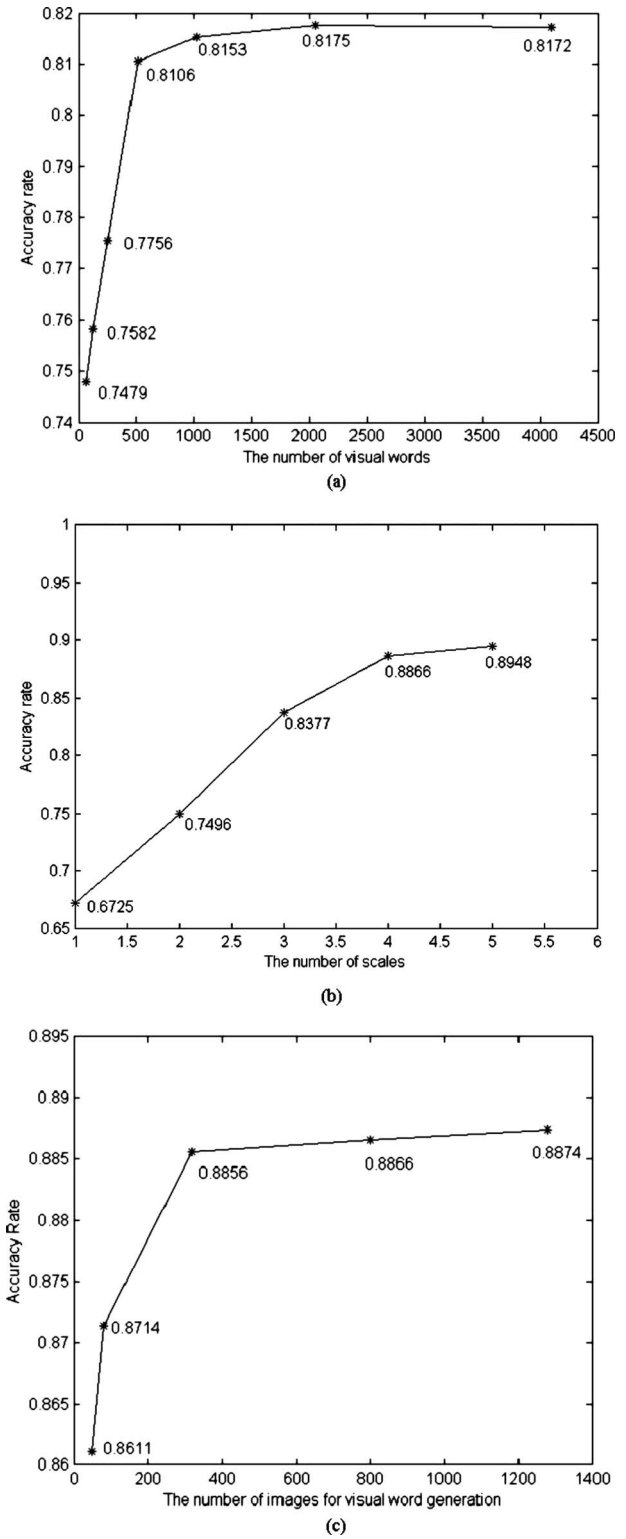


Fig. 9 (a) Accuracy rate versus the number of visual words for scale 3; (b) accuracy rate versus the number of scales; (c): accuracy rate versus the number of images for visual-word generation.

images for visual-word generation on the performance is small compared with the number of visual words and the number of scales. Using only 48 (6 in each category) images, we have obtained an 86.11% accuracy rate, and only a little improvement (88.74%) by including 1280 (160 in each category). By considering the balance between performance and computational burden, we choose 800 images (100 in each category) for the generation of visual words.

4 Experimental Results

The performance of the proposed scene classification method was tested on two data sets, which have been widely used in the previous research.[16,19,24,25] For simplicity's sake, we focus our analysis and discussion on data set 1, whereas we only report the overall results for data set 2.

