Feature fusion at the local region using localized maximum-margin learning for scene categorization

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

Abstract

In the field of visual recognition such as scene categorization, representing an image based on the local feature (e.g. the bags of visual words (BOW) model and its variants) has become popular and one of the most successful methods. In this paper, we propose a method that uses localized maximum-margin learning to fuse different types of features during the BOW modeling process for eventual scene classification. Unlike previous feature fusion methods for visual recognition, which combines the features after generating the entire set of representations based on different types of features from local regions, the proposed method fuses different features at the stage when the best visual word is selected to represent a local region (hard assignment) or the probabilities of the candidate visual words used to represent the unknown region are estimated (soft assignment). The merits of the proposed method are that (1) errors caused by the ambiguity of single feature when assigning local regions to the best representative visual words can be corrected or the probabilities of the candidate visual words used to represent the region can be estimated more accurately; and that (2) it offers a more flexible way in fusing these features through determining the similarity-metric locally by localized maximum-margin learning. The proposed method has been evaluated experimentally and the results indicate its effectiveness.

Keywords: Scene categorization, image recognition, feature fusion,

*Corresponding author. Tel. +852 28578414; Fax. +852 25598738

Email addresses: jzhqin@eee.hku.hk (Jianzhao Qin), nyung@eee.hku.hk (Nelson H. C. Yung)

1. Introduction

Scene categorization concerns with automatically labeling or classifying a given image to a specific scene category (e.g., coast, forest, highway, office, kitchen, street, sitting room and etc.). Automatic categorization of an image to a scene can be used to manage picture libraries and retrieve images from Internet or in image databases [1, 2, 3]. Moreover, being able to recognize the scene category of a place is vital for an intelligent vehicle or robot to locate its position and take appropriate actions under different scenes [4, 5]. Furthermore, scene categorization can also provide critical contextual information to many computer vision tasks, such as object recognition, image segmentation and etc. [6, 7]. It is also essential for an intelligent video surveillance system in the future, which can help define what abnormal conditions are for detection and tracking (e.g. abnormal objects, abnormal behaviors). For instance, a person running can be considered as abnormal in a 'street' scene, but normal in a 'sport ground' scene.

In the early research work for scene categorization, many global feature based methods [2, 3, 4, 8] have been proposed. In these methods, an image is taken as a whole, and the distribution(s) of color [2, 3] and/or texture [2] and/or gradients [4, 8] over the entire image region is (are) employed to describe the scene image. They have achieved certain success, especially in separating outdoor scenes from indoor scenes. However, when they are employed to classify scenes that have similar global properties (e.g. bedroom vs. sitting room; open country vs. coast), they often result in poor success rate. In recent years, local semantic feature based methods [9, 10, 11, 12, 13] become more popular because of its robustness towards occlusions, illumination variations and slight geometric deformation. They model a scene image by the co-occurrences of a number of visual components or the co-occurrences of a certain number of visual topics (intermediate representation). One of the most popular and successful models is called 'the bags of visual words' (BOW) [11, 12, 13]. Many variants of this model have been proposed [14, 15, 16, 17, 18]. In [14, 15, 16], latent variables, which can be taken as a group of visual words, are learned using the techniques called probabilistic Latent Semantic Analysis (pLSA) [19] or Latent Dirichlet Allocation [20]. In [17], Lazebnik et al. proposed a spatial pyramid matching method,

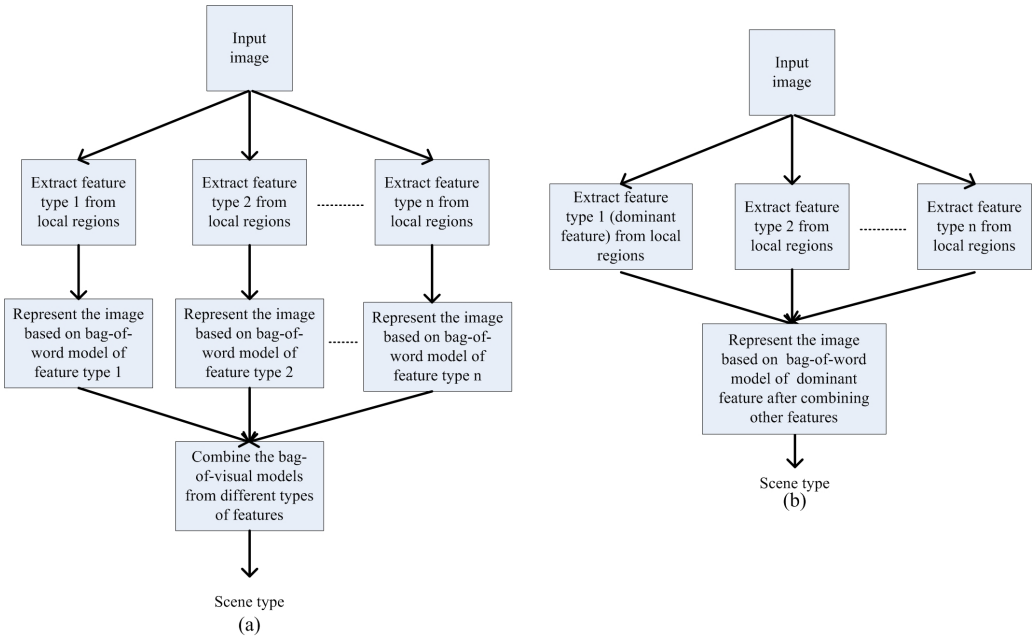

Figure 1: (a) Global feature combination; (b) Local feature combination based on dominant feature.

which matches the distributions of visual words at different spatial resolution between paired images then used it as a similarity measurement.

