Scene categorization with multi-scale category-specific visual words

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

ABSTRACT

In this paper, we propose a scene categorization method based on multi-scale category-specific visual words. The proposed method quantizes visual words in a multi-scale manner which combines the global-feature-based and local-feature-based scene categorization approaches into a uniform framework. Unlike traditional visual word creation methods which quantize visual words from the whole training images without considering their categories, we form visual words from the training images grouped in different categories then collate the visual words from different categories to form the final codebook. This category-specific strategy provides us with more discriminative visual words for scene categorization. Based on the codebook, we compile a feature vector that encodes the presence of different visual words to represent a given image. A SVM classifier with linear kernel is then employed to select the features and classify the images. The proposed method is evaluated over two scene classification datasets of 6,447 images altogether using 10-fold cross-validation. The results show that the classification accuracy has been improved significantly comparing with the methods using the traditional visual words. And the proposed method is comparable to the best results published in the previous literatures in terms of classification accuracy rate and has the advantage in terms of simplicity.

Keywords: scene categorization, multi-scale, visual words, category-specific

1. INTRODUCTION

Automatic labeling or classification of an image to a specific scene category (e.g. indoor, outdoor, forest, coast and etc.) is an important and challenging problem, but finds wide ranging applications in disciplines such as image retrieval[1-3] and intelligent vehicle/robot navigation[4, 5]. Scene classification not only helps to organize image databases but also informs an intelligent agent the nature of its environment, which is a critical piece of information for the agent to interact with its environment. Additionally, the category of a scene also provides vital contextual information for object recognition[6, 7], visual surveillance or other computer vision tasks.

The challenge of scene labeling/classification comes from the ambiguity and variability in the content of scene images, which is further worsen by the variation in illumination and scale. In previously published literatures, some popular approaches for scene classification have employed global features of some kind to represent the scene. In principle, they consider the whole image as an entity then relies on low-level features (e.g. 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 and saturation values, edge direction histograms to describe the scene instead. Siagian et al[4, 8], proposed a global feature called 'Gist' to represent the scene. The 'Gist' feature 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 certain types of scenes with obvious different global properties such as color (e.g. forest vs. inside city) and edge (e.g. tall building vs. mountain). If we want to differentiate scenes with similar global characteristic (e.g. office vs. inside city or bedroom vs. sitting room), however, the discriminative ability of global feature is obviously not good enough. Thus, recently, features extracted from local regions in a scene are employed for classification[9, 10]. Luo et al and vogel et al[9, 10] identified the types of objects exist in current scene, such as sky, grass, water, trunks, foliage, field, rocks, flowers, sand and etc. The labels of the local regions are man labeled or automatically labeled by semantic concept classifier based on the low-level image features. Theoretically, if the labels of the local regions can be successfully identified (i.e. object recognition), the classification of the scene may become a trivial problem. In practice, however, robust 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

Intelligent Robots and Computer Vision XXVI: Algorithms and Techniques, edited by David P. Casasent,
Ernest L. Hall, Juha Röning, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7252, 72520N
© 2009 SPIE-IS&T · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.805850

labeling the training images for object recognition is a time-consuming, expensive and tedious process. In order to avoid these, Fei-Fei and Perona[11] and Quelhas et al[12] 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 'theme'[11] also is learned and taken as the intermediate representation of the scene. A comparative study conducted by Bosch et al[13] has pointed out that using visual-word representations jointly with different techniques is the one which obtains the best classification results for scene classification.

In this paper, we propose a scene categorization algorithm based on multi-scale 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 single scale or randomly selected scales with limited range as shown in previous approaches[11, 14] , we propose to quantize multi-scale features into visual words that cover from the coarsest (global) to the finest (local) regions. The multi-scale approach provides 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 traditionally visual word creation strategy. Fig. 1 depicts the framework of the proposed method. It consists of 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 Scale-Invariant Feature Transform (SIFT) features[15] are extracted. Given the SIFT features, clustering is performed according to different scales and scene categories to create representative visual words, which are denoted 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 used of 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 during training, a list of visual word is generated that best represent the image a feature vector is compiled according to this list, which is classified by the SVM to obtain the scene type.