Data set 1. 2688 color images from 8 categories: *coast* (360 samples), 328 *forest* (328 samples), *mountain* (274 samples), *open country* (410 samples), *highway* (260 samples), *inner city* (308 samples), *tall buildings* (356 samples), and *streets* (292 samples). The average size of each image is 256×256. Gray version of the images is used for our experiments.

Data set 2. 3759 images from 13 categories: *coast* (360 samples), *forest* (328 samples), *mountain* (274 samples), *open country* (410 samples), *highway* (260 samples), *inner city* (308 samples), *tall buildings* (356 samples), *streets* (292 samples), *bedroom* (216 samples), *kitchen* (210 samples), *living room* (289 samples), *office* (215 samples), and *suburb* (241 samples). This data set is an extension of data set 1 by adding five new scene categories.

Figure 10 depicts some samples from data set 1. The experiments reported in Refs. 16, 19, and 24 only use the holdout method to estimate the accuracy rate, i.e., the accuracy rate is estimated by only a single split into training set and test set. This may underestimate or overestimate the accuracy rate. In our experiments, we perform a 10-fold cross-validation in order to achieve a better performance estimation. Moreover, in order to have a reliable comparison between different visual-word creation strategies, as well as a comparison between the multiscale method, single-scale method, and randomly-selected-scale method, we also performed the paired Student *t*-test[26] on the accuracy rates from 10-fold cross-validation at 5% statistical significance level. That is, we want to verify that the mean of the 10 accuracy rates obtained by a method is statistically different from the mean obtained by the other methods. We assume that the distribution of the accuracy rates obtained by using different training sets and test sets is normally distributed, because, given a classification algorithm, accuracy rates are mainly affected by two factors: (1) the features of samples in the training set and (2) the features of samples in the test set. We can safely assume that the distribution of features of the images from a specific scene category is a normal distribution. This is because we can assume a prototype of a scene category (e.g., *forest*). Then the samples are varities of this prototype, including variations in the content (e.g., different kinds of trees, different numbers of trees), in illumination, in scale, etc. The features of the training set are formed by randomly selecting samples from the normal feature distribution. By doing

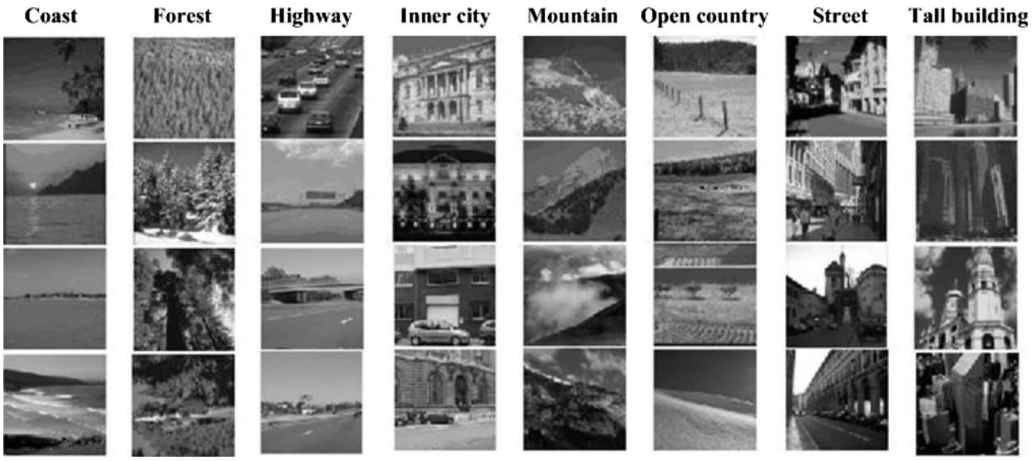


Fig. 10 Samples from data set 1.

so 10 times to form different groups of samples, we get the 10 different training sets for the 10-fold cross-validation. It is obvious that the means of the features of different training sets are normally distributed, which indicates that the characteristics of the different training sets are normally distributed. It is the same for the different test sets. Thus, the combination of these two factors is also normally distributed. Therefore, we can safely assume that the distribution of the accuracy rates obtained by using different training sets and test sets is normally distributed.