To further enhance the performance of the BOW based methods, algorithms have been proposed to combine different types of features [21, 22, 23] in the field of object recognition. The methods proposed by Varma et al. [21] and Bosch et al. [23] create several spatial pyramid representations of the BOW model that correspond to different types of features, and then a multiple-kernel learning (MKL) approach is employed to learn the linear weighting of different kernels that correspond to different spatial pyramid representations. Gehler et al. [24] proposed an enhanced multiple kernel method called 'LPboost' (linear programming boosting), which allows the support vector machines (SVM)' parameters trained for different types of features to be different. From the combination procedures of the aforementioned methods, we can see that the combination of features occurs after the image has been represented by the BOW model or its variants. In other words, such feature combination is carried out globally. Figure 1 depicts the

differences between the global feature combination method and the proposed local combination method. One of the weaknesses of the global combination method is that the ambiguity of the local patches caused by the single feature representation would unlikely be resolved by other globally introduced features. This is, because as the other features are globally coded, they do not provide information about a specific local image region. For example, just based on the SIFT feature, a region of an image that represents the grass land of the 'Open country' scene may be incorrectly represented by the visual word that represents a part of the sea water from the 'Coast' scene, which may result in incorrect classification of the image from 'Open country' to 'Coast'. Although, combining BOW model of SIFT and BOW model of color feature may alter the final classification result ('Open country' may have more green regions while 'Coast' may have more blue regions). However, in some cases, the image of 'Open country' may have an equally large region that represents the blue sky, which is similar to sea water in color. Conversely, 'Coast' may show a large region of green trees. As such, 'Open country' scenes and 'Coast' scenes can share very similar BOW model of color feature. The other weakness of the global approach is that they can only give fixed weightings to the considered features. However, in practice, in order to differentiate a region from other regions, we may give more weight to a particular feature. For instance, in order to differentiate the shore in the 'Coast' scene from the grass land in the 'Open country' scene, we can put more weights on the color feature. However, if the target is to differentiate the grass land in the 'Open country' scene from the trees in the 'Forest' scene, gradient and texture features should be given more weighting values instead.

In this paper, we propose a local feature fusion method using localized maximum-margin learning. Given an unknown image region, the dominant feature (we choose SIFT feature as the dominant feature due to its successfulness and popularity in the area of visual recognition [14, 25, 26].) and other features (we choose color feature and (or) local binary pattern (LBP) feature [27]) are extracted from the image region. Based on the dominant feature, we select a set of nearest neighbor visual words based on Euclidean distance measured from the dominant feature. Then, these candidate visual words are considered as classes. The dominant features and other features of the training image regions that form these visual words are taken as the features of the samples belonging to each class. Next, maximum margin learning is employed to search for a combination of the multiple features that result in maximum-

margin separation between each class (Since this maximum-margin learning is performed on the visual words which are closest to the feature extracted from unknown region, we call it localized maximum-margin learning.). This can also be treated as a support vector machine (SVM) classifier training based on the classes formed by the candidate visual words. The trained classifier is used to measure the similarity between features of the unknown image region and the candidate visual words. The classification result determines which visual word best represents the unknown region, or estimates the probabilities of the candidate visual words used to represent the unknown region. Based on the maximum-margin criteria, the classifier adaptively determines different combination strategies for different features. After a list of visual words, a 0-1 binary vector that represents the existence of the visual word is compiled to form the BOW representation of the image and used for classifier training or testing. In order to enhance the generalization ability of the method, we use a soft multiple visual words assignment strategy in which the output of the maximum-margin learning is used to estimate the probabilities of the visual words. Then, a float type vector representing the probabilities of the visual words is compiled to form the BOW representation of the image. We have evaluated the performance of the proposed method based on two widely used scene categorization datasets consisting of 8 scene categories with 2688 images and 15 scene categories with 4485 images respectively, using 10-fold cross-validation. The experimental result shows the superiority of the proposed adaptive feature fusion method to the single feature method and other representative scene categorization methods. And the experimental result also illustrates that the performance of the proposed method is better than (or equivalent to) the MKL based global combination method.

The rest of this paper is organized as follows: Section 2 formulates the problem. Section 3 gives an overview of the whole scene categorization system. Section 4 describes the local feature extraction procedure. Section 5 reviews the category-specific visual words creation method. Section 6 introduces the proposed localized maximum-margin learning method for feature fusion. Section 7 introduces a caching strategy to reduce the computational cost. Section 8 describes the soft multiple visual words assignment method. Section 9 reports the experimental results. And this paper is concluded in Section 10.