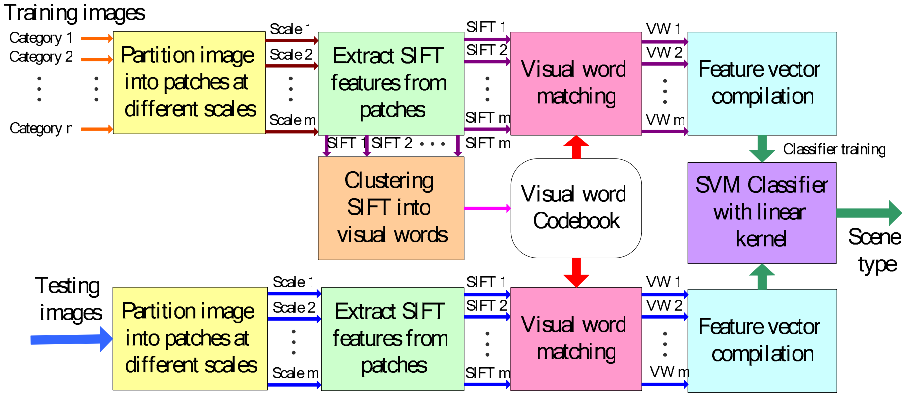
Training images
Category 1
Category 2
Category n
Partition image into patches at different scales
Scale 1
Scale 2
Scale m
Extract SIFT features from patches
SIFT 1
SIFT 2
SIFT m
Visual word matching
VW 1
VW 2
VW m
Feature vector compilation
Classifier training
SIFT 1
SIFT 2 • • •
SIFT m
Clustering SIFT into visual words
Visual word Codebook
SVM Classifier with linear kernel
Scene type
Testing images
Partition image into patches at different scales
Scale 1
Scale 2
Scale m
Extract SIFT features from patches
SIFT 1
SIFT 2
SIFT m
Visual word matching
VW 1
VW 2
VW m
Feature vector compilation

Fig. 1: Framework of the proposed method

The reminder of this paper is organized as follows. In Section 2, we formulate the scene classification problem based on the visual words representation, then present the method and various steps involved to generate the multi-scale visual words using the category-specific creation strategy, the feature extraction process and the classifier training. Section 3 demonstrates the experimental results. Finally, this paper is concluded in Section 4.

2. PROPOSED METHOD

2.1 Representation of the scene image using multi-scale visual words

Representing an image by a collection of local image patches of certain size[16-18] has become very popular and achieved certain success in visual recognition, image retrieval, scene modeling/categorization, etc., due to its robustness to occlusions, 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-occurrences of a large number of visual components or the co-occurrences of a certain number of visual topics (intermediate representation)[12, 19]. 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.

The scene classification problem based on visual words can be formulated in the following manner: given an image $\mathbf{I}\in\Re^{m\times n}$ and a set of scene categories $\mathbf{c}=\{c_1,c_2,\cdots,c_m\}$, we first represent the image $\mathbf{I}$ by a codebook $\mathbf{V}$ consisting a set of visual words $\mathbf{V}=\{v_1,v_2,\cdots,v_k\}$. We denote this representation by $\mathrm{R}(\mathbf{I})$, which is a vector $\mathbf{r}=\mathrm{R}(\mathbf{I}),\mathbf{r}\in\Re^k$ which indicates the distribution or the presence of the visual words. Then, we look for a projection $\mathrm{f}:\mathrm{v}(\mathbf{I})\rightarrow\mathbf{c}$ to project the visual words representation of the image to the scene category $c_i, i=1,\cdots,m$ which it belongs. How to generate the code book $\mathbf{V}$ and how to represent the image by this code book will significantly influence how accurate we can classify the images into their correct categories.

In this subsection, we introduce the concept of using multi-scale visual words to represent a scene image. Instead of creating the visual words from single scale or randomly selected scales with limited range (from 10 to 30 pixels which is aiming at adapting the scale variation of the visual word) as shown in previous approaches[11, 14], which may fail to describe the image regions at other scales (especially the global characteristic of the entire image). We propose to quantize visual words in multi-scale, which covers from the coarsest (global) to the finest regions. The visual words at the scale as large as the whole image are capable of describing the global characteristic of the image scene. And the visual words with the consecutive smaller scales are able to represent local features with different scales. Therefore, multi-scale visual word combines the global features and local-features into a uniform framework which can give us a richer representation of the scene image.