Table 1 shows the classification accuracy rates for visual words at each scale created by the traditional visual-word creation strategy and by the category-specific visual-word creation strategy. (Since the traditional visual word creation strategy performs clustering on the whole set of patch features, the memory and time complexity for the clustering operation for scale 5 is huge. Thus, we only do the comparison up to scale 4.) As can be seen, the category-specific visual-word creation strategy outperforms the traditional one significantly and consistently for every scale.

Using the parameter setting procedure introduced in Sec. 3, we generate 32, 128, 2048, and 4096 (total: 6304) visual words for scales 1 to 4, respectively, when the category-specific creation strategy is applied, and 16, 512, 64, and 2048 (total: 2640) visual words when the traditional creation strategy is applied. In order to evaluate the usefulness of the visual words generated by the two different strategies to the classification problem, we also employ the discriminative-ability measurement formulated as Eq. (2). Since the numbers of visual words generated by the two strategies are different, we only extract the most discriminative 1000 visual words generated by each strategy for comparison. In this case, we cannot assume that the distribution of the discriminative-ability measure of the 1000 visual words is normal. Thus, the Wilcoxon rank sum test[27] is performed to test whether the discriminative ability of the visual words generated by the category-specific strategy is better than those generated by the traditional strategy at 5% significance level. The medians of the discriminative value of these visual words are 131.51 for the category-specific strategy and 123.93 for the traditional strategy. The Wilcoxon rank sum test shows that the discriminative ability of the visual words generated by the category-specific strategy is better than that of those generated by the traditional strategy at the 5% significance level.

Table 1 Classification accuracy for each scale of visual words.

Creation strategy	Accuracy (%)			
	Scale 1	Scale 2	Scale 3	Scale 4
Traditional	32.07±3.34	65.98±3.87	68.96±3.83	82.68±2.88
Category-specific	68.09±4.67	73.64±3.04	80.45±3.49	86.56±3.58

Figure 11 depicts the distributions of the multiscale category-specific visual words in the test set. The legend at the right of the figure shows the correspondence between the filled patterns and visual words created from specific scene categories. Figure 11(a) demonstrates the distribution of the visual words at scales 1, 2, 3, and 4 in the test images from *coast* scenes. The four bars filled with light horizontal strokes denote the frequency of visual words specific to the *coast* samples in the training set at scales 1, 2, 3, and 4. As shown, visual words specific to the *coast* samples are the most frequently detected visual words from the *coast* samples in the test set across all the four scales. These bars filled with light horizontal Strokes are obviously higher than the bars filled with other patterns, while they are lower than the other bars in Fig. 11(b)–11(h), which means that the *coast*-specific visual words are most likely to be detected from *coast* scenes while being less likely to be detected from other scenes. Likewise, the same property is also depicted in the distribution of visual words created from other scene categories, which indicates the discriminative ability of the proposed category-specific generation strategy. This distribution figure also offers other useful in-