2. Problem formulation

The scene categorization problem based on the BOW model representation 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 of a set of visual words $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$. We denote this representation by $\mathrm{R}(\mathbf{I})$, which is a vector $\mathbf{r} = \mathrm{R}(\mathbf{I}),\ \mathbf{r} \in \Re^k$ that indicates the distribution of the visual words, or the presence of the visual words, or the probability of the presence of the visual words. The problem then becomes the issue of finding a projection:

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

which projects the visual words representation of the image to the scene category $c_i,\ i = 1, \cdots, m$ where it belongs.

The target of the local feature fusion method proposed in this paper is to form a better BOW model, $\mathrm{R}(\mathbf{I})$. That is, given a local image region $\mathbf{I}_r$ and a list of visual words $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$ created by a single dominant feature, we search for a projection p from a local image region to a visual word, $p : \mathbf{I}_r \rightarrow \mathbf{v}_i$, which can represent the local image region more accurately by fusing other types of features or estimate the probabilities of the visual words used to represent the local image region. Based on the improved BOW representation model, we expect that the classification accuracy rate can be increased.

3. Overview

Figure 2 depicts the overall framework of the proposed scene classification method. In the training stage, each training image is divided into regular patches at different scales for visual word creation. Then, dominant feature and other features are extracted from each patch. The dominant and other features are combined with the dominant and other features respectively from the coarser scale and neighborhood regions (contextual information) to describe the region of interest (ROI). Based on the dominant features, clustering is performed according to different scales and scene categories to create representative visual words denoted by the means of the clusters, which are subsequently entered into a codebook. Next, each image in the training set is evaluated against the visual word codebook using the proposed localized maximum-margin learning method to combine other features in order to

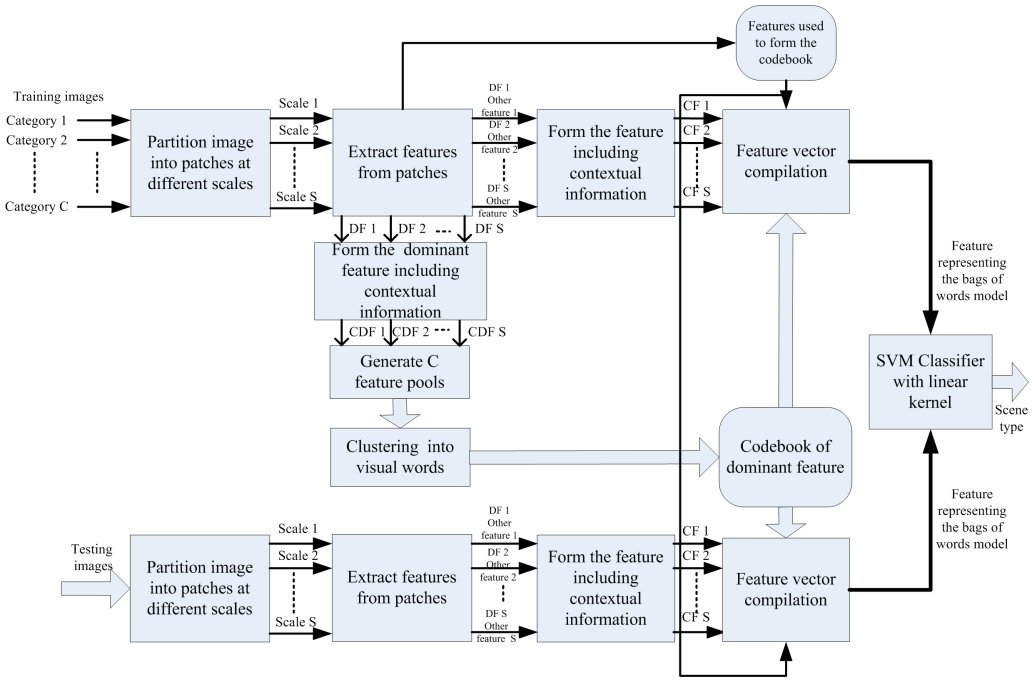

Figure 2: The framework of the proposed method (DF denotes dominant feature, CDF is the dominant feature with contextual information and CF denotes the multiple features with contextual information).

select a list of visual words that best represents the patches of the image (or to estimate the probabilities of the visual words that can be used to represent the patches). This list is further compiled into a feature vector (a 0-1 binary feature vector representing the existence of the visual words or a feature vector recording the probabilities of the visual words which can be used to represent the local regions), which is used for training a Support Vector Machine (SVM) classifier. In the classification of an unknown image, the image is partitioned into patches at different scales and their features calculated. As in training, using the localized maximum margin learning to combine the features, a list of visual words that best represents the local regions of the image is selected (or the probabilities of the visual words used to represent the local regions are calculated) to compile a feature vector. Finally, the feature vector is classified by the SVM to obtain the scene type.