Using the multi-scale visual words, we represent the images in the following way: given an image $\mathbf{I}\in\Re^{m\times n}$, we represent the image $\mathbf{I}$ by a codebook $\mathbf{V}$ consisting a set of multi-scale visual words $\mathbf{V}=\{\mathbf{V}_s, s=1,2,\cdots,S\}$ where $\mathbf{V}_s=\{\mathbf{v}_{i(s)}, i=1,2,\cdots n_s\}$ denotes a set of visual words at Scale s. We denote this representation by $\mathrm{M}(\mathbf{I})$, which is a vector $\mathbf{r}=\mathrm{M}(\mathbf{I}), \mathbf{r}=\{\mathbf{r}_s\in\Re^{n_s}, s=1,2,\cdots,S\}\in\Re^{\sum_{s=1}^{S}n_s}$ which indicates the distribution or the presence of the visual words at different Scale $s, s=1,2,\cdots,S$ (S is the number of scales while n_s represents the number of visual words at Scale s).

To create the multi-scale visual words, parts of the images are randomly selected from the training set. These images are regularly divided into overlapped patches with different scales. For Scale s, the width and height of the overlapped square patches are $\frac{W}{2^{s-1}}$ and $\frac{H}{2^{s-1}}$ respectively where W is the width of the image and H is the height of the image. Fig. 2 depicts the sampling strategy for Scale 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 9 overlapped patches. The features extracted from the patches represent the characteristics of the regions in the scene image with scale $\frac{W}{2}$ horizontally and $\frac{H}{2}$ vertically. (The number of patches for Scale s is $(2^s-1)^2$). Similarly, the patches from subsequent scale levels represent the characteristics of the regions in the scene with consecutively smaller scales.

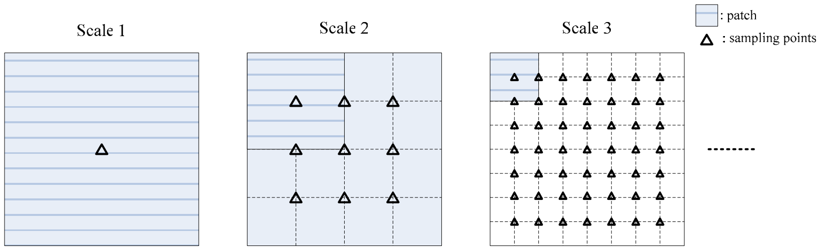

Fig. 2: Overlapped patches at different scales (Note: Triangles denotes the sampling points)

2.2 Category-specific visual word creation strategy

The traditionally visual word creation strategy divides an image into patches regularly or based on the interest point detectors (e.g. scale-invariant interest point detector), then the 128-dim Scale-Invariant Feature Transform (SIFT) features are extracted from the overlapped square patches to describe their gradient features (other features also can be used) (Fig.3(a)). The SIFT features then form a feature pool $\mathbf{P} = \left[\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \cdots, \mathbf{f}_m\right]$, which is subsequently quantized into N visual words. The visual words are represented by the centroids of the clusters using clustering algorithm (e.g. *k*-means). They describe the set of patches with similar features. There are two disadvantages of the traditionally visual word creating strategy. Firstly, since each image is often divided into hundreds of local patches (about 500 in Fei-Fei's method), 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-dim features. Secondly, 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 grasses in *'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 gradient property (Fig. 4). The clustering algorithm will likely quantize the two different features into one visual word. Therefore, this visual word will lose its discriminative ability in separating these two scenes. Although some algorithms (e.g. Mutual information and linear SVM hyperplane normal coefficient[20]) in the field of object recognition have been proposed to select the visual words with better discriminative ability after getting the visual words, some useful visual words may disappear during the traditional visual word creation process, which cannot be compensated for by the visual words selection process. It is because even if we use visual word selection process to select the most discriminative visual word after their creation, it still would be judged as a useless visual word for classification, which would be eliminated in the selection process. In order to generate more discriminative visual words and reduce the memory demand in each clustering performance, we propose a category-specific visual word creation strategy.

The proposed category-specific visual word creation process works as depicted in Fig. 3(b). Instead of quantizing the features from the whole feature pool, we firstly generate C feature pools from C-category scene images separately. Then quantize the features to create visual words independently from each feature pool. Finally, the visual words are collated to form the final visual words/codebook.