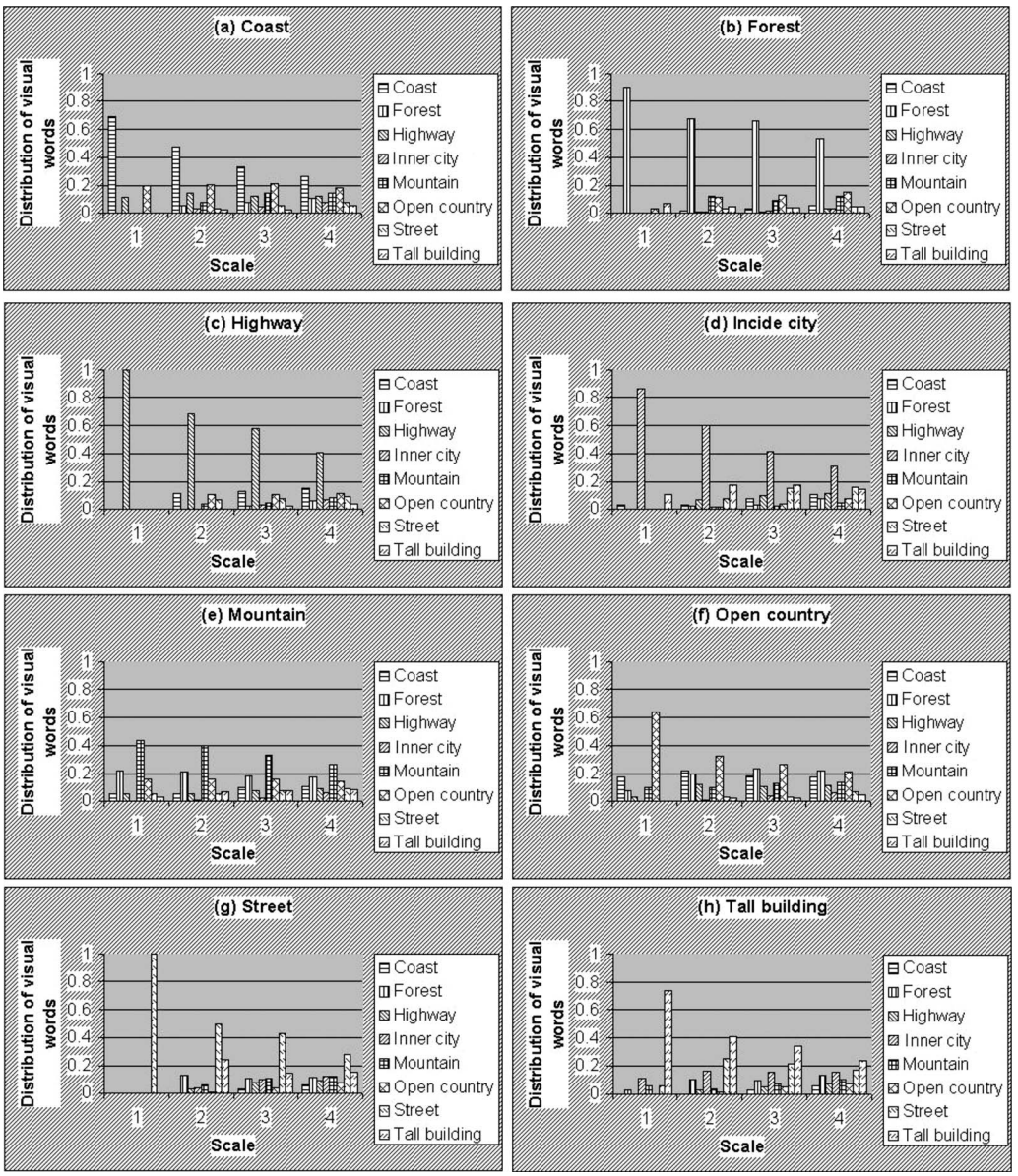


Fig. 11 Distributions of the multiscale category-specific visual words of samples from different scene categories in the test set.

formation. For instance, from Fig. 11(a), we can see that the visual words from *open country* are the second most frequently detected visual words for the samples from *coast*, which means that the visual words from *open country* are also likely to be detected in the *coast* scene. This reveals the similarity between *coast* and *open country* (since we have not included color information in the classification, it is not easy to separate sea water in the *coast* scene from the grassland in the *open country* scene) and also reveals that the *coast* scene may be easier to confuse with the *open country* scene than other scenes are. Similarly, Fig. 11(b) shows that *forest* may be easily misclassified as *open country*. Additionally, from Fig. 11(f), we observe that the distribution difference between the visual words from *open country* and the visual words from other categories are less obvious. It means that the *open country* is much more difficult to differentiate from other scenes, which explains its lowest classification accuracy rate among all the other scenes. This observation is further supported by the following analysis of the confusion table in Table 3.