4. Image region feature extraction

In this paper, we extract three types of features from each image region, i.e., Scale-invariant feature transform (SIFT) [25] or and local binary pattern (LBP) [27] features. Among these three types of features, SIFT describes the distribution of the gradients at different orientations, which has been widely used to object recognition and scene categorization, among others. We choose SIFT feature as the dominant feature due to its successfulness and popularity in the area of visual recognition [14, 25, 26]. The color feature provides the color information and the LBP feature represents the texture information using the histogram of uniform local binary patterns. It has been successfully used for problems such as texture classification and face detection. The uniform patterns of the LBP feature represent primitive microstructures, such as edges, corners and spots, which are complementary to the features describing gradient distribution [28].

For SIFT feature extraction, we adopt the method introduced by Lowe [25], i.e., the ROI is divided into 4×4 blocks; then, the histogram of the gradient magnitude at different orientations is calculated for each block; after weighted by a Gaussian function, the histograms of the 16 blocks are concatenated to form a SIFT feature vector. For color feature, we adopt a similar procedure as well. After transforming the image into Lab color space, the ROI is divided into 4×4 blocks. Then, the mean values of the L, a, b components are calculated. Next, these mean values for each block are concatenated together after weighted by a Gaussian function. Similarly, for the

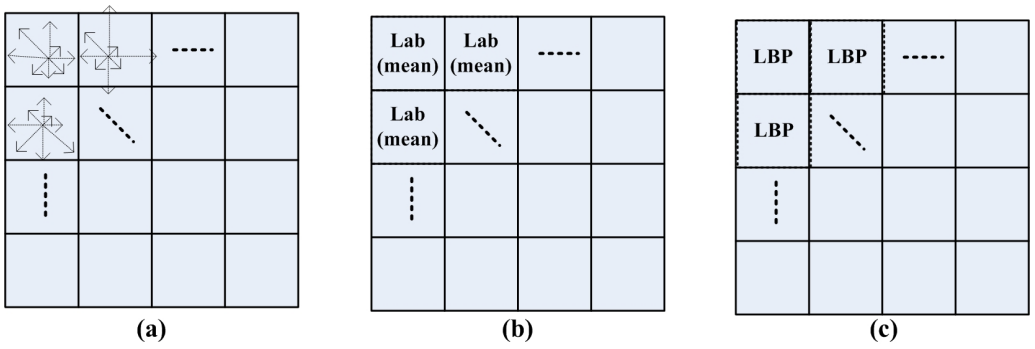

Figure 3: Extraction of (a) SIFT feature; (b) color feature; (c) LBP feature.

LBP feature extraction, the ROI is also divided into 4×4 blocks, and then the histogram of local binary patterns (9 bins correspond to 9 uniform patterns and 1 bin corresponds to the non-uniform patterns) is calculated from each block. After that, the histograms from the blocks are concatenated and weighted by a Gaussian function to form the LBP feature vector. Figure 3 depicts the extraction for these three types of features. In order to capture image information from different scales, the image is regularly divided into patches at different scales from the coarsest scale (i.e. the whole image) to consecutive finer scales. The three image features described above are extracted from all these patches. Meanwhile, the contextual information is also extracted and integrated to describe the ROI. Such contextual information provides useful cue about the ROI [18], and can potentially reduce ambiguity when using visual words to represent local regions. We combine the image feature from the region at coarser scale (but with the same sampling point) and the image features from the neighbor regions at the same scale with the feature of ROI altogether to finally describe the ROI. Figure 4 depicts the sampling points of the image patches at different scales and the regions for context information extraction.

Let $\mathbf{P}_L \in^{m_L \times n_L}$ denotes the ROI, $\mathbf{P}_C \in^{m_C \times n_C}$ denotes the region having the same sampling point as the ROI but at a coarser scale level and $\mathbf{P}_N \in^{m_N \times n_N}$ denotes the neighbor regions of the ROI at the same scale level. For local visual word, the ROI is represented by $\mathbf{f} = f(\mathbf{P}_L)$ where f denotes the feature extraction function. For the contextual visual word, we represent the ROI as $\mathbf{f} = f(\mathbf{P}_L, \mathbf{P}_C, \mathbf{P}_N)$. We linearly combine these features. The

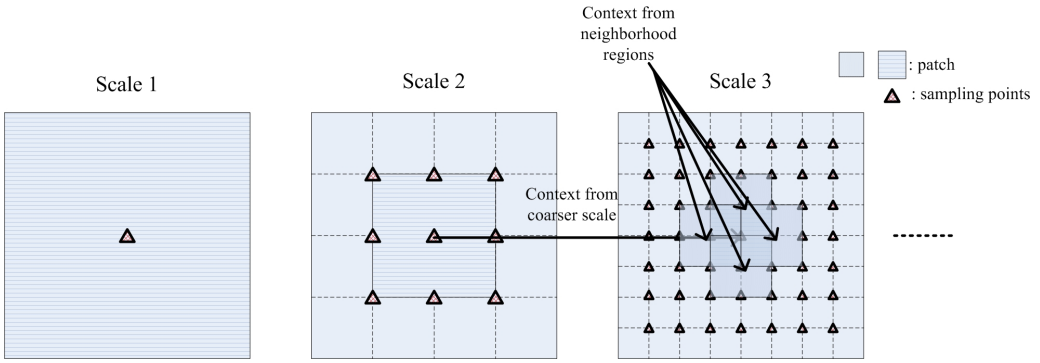

Figure 4: Sampling points of the image patches at different scales and the regions for context information extraction.