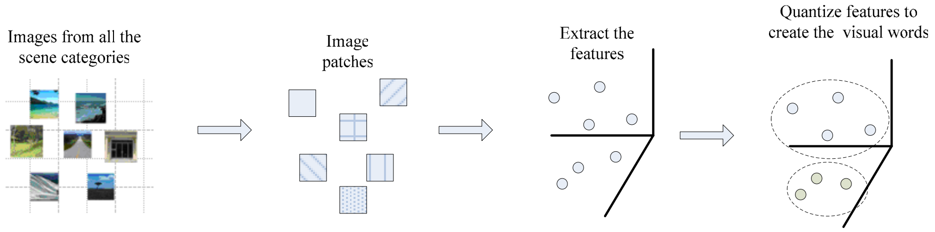
Images from all the scene categories
Image patches
Extract the features
Quantize features to create the visual words

Fig. 3(a): Traditional visual word creation procedure

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 Section 2.1.

Step 2: Extract SIFT features from the patches.

Step 3: Generate C (the number of categories) feature pools at Scale s, $s = 1, \cdots, S$ where S is the number of scales, i.e. $\mathbf{P}_1^s = \{\mathbf{f}_1^1, \mathbf{f}_2^1, \mathbf{f}_3^1, \cdots\}, \mathbf{P}_2^s = \{\mathbf{f}_1^2, \mathbf{f}_2^2, \mathbf{f}_3^2, \cdots\}, \cdots, \mathbf{P}_C^s = \{\mathbf{f}_1^C, \mathbf{f}_2^C, \mathbf{f}_3^C, \cdots\}$. The features in each pool are patch features belonging to the same scene 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 Scale s, $\mathbf{v}_{1(s)}^c, \mathbf{v}_{2(s)}^c, \mathbf{v}_{3(s)}^c, \cdots, \mathbf{v}_{n(s)}^c,\ c = 1, \cdots, C,\ s = 1, \cdots, S$.

Step 5: Group the visual words together to form the final codebook $\mathbf{V} = \{\mathbf{v}_{1(s)}^c, \mathbf{v}_{2(s)}^c, \mathbf{v}_{3(s)}^c, \cdots, \mathbf{v}_{n(s)}^c\},\ c = 1, \cdots, C, s = 1, \cdots, S$.

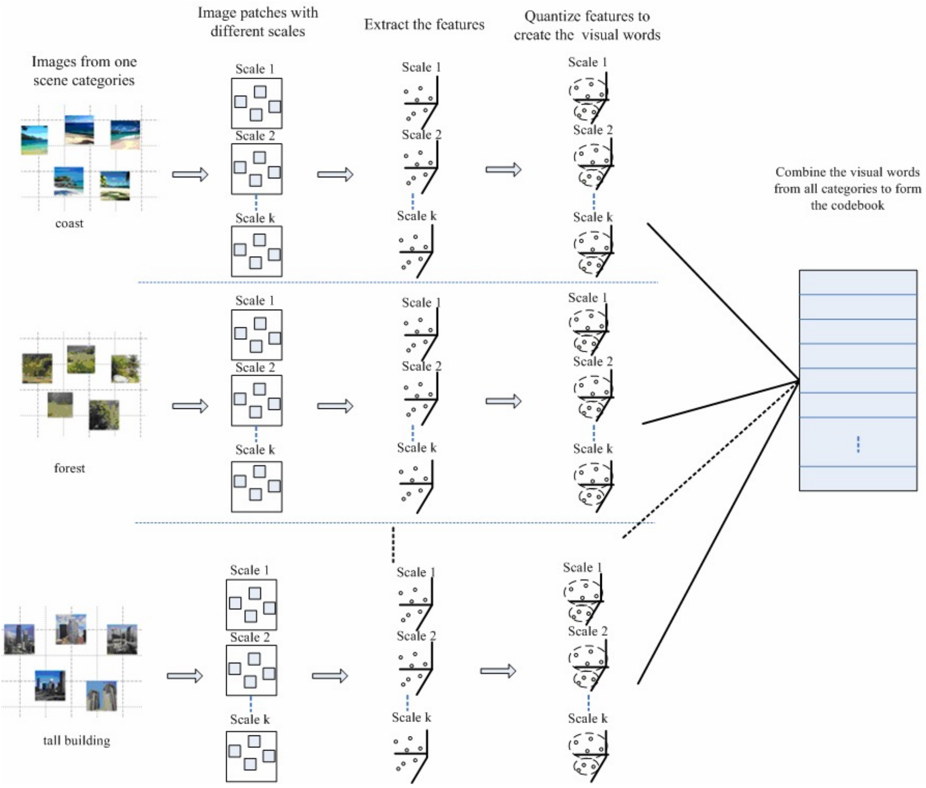
Images from one scene categories
Image patches with different scales
Extract the features
Quantize features to create the visual words
Scale 1
Scale 2
Scale k
coast
forest
tall building
Combine the visual words from all categories to form the codebook