Table 2 shows the classification accuracy rates after combining the four-scale visual words created by the traditional visual-word creation strategy and the category-specific visual-word creation strategy. The paired *t*-test shows that the category-specific visual-word creation strategy outperforms the traditional visual-word creation strategy after combining the visual words from four scales, with 5% statistical significance level. Comparing the accuracy rates of Table 1 achieved by the single-scale visual words

Table 2 Classification accuracy after combining the four-scale visual words.

	Accuracy (%)	
Creation strategy	Scale 1+2+3+4	Randomly selected scales
Traditional	84.33±2.67	—
Category-specific	88.15±3.18	82.68±4.38

with the accuracy rates of Table 2 achieved by the multiscale visual words, it is noted that the performance is obviously improved by using the multiscale visual words. Moreover, we also compared the performance of the multiscale visual words with the performance of the visual words with randomly selected scales in the range between 10 to 30 pixels as used in Ref. 16. (The positions of the sampling points for the visual words with random scales are the same as the positions of the sampling points at scale 4.) The accuracy rate for the visual words with randomly selected scales is $(82.68 \pm 4.38)\%$, which is poorer than with the proposed multiscale visual words.

Table 3 depicts the confusion table estimated by 10-fold cross-validation for data set 1. The diagonal entries are the average accuracy rates for each scene category. The off-diagonal entries are the percentages of images that are misclassified into other categories. For instance, the element in row 1 and column 6 (not including the title fields) shows that 10.00% of the images that belong to *coast* are misclassified into *open country* (a similar result is shown in row 6, column 1). From row 1 of Table 3, we can see that most of the misclassified images of the category *coast* are mislabeled as *open country*.

Figure 12 presents some of the correctly classified samples, and Fig. 13 presents some of the misclassified samples. The images in the first row of Fig. 13 illustrate some *coast* samples that are incorrectly classified. Most of them are confused with *open country* images. The misclassified *coast* images show a certain similarity to the *open country* images at first glance, especially when there is no other information to help us separate sea water from grass-

Fig. 12 Correctly classified samples.

Table 3 Confusion table of data set 1 (%).

	Coast	*Forest*	*Highway*	*Inner city*	*Mountain*	*Open country*	*Street*	*Tall building*
Coast	87.78	0.56	0.28	0.28	1.11	10.00	0	0
Forest	0	94.87	0	0	2.94	2.19	0	0
Highway	4.23	0.38	85.38	2.31	1.54	4.23	1.92	0
Inner city	0.33	0	0.26	89.51	0	0.26	5.05	4.58
Mountain	2.16	2.43	0.54	0	88.40	6.19	0	0.27
Open country	10.00	2.44	0	0	3.90	83.17	0	0.49
Street	0	0	3.10	6.51	0	0	87.37	3.01
Tall building	0	0.57	0	2.73	1.67	0.29	0.29	94.45

Fig. 13 Misclassified samples. (Words above the images are the labels predicted by the classifier, and the word on the left of each row is the true label of the images in that row.)

land. We also can find the similar case in the sixth row of Fig. 13. It is difficult to identify whether the bottom region of the last image of the sixth row of Fig. 13 is grassland or sea. Figure 14 shows that in this misclassified image, a grassland patch is incorrectly represented by a visual word from *coast* that denotes the sea. This may result in ambiguity between a *coast* scene and an *open country* scene. If color or other features are introduced appropriately to describe the characteristics of the patches, the confusion between *coast* and *open country* may be reduced, which indicates that the SIFT feature is not sufficient to represent the

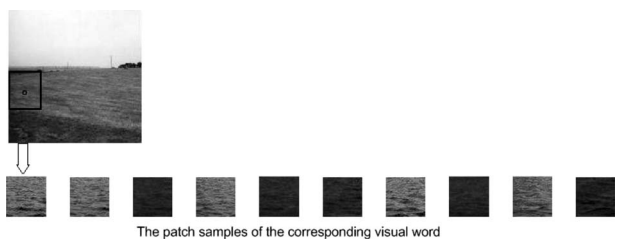


Fig. 14 The grassland patch of the *open country* image is incorrectly represented by the visual word from *coast* that denotes the sea.

Table 4 Results obtained by the proposed method versus previous ones.