feature of the ROI is then represented as:

$$\mathbf{f} = [f(\mathbf{P}_L), w_C \cdot f(\mathbf{P}_C), w_N \cdot f(\mathbf{P}_N)], \tag{2}$$

where w_C and w_N are the weighting parameters that control the significance of features from the coarser scale and the neighborhood regions. The weighting parameters for different contextual information are determined using cross-validation (Our experiments showed that the weighting parameters are usually set around 0.7.).

5. Category specific visual words creation

We train the visual words in a category-specific manner, that is, the features are extracted from a subset of training images (100 images each class) at different scene categories, and then features of different scene categories are clustered by K-means. After that, the visual words created from different scene categories are concatenated to form the codebook. Figure 5 depicts how the category-specific visual words are created. It has been proven that visual words created this way are more discriminative [29]. Since images from different categories may have similar regions, creating the visual words in this way may result in redundancy as well (several visual words represents the image regions with very similar characteristic), which reduces the generalization ability of the codebook. In Section 8, we will introduce a soft-assignment strategy that solves the problem.

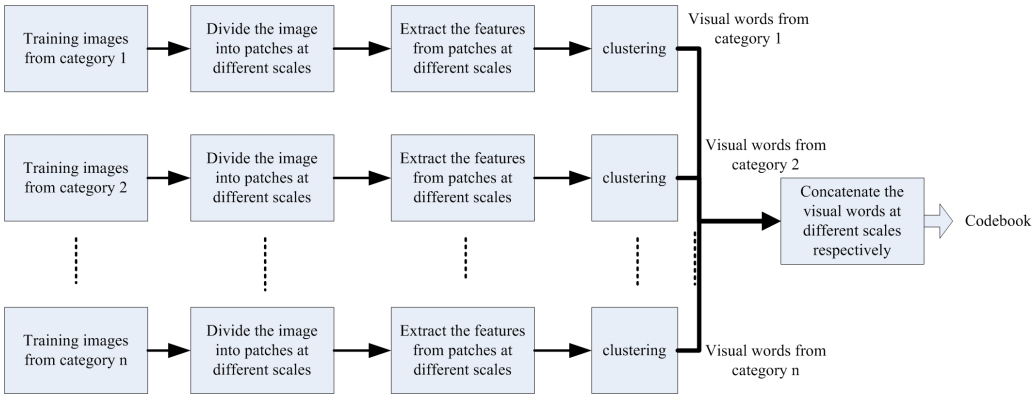

Figure 5: Category-specific visual words creation.

6. Localized maximum-margin learning for feature fusion

Figure 6 depicts the procedure of combining dominant feature with other features using localized maximum-margin learning to select the best visual words. From Section 4, using the dominant feature vector $\mathbf{f}^D$ (this feature is the combination of the local image feature with the contextual feature and transformed by the PCA transformation matrix), we first calculate its similarity with the visual words in the codebook based on the Euclidean distance measurement. Constrained by the ratio to the shortest distance and maximum number, K-nearest neighbor candidate visual words are chosen. Then, the dominant features which were clustered to form these visual words and their corresponding other features are retrieved and concatenated to form a feature vector $\mathbf{f} = [\mathbf{f}^D, \mathbf{f}^O]$, where $\mathbf{f}^O$ denotes other features. Next, maximum-margin learning is employed to learn weighting values of the elements in the feature vector. The 2-class maximum-margin learning problem is equivalent to the 2-norm minimization problem as depicted in Equation (3),

$$\begin{aligned} &\min_{\mathbf{w},\eta} P \sum_{i=1}^{L} \eta_i + \tfrac{1}{2} \|\mathbf{w}\|_2 \\ &s.t. \quad y_i \mathbf{w}^T [\phi(\mathbf{f}_i^T) \;\; 1]^T + \eta_i \geq 1, \quad \eta_i \geq 0, \; i = 1, \ldots, L \end{aligned} \tag{3}$$

where $\mathbf{w}$ can be taken as the weighting vector, η_i is called slack variables which allows to handle non-separable feature vectors , ϕ denotes a linear or nonlinear transform on the feature vector, $\mathbf{f}_i$, L is the total number of training samples, y_i takes two values, i.e. 1 or -1, corresponding to two