Fig. 3(b): Multi-scale category-specific visual word creation procedure

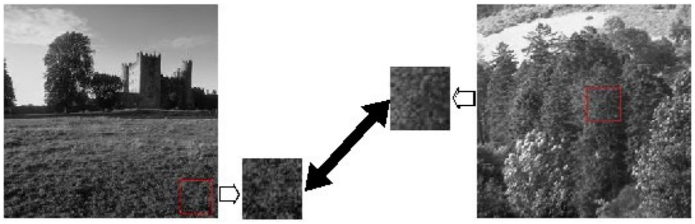

Fig. 4: Two similar patches from different scenes. (Note: The left patch is from ‘forest’ which depicts the trees and the right patch is from ‘open country’ which depicts the grassland)

As seen from the category-specific visual word creation process, the visual words are quantized from the features belonging to the same category. It avoids mixing features which may provide important differentiating cues for classification from different categories. Take the example as mentioned previously, the traditional visual word creation strategy will mix the features of trees and grasses to form a visual word which losses 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 into C separate feature pools, the clustering operation is performed on the feature pool with $\frac{1}{C}$ size comparing with the traditional strategy each time. Thus, the usage of memory is more efficient for the proposed strategy. That is, if W bytes of memory space is needed for performing clustering in the traditional strategy, only $\frac{W}{C}$ bytes of memory space is consumed by the category-specific strategy at each running of the clustering algorithm.

Inevitably, this visual word creation process may generate redundant visual words. These redundant visual words can be reduced using the visual words selection method proposed by Nowak[20]. In our method, instead of introducing a separate visual words selection process, we choose to combine the selection process with the classifier training process which will be discussed in Section 2.3.

2.3 Feature extraction and classifier training

This section 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. Assuming that the codebook has n visual words, then the presence of the visual words is represented by a 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 $s, s = 1,2,\cdots,S$.

Step 2: Extract m_s SIFT features at Scale $s, s = 1,2,\cdots,S$ from the patches.

Step 3: Set $k = 1$.

Step 4: For the k^{th} SIFT feature $\mathbf{f}_k$ which is at Scale s , we calculate its distance, $d_{kj} = \left\|\mathbf{f}_k - \mathbf{V}_j\right\|_2$, $j = s_1,\cdots,s_n$ ($s_1,\cdots,s_n$ is the index of visual words at Scale s in the codebook), to each visual word in the codebook from 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 patch, i.e. $\ell = \min_j \left\|\mathbf{f}_i - \mathbf{V}_j\right\|_2$.

Step 5: Set the l^{th} element of $\mathbf{x}$ to 1

Step 6: If k equals n (the number of visual words), terminate the process, otherwise $k = k+1$ go back to Step 4.

In the training process, images in the training set can be represented as a set of n -dim features, $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\cdots,\mathbf{x}_M\}$, where M denotes the number of training images. To formulate it into a SVM classifier, given a training set of labeled data, $\{(\mathbf{x}_i, y_i) \,|\, (\mathbf{x}_i, y_i) \in \Re^n \times \{\pm 1\}, i = 1,\cdots,M$, where $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_M$ are the n dimensional features that have been labeled as $y_1, y_2, \ldots y_M$, the training of SVM classifier with linear kernel can be formulated as the following optimization problem for a 2-class classification problem:

$$\begin{aligned} &\min_{\mathbf{w},\eta} P\sum_{i=1}^{M} \eta_i + \frac{1}{2}\left\|\mathbf{w}\right\|_2 \\ &s.t. \quad y_i \mathbf{w}^T [\mathbf{x}_i^T \; 1]^T + \eta_i \geq 1, \quad \eta_i \geq 0, i = 1,\ldots,M \end{aligned} \quad , \tag{1}$$

where $P > 0$ is a penalty parameter, 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 2-class SVM classifier can be extended to a multi-class situation using the one-against-all method. 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 above optimization problem for m (the number of classes) times. From the formulation of the SVM classifier, we can see that the absolute value of the elements 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. EXPERIMENTAL RESULTS

Performance of the proposed scene classification method is tested based on two datasets which has been widely used in previous research[11, 14, 21]. For simplicity sake, we focus our analysis and discussion on Dataset 1, whereas we only report the overall results for Dataset 2.