	Accuracy (%)	
Data set	Proposed method	Other methods
1	88.81±3.74	87.8,[25] 86.65[19]
2	85.05±2.16	85.9,[25] 65.2,[16] 73.4[19]

local image in some cases. The entries in row 3 of Table 3 show that the *highway* images can be confused with the *coast* and *open country* images due to the similar spatial structure of the scenes (they are scenes with the openness characteristic). Furthermore, the road surface in *highway* also may be confused with the surface of the sea in *coast* or the grassland in *open country*. This ambiguity also may result in misclassification. From row 7 of Table 3, we can observe that a *street* scene may be misclassified as *inner city* or *tall building*. The reason may be that buildings, roads, cars, and pedestrians may exist in all three types of scenes. If buildings occupy a large part of a *street* image, it may be confused with *tall building* or inner city. If roads or cars occupy a large part of a street image, it may be confused with *highway* (e.g., row 7 of Fig. 13).

Table 4 shows the performance of the proposed method versus other results in the literature using the same data sets. We can see that the proposed method achieves comparable performance to the best results obtained among those methods, in terms of accuracy rate. The best results reported in Ref. 25 were obtained with a hybrid generative-discriminative approach, which is much more complex than the proposed method. The performance of the proposed method is slightly superior to their result in data set 1 (by 1.01%), but poorer than their result in data set 2 (by 0.85%). Note that the results reported in Ref. 25 were estimated by only a single training-set–test-set split. This estimation of the performance using the single split is not as reliable as the estimation obtained by the 10-fold cross-validation. In the ten classification rates generated by the ten different training-set–test-set splits using 10-fold cross-validation, for data set 1, the best result we obtained is 95.11%, and the worst result 83.33%. For data set 2, the best result we obtained is 88.98%, and the worst result 81.89%. This shows that without using k-fold cross-validation or other more reliable methods to estimate the performance, a single training–test-set split may generate bias in performance estimation.

5 Conclusion

In this paper, we have presented a scene categorization approach based on the multiscale category-specific visual word. The multiscale visual words give us a richer representation of the scene images, which represents each of them from the whole image down to consecutive smaller regions of it. This representation combines the global-

feature-based approach and the local-feature-based approach into a uniform framework, which can help us to differentiate images from different scene categories. Moreover, the category-specific visual-word creation strategy is capable of generating visual words with better discriminative abilities than the traditional strategy. The theoretical analysis highlights the advantages of the proposed category-specific strategy.

We have tested the proposed method on two data sets with 8 and 13 scene categories, respectively, which are used widely by other research groups. The experimental results have shown that the proposed multiscale category-specific visual words significantly outperform the traditional visual words and that their performance is comparable to the best results reported in the previously published literature in terms of classification accuracy rates. In the proposed method, only using a list of visual words without training other complex models (such as the pLSA or hybrid generative-discriminative model), we have obtained comparable performance, which shows the advantage of the proposed method in simplicity.

As it is, our proposed method has not included the spatial correlations between visual words yet. We believe that spatial information would further improve the proposed method's performance. In our future work, we will consider modeling the spatial correlations between visual words.

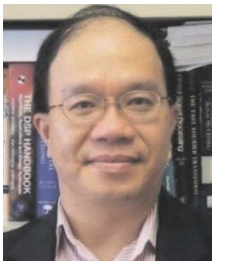
Acknowledgments

The authors would like to thank Antonio Torralba and Fei-Fei Li for providing their data sets, and Dr. Clement Chun Cheong Pang for discussions and suggestions.