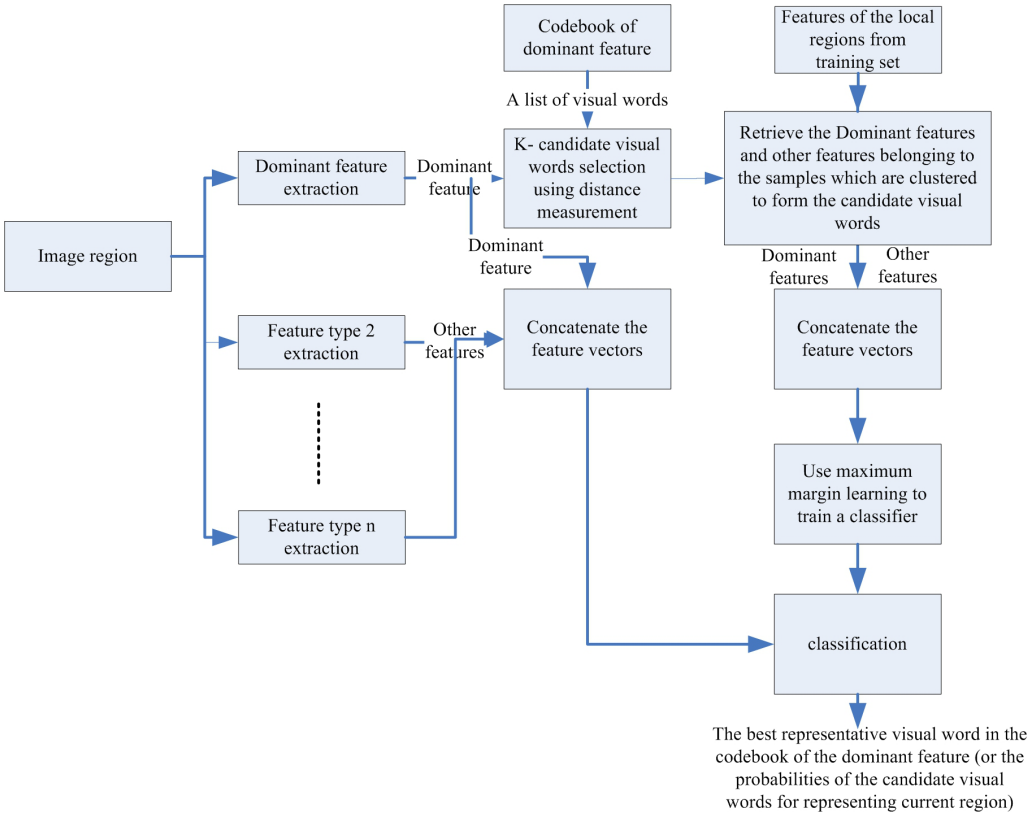

Figure 6: Fusion of features using localized maximum-margin learning.

classes respectively and P is the penalty parameter. The dual form of the optimization problem in equation (3) is:

$$\begin{aligned} & \mathrm{Q}(\alpha) = \sum_{i=1}^{L} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{L} \sum_{i=1}^{L} \alpha_i \alpha_j y_i y_j \phi(\mathbf{f}_i)^T \phi(\mathbf{f}_j) \\ & s.t. \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & \qquad 0 \leq \alpha_i \leq P \qquad \text{for } i = 1, 2, \ldots, L \end{aligned} \tag{4}$$

where $\{\alpha_i\}_{i=1}^{L}$ are the Lagrange multipliers. The inner product, $\phi(\mathbf{f}_i)^T \phi(\mathbf{f}_j)$, of two transformed feature vectors can be effective calculated using a kernel trick [30, 31], $K(\mathbf{f}_i, \mathbf{f}_j) = \phi(\mathbf{f}_i)^T \phi(\mathbf{f}_j)$. When ϕ is linear, the linear kernel,

$$K(\mathbf{f}_i, \mathbf{f}_j) = {\mathbf{f}_i}^T \mathbf{f}_j, \tag{5}$$

is usually used. When ϕ is non-linear, the polynomial kernel,

$$K(\mathbf{f}_i, \mathbf{f}_j) = ({\mathbf{f}_i}^T \mathbf{f}_j + 1)^{pw}, \tag{6}$$

where the power pw is specified a priori by the user, and radial-basis function kernel,

$$K(\mathbf{f}_i, \mathbf{f}_j) = \exp(-\frac{1}{2\delta^2} \|\mathbf{x} - \mathbf{x}_i\|^2), \tag{7}$$

where the width δ is specified a priori by the user, is commonly used (We will discuss the impact of choosing different kernels to the performance in the experimental result part.). The 2-class problem can be easily extended to multiple-class problem using the one against one strategy [32]. Since maximum margin learning is based on the feature vectors that form the K-nearest neighbor candidate visual words, we call it localized maximum-margin learning. The localized learning enables us to find a weighting vector $\mathbf{w}$ which maximizes the distances between the candidate visual words after introducing the other features. After that, the learned weighting value is used to measure the similarity of the feature of the unknown region to the candidate visual words for selecting the best representative visual words for the unknown region. Obviously, this weighting vector $\mathbf{w}$ can change and adapt to different nearest neighbor structures when finding a suitable weighting values, which can maximize the margin between the feature vectors belonging to different visual words. For instance, if the SIFT features of the candidate visual words are very similar, larger weight is given to the color feature in

order to separate the samples belonging to different candidate visual words. In contrast, if the color features are similar, larger weight is given to the SIFT feature. The steps for selecting the best representative visual words by combining the other features using the localized maximum-margin learning are as follows:

Step 1: Given a list of visual words, $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ formed by clustering the dominant features from a subset of training set, and the dominant feature of an image patch, $\mathbf{f}^D$, calculate the Euclidean distances between the feature and the visual words, $\{d_1, d_2, \ldots, d_n\}$.