Dataset 1: Consists of 2688 color images from 8 categories: coast (360 samples), 328 forest (328 samples), mountain (274 samples), open country (410 samples), highway (260 samples), inside city (308 samples), tall buildings (356 samples), and streets (292 samples). Gray version of the images is used for our experiments.

Dataset 2: Contains 3759 images from 13 categories: coasts (360 samples), forest (328 samples), mountain (274 samples), open country (410 samples), highway (260 samples), inside 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 dataset is an extension of Dataset 1 by adding 5 new scene categories.

Fig. 5 depicts some samples from Dataset 1. The experiments conducted in[11, 14, 22] only use the holdout method to estimate the accuracy rate, i.e. the accuracy rate is estimated by only a single training set and test set split. This may overestimate the accuracy rate. In our experiments, we perform 10-fold cross-validation in order to achieve better performance estimation. Moreover, to compare different visual words creation strategies, we also calculated statistical significance level of 95% using paired Student t-test[23].

Table 1 shows the classification accuracy rates (%) for visual words at each scale created by the traditional visual word creation strategy and the category-specific visual word creation strategy respectively. (Since the traditional visual word creation strategy performs clustering in the whole set of patch features, the memory and time-consumption for the clustering operation to the patch features at 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 visual word creation strategy significantly and consistently for every scale with 95% significance level.

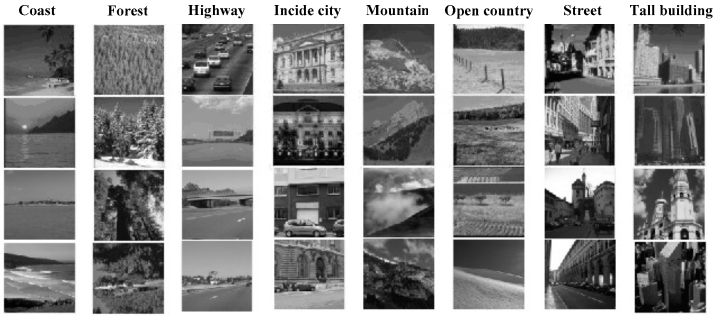

Fig. 5: Samples from Dataset 1

Table 1: Classification accuracy (%) for visual words at scale 1, 2, 3, 4 respectively

Scale 1	Scale 2	Scale 3	Scale 4
Traditional visual word creation strategy			
32.07 ± 3.34	65.98 ± 3.87	68.96 ± 3.83	82.68 ± 2.88
Category-specific visual word creation strategy			
68.09 ± 4.67	73.64 ± 3.04	80.45 ± 3.49	86.56 ± 3.58

Table 2 shows the classification accuracy rates (%) after combining the 4-scale visual words created by the traditional visual word creation strategy and the category-specific visual word creation strategy respectively. The paired t-test show that the category-specific visual word creation strategy outperforms the traditional visual word creation strategy after combining the visual words from 4 scales with statistical significance level 95%. Comparing the accuracy rates of Table 1 achieved by the single scale visual words with the accuracy rates of Table 2 achieved by the multi-scale visual words, it is noted that the performance is obviously improved by using the multi-scale visual words. Moreover, we also compared the performance of the multi-scale visual words with the performance of the visual words with randomly selected scales in the range between 10 to 30 pixels used in[11] (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 performance of the visual words with randomly selected scales is 82.68 ± 4.38 % accuracy rate, which is poorer than the proposed multi-scale visual words.

Table 2: Classification accuracy after combining the 4-scale visual words created

Scale 1+2+3+4	Randomly selected scales
Traditional visual word creation strategy	
84.33 ± 2.67%	---
Category-specific visual word creation strategy	
88.15 ± 3.18%	82.68 ± 4.38%

Fig. 6(a) presents some of the correctly classified samples. The images in the first row of Fig. 6(b) illustrate some misclassified samples. The misclassified *'Coast'* images show certain similarity to the *'Open country'* images at first glance especially when there is no color information to help us separate the sea water from the grassland. We also can find the similar case in the 6th row of Fig. 6(b). It is difficult to identify whether the bottom region of the last image of the 6th row of Fig. 6(b) is grassland or sea. This may result in the ambiguity between the *'Coast'* scene and *'Open Country'* scene. If color is introduced appropriately to describe the characteristics of the patches, the confusion between *'Coast'* and *'Open country'* may be reduced. From the Row 7 of Table 3, we can observe that the *'Street'* scene may be misclassified into *'Inside city'* scene and *'Tall building'* scene. The reason may be that buildings, roads, cars, pedestrians may exist in all three types of scenes. If buildings occupy large part of the street image, it may be confused with *'Tall building'* or *'Inside city'*. If roads, cars occupy large part of the street images, it may be confused with *'Highway'* (e.g. Row 7 of Fig. 6(b)).