References

1. J. Z. Wang, L. Jia, and G. Wiederhold, "SIMPLIcity: semantics-sensitive integrated matching for picture libraries," *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(9), 947–963 (2001).
2. E. Chang, K. Goh, G. Sychay, and W. Gang, "CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.* **13**(1), 26–38 (2003).
3. A. Vailaya, M. Figueiredo, A. Jain, and H. J. Zhang, "Content-based hierarchical classification of vacation images," in *IEEE Int. Conf. on Multimedia Computing and Systems*, M. Figueiredo, Ed., pp. 518–523 (1999).
4. C. Siagian and L. Itti, "Gist: a mobile robotics application of context-based vision in outdoor environment," in *2005 IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition*, L. Itti, Ed., pp. 1063–1069 (2005).
5. R. Manduchi, A. Castano, A. Talukder, and L. Matthies, "Obstacle detection and terrain classification for autonomous off-road navigation," *Auton. Rob.* **18**(1), 81–102 (2005).
6. A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.* **53**(2), 169–191 (2003).
7. A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in *Advances in Neural Information Processing Systems 17 (NIPS)*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. pp. 1401–1408, MIT Press, Cambridge, MA (2005).
8. C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(2), 300–312 (2007).
9. J. Luo and A. Savakis, "Indoor vs. outdoor classification of consumer photographs using low-level and semantic features," in *2001 Int. Conf. on Image Processing*, A. Savakis, Ed., pp. 745–748 (2001).
10. J. Vogel and B. Schiele, "A semantic typicality measure for natural scene categorization," in *DAGM-Sump. 2004*, pp. 195–203, Springer-Verlag (2004).
11. R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Tenth IEEE Int. Conf. on Computer Vision*, L. Fei-Fei, Ed., pp. 1816–1823 (2005).
12. S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(11), 1475–1490 (2004).
13. J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Ninth IEEE Int. Conf. on Computer Vision*, pp. 1470–1477 (2003).
14. P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *Tenth IEEE Int. Conf. on Computer Vision (ICCV 2005)*, pp. 883–890 (2005).
15. F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, P. Perona, Ed., pp. 524–531 (2005).
16. L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, pp. 524–531 (2005).
17. P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *Tenth IEEE Int. Conf. on Computer Vision*, pp. 883–890 (2005).
18. A. Bosch, X. Munoz, and R. Marti, "Which is the best way to organize/classify images by content?," *Image Vis. Comput.* **25**(6), 778–791 (2007).
19. A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *9th Eur. Conf. on Computer Vision, ECCV 2006*, pp. 517–530 (2006).
20. D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Seventh IEEE Int. Conf. on Computer Vision*, pp. 1150–1157 (1999).
21. E. Nowak and F. Juries, "Vehicle categorization: parts for speed and accuracy," in *2nd Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 277–283 (2005).
22. V. Vapnik, Ed., *The Nature of Statistical Learning Theory*, Springer-Verlag (1995).
23. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Int. Joint Conf. on Artificial Intelligence*, pp. 1137–1145 (1995).
24. A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *Int. J. Comput. Vis.* **42**(3), 145–175 (2001).
25. A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(4), 712–727 (2008).
26. D. B. Wright, Ed., *Understanding Statistics: An Introduction for the Social Sciences*, SAGE (1997).
27. W. J. Conover, *Practical Nonparametric Statistics*, John Wiley & Sons (1980).

Jianzhao Qin received his BEng degree and MEng degree from South China University of Technology in 2003 and 2006, respectively. Currently he is a PhD candidate in the Department of Electrical and Electronic Engineering, University of Hong Kong. His research interests include pattern recognition, machine learning, image processing, and brain-computer interfacing.



Nelson H. C. Yung received his BSc and PhD degrees from the University of Newcastle-upon-Tyne. He was a lecturer at the same university from 1985 to 1990. From 1990 to 1993, he worked as a senior research scientist at the Department of Defence, Australia. He joined the University of Hong Kong in late 1993 as an associate professor. Dr. Yung has coauthored five books and book chapters and has published more than 150 journal and conference papers in the areas of digital image processing, parallel algorithms, visual traffic surveillance, autonomous vehicle navigation, and learning algorithms. He acts as consultant to government units and a number of local and international companies. He was a guest editor of the *SPIE Journal of Electronic Imaging*. He also serves as reviewer for a number of IEEE, IET, and SPIE journals. He is a chartered electrical engineer, a member of the HKIE and the IEE, and a senior member of the IEEE.