Step 2: Choose the minimum distance, $d_m = \min\{d_1, d_2, \ldots, d_n\}$, then calculate the ratios of the distance to the minimum distance, $r_i = \frac{d_i}{d_m}, i = 1, 2, \ldots n$.

Step 3: Select the candidate visual words whose distance to $\mathbf{f}^D$ satisfy the ratio $r_j \leq T_r, j = 1, 2, \ldots, K$ ($T_r = 1.15$ *in this paper*) and its corresponding visual words, $\mathbf{v}_j,\ j = 1, 2, \ldots, K$ (in order to reduce the computational burden, the maximum number of the preliminary selected visual words can be upper-bounded by number N_b).

Step 4: Retrieve the dominant features $\{\mathbf{f}^D_{1j}, \mathbf{f}^D_{2j}, \ldots, \mathbf{f}^D_{n_j j}\},\ j = 1, 2, \ldots, K$ and the corresponding other features $\{\mathbf{f}^O_{1j}, \mathbf{f}^O_{2j}, \ldots, \mathbf{f}^O_{n_j j}\},\ j = 1, 2, \ldots, K$ ($\mathbf{f}^O_{ij}$ denotes the concatenated feature vectors of other features, i.e. $\mathbf{f}^O_{ij} = [\mathbf{f}^{O_1}_{ij}; \mathbf{f}^{O_2}_{ij}; \cdots \mathbf{f}^{O_m}_{ij}]$, n_j is the number of features of visual word j and m is the number of types of other features.) from the training set that are clustered to form the candidate visual words, $\mathbf{v}_j,\ j = 1, 2, \ldots, K$. Then, normalize these features by their l_2 norms, i.e., $\mathbf{f}^D_{ij} = \frac{\mathbf{f}^D_{ij}}{\left\|\mathbf{f}^D_{ij}\right\|_2}, \mathbf{f}^{O_1}_{ij} = \frac{\mathbf{f}^{O_1}_{ij}}{\left\|\mathbf{f}^{O_1}_{ij}\right\|_2}, \mathbf{f}^{O_2}_{ij} = \frac{\mathbf{f}^{O_2}_{ij}}{\left\|\mathbf{f}^{O_2}_{ij}\right\|_2}, \cdots, \mathbf{f}^{O_m}_{ij} = \frac{\mathbf{f}^{O_m}_{ij}}{\left\|\mathbf{f}^{O_2}_{ij}\right\|_2}$ (If linear kernel is used, no normalization can give better performance when soft multiple visual words assignment is used.). After that, concatenate them to form the feature vector for learning, $\mathbf{f}_{ij} = [\mathbf{f}^D_{ij}; \mathbf{f}^{O_1}_{ij}; \mathbf{f}^{O_2}_{ij}; \cdots \mathbf{f}^{O_m}_{ij}]$.

Step 5: Take the retrieved features $\{\mathbf{f}_{1j}, \mathbf{f}_{2j}, \ldots, \mathbf{f}_{n_j j}\},\ j = 1, 2, \ldots, K$ as the training set of K classes. Then, train a classifier based on Equation (3) .

Step 6: Classify the given patch feature $\mathbf{f}$ by the trained classifier to a class c. Then, the image patch is represented by the visual word corresponding to class c.

In terms of probability, the best visual words selection based on feature fusion can be formulated as $\mathbf{v}_{best}^{D} = \max_{\mathbf{v}_j} p(\mathbf{v}_j^D|\mathbf{f}^D, \mathbf{f}^O)$ while the best visual words selection based on single dominant feature is formulated as $\mathbf{v}_{best}^{D} = \max_{\mathbf{v}_j} p(\mathbf{v}_j^D|\mathbf{f}^D)$. From the best representative visual words from each patch, a bag of visual words model BOW_D, which is denoted as a feature vector (the compilation of BOW_D will be discussed in Section 8) is produced. Based on BOW_D of the images from the training set, a linear SVM classifier is trained for classifying the test images.

7. Caching strategy for reducing the computational cost

In the proposed method, localized maximum-margin learning is used to find the best representative visual word for each patch in the image. Solving the optimization problem of Equation (3) largely increases the computational cost comparing with using the Euclidean distance. Assume an image has been divided into N patches and each patch has K candidate visual words on average. The computational complexity for maximum-margin learning is $\mathrm{O}(i(K-1)KPd)$, where i is the iteration number, P is the number of training samples and d is the dimension of the feature vector. Thus, the computational complexity for the localized maximum-margin learning for an image is $\mathrm{O}(i(K-1)KPdN)$.