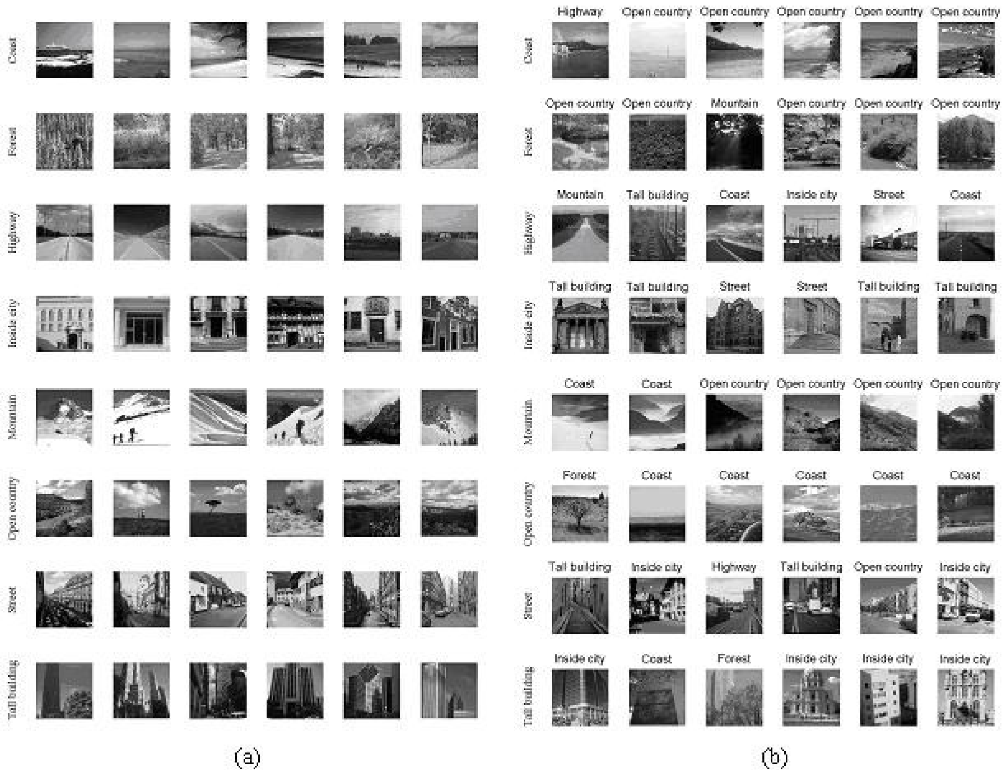

Fig. 6: (a) Correctly classified samples; (b) Misclassified samples (Note: The words above the images are the predicted labels by the classifier and the word on the left of each row are the true labels of the images in that row.)

Table 3 shows the performance of the proposed method versus other results in the literatures using the same datasets. We can see that the proposed method achieves comparable performance to the best results obtained among these methods in terms of accuracy rate. The best results reported in[22] proposed a hybrid generative/discriminative approach which is more complex than the proposed method. The performance of the proposed method is lightly superior to their result in dataset 1 by 1.01% but poorer than their result in dataset 2 by 0.85% (Note that the results reported in the paper[22] achieving the best result were estimated by only a single training set and test set split. This may overestimate the accuracy rate.)

Table 3: Results obtained by the proposed method versus previous (Note that the results reported in the previous literatures were estimated by only a single training set and test set split. This may overestimate the accuracy rate.)

	Proposed method	Other methods
Dataset 1	88.81%	87.8%[22], 86.65% [14]
Dataset 2	85.05%	85.9%[22], 65.2%[11], 73.4%[14]

4. CONCLUSIONS

In this paper, we have presented a scene categorization approach based on the multi-scale category-specific visual word. The multi-scale visual words give us a richer representation of the scene images which represent the scene image from the whole image to consecutive smaller regions of the image. This representation combines the global-feature-based approach and the local-feature-based approach into a uniform framework. Moreover, the category-specific visual word creation strategy is capable of generating visual words with better discriminative abilities than the traditional strategy. We test the proposed method on two datasets with 8 and 13 scene categories respectively which are used in the previous literatures. The experimental results have shown the proposed method is comparable to the methods with the best results published in the previous literatures in terms of classification accuracy rate. And in the proposed method, except for a list of visual words, no other complex intermediate model is trained. Thus, the proposed method has the advantage in terms of simplicity. As it is, our proposed method has not included the spatial correlations between visual words yet. We believe this spatial information would help 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, Fei-Fei Li for provding their data sets and Dr. Clement Pang for discussions.

REFERENCES

[1] Wang, J. Z., Jia L. and Wiederhold G., "SIMPLIcity: semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(9), 947-963 (2001)

[2] Chang, E., Kingshy G., Sychay G. and Gang W., "CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines," *IEEE Transactions on Circuits and Systems for Video Technology* 13(1), 26-38 (2003)

[3] Vailaya, A., Figueiredo M., Jain A. and Zhang H. J., "Content-based hierarchical classification of vacation images," in *IEEE International Conference on Multimedia Computing and Systems* M. Figueiredo, Ed., pp. 518-523 (1999).

[4] Siagian, C. and Itti L., "Gist: A Mobile Robotics Application of Context-Based Vision in Outdoor Environment," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* L. Itti, Ed., pp. 1063-1069 (2005).

[5] Manduchi, R., Castano A., Talukder A. and Matthies L., "Obstacle Detection and Terrain Classification for Autonomous Off-Road Navigation," *Autonomous Robots* 18(1), 81-102 (2005)

[6] Torralba, A., "Contextual priming for object detection," *International Journal of Computer Vision* 53(2), 169-191 (2003)

[7] Torralba, A., Murphy K. P. and Freeman W. T., "Contextual models for object detection using boosted random fields," in *Adv. in Neural Information Processing Systems 17 (NIPS)*, pp. 1401-1408, MIT Press, Lawrence K. Saul and Yair Weiss and Léon Bottou (2005).

[8] Siagian, C. and Itti L., "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(2), 300-312 (2007)

[9] Luo, J. and Savakis A., "Indoor vs outdoor classification of consumer photographs using low-level and semantic features," in *2001 International Conference on Image Processing* A. Savakis, Ed., pp. 745-748 vol.742 (2001).

[10] Vogel, J. and Schiele B., "A Semantic Typicality Measure for Natural Scene Categorization," in *2004 DAGM* pp. 195-203, Springer-Verlag (2004).

[11] Fei-Fei, L. and Perona P., "A Bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 524-531 vol. 522 (2005).

[12] Quelhas, P., Monay F., Odobez J. M., Gatica-Perez D., Tuytelaars T. and Van Gool L., "Modeling scenes with local descriptors and latent aspects," in *Tenth IEEE International Conference on Computer Vision*, pp. 883-890 Vol. 881 (2005).

[13] Bosch, A., Munoz X. and Marti R., "Which is the best way to organize/classify images by content?," *Image and Vision Computing* 25(6), 778-791 (2007)

[14] Bosch, A., Zisserman A. and Munoz X., "Scene classification Via pLSA," in *ECCV 2006*, pp. 517-530 (2006).

[15] Lowe, D. G., "Object recognition from local scale-invariant features," in *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, pp. 1150-1157 vol.1152 (1999).

[16] Fergus, R., Fei-Fei L., Perona P. and Zisserman A., "Learning object categories from Google's image search," in *Tenth IEEE International Conference on Computer Vision* L. Fei-Fei, Ed., pp. 1816-1823 Vol. 1812 (2005).

[17] Agarwal, S., Awan A. and Roth D., "Learning to detect objects in images via a sparse, part-based representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(11), 1475-1490 (2004)

[18] Sivic, J. and Zisserman A., "Video Google: a text retrieval approach to object matching in videos," in *Ninth IEEE International Conference on Computer Vision*, pp. 1470-1477 vol.1472 (2003).

[19] Li, F. and Perona P., "A Bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* P. Perona, Ed., pp. 524-531 vol. 522 (2005).

[20] Nowak, E. and Juries F., "Vehicle categorization: parts for speed and accuracy," in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 277-283 (2005).

[21] Oliva, A. and Torralba A., "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision* 42(3), 145-175 (2001)

[22] Bosch, A., Zisserman A. and Muoz X., "Scene Classification Using a Hybrid Generative/Discriminative Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(4), 712-727 (2008)

[23] Wright, D. B., Ed., [*Understanding Statistics: An Introduction for the Social Sciences*], SAGE (1997).