In order to reduce the computational cost for the localized maximum margin learning, we propose a caching strategy for training and testing. Since, the localized maximum learning for each patch can be taken as a multiple-class SVM model training process and 1 vs. 1 strategy is used for the multiple-class training, the problem can be decomposed into $K(K-1)/2$ paired SVM model training. And since these training images are divided into local patches, similar patches must co-exist within an image and among these images and are represented by the same visual word. For these similar patches, they must have similar candidate visual words when Euclidean distance measurement is employed to select the candidate visual words. If we follow the above steps to perform the training, it will result in large redundant calculations. This redundancy inspires us to employ the caching strategy to reduce the computational cost. The caching strategy is as follows:

Given a cache region, M, which stores the paired SVM models that have been trained. These models are indexed by the paired visual words indices, (i, j).

Step 1: For a given local patch, K candidate visual words $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_K\}$ are selected by the Euclidean distance measurement.

Step 2: Before conducting the paired model training, examine whether the indexes of the two visual words exist in the cache. If there is matched index in the cache, and then the model parameters can be directly read from the cache. Otherwise, paired maximum-margin learning is performed and the corresponding parameters are stored in the cache.

In the extreme case, the cache stores $n(n-1)/2$ models, where n is the total number of visual words. In fact, since the candidate visual words for each local patch share similar characteristics, paired maximum-margin learning usually occurs among similar visual words, which means that, in most cases, paired maximum-margin learning between the visual words with different characteristics would not be trained and stored. This significantly reduces the number of paired-models for training and storing.

8. Soft multiple visual words assignment

In the previous works using BOW model, a given image is encoded as the distribution of the visual words [14, 16, 17], or 0-1 binary coding [18], which denotes whether the visual word exists in the image. As mentioned in Section 5, the category-specific visual words creation strategy is able to increase the discriminative ability. But the side effect is that the redundancy in visual words may reduce its generalization ability. Figure 7 depicts a list of visual words that are created from the training images from scene category 1 and scene category 2. In this figure, we assume visual words V3 and V8 represents the regions with similar characteristic, e.g. the sky region that exists in both scene category 1 and scene category 2. When these visual words are used to encode the image from category 1, it may be encoded as Case 1 (1010100000) or Case 2 (1000100100) (Black square denotes this visual word is used to represent a region in this image and marked as 1), meaning that the similarity between images from the same category is reduced.

In order to increase the generalization ability of the method, a soft-assignment strategy is proposed. That is, the probabilities of the visual

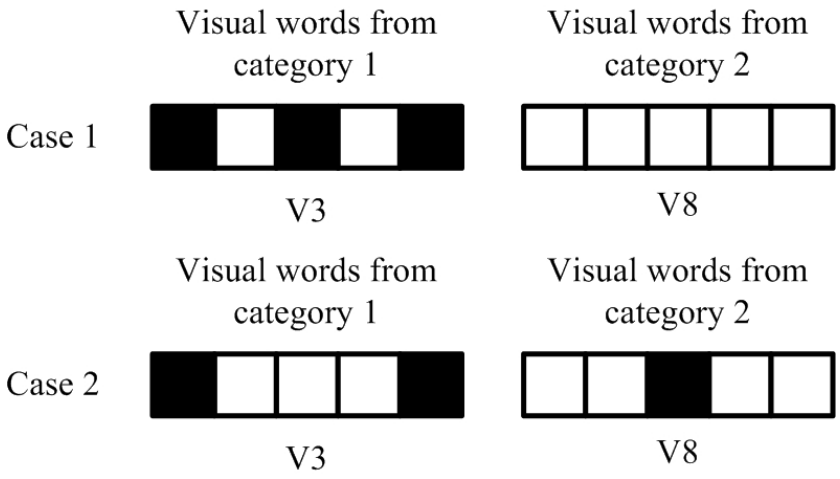

Figure 7: A toy example of the 0-1 binary coding.

words used to represent the image regions are calculated instead of 0-1 coding. The i^{th} element of the vector used to represent the image is set as $\max p(\mathbf{v}_i|\mathbf{f}_j), j = 1, \cdots, N; i = 1, \cdots, M$, where M is the number of visual words and N is the number of the patches the image has been divided into. By doing this, the visual vectors V3 and V8 mentioned in the above example have similar probabilities, which result in the distance between the representation vectors for the images from the same category in Case 1 (It may be encoded as [1.0, 0.0, 0.51, 0.0, 1.0, 0.0, 0.0, 0.49, 0.0, 0.0].) and Case 2 (It may be encoded as [1.0, 0.0, 0.49, 0.0, 1.0, 0.0, 0.0, 0.51, 0.0, 0.0].) becomes shorter. Meanwhile, since the visual words are created in a category-specific manner, the visual words used to represent the image regions, which are unique to a specific scene category, shall have higher probability than the visual words created from other scene categories. This can preserve the discriminative ability of the method. The probability of the visual word given the feature of the image region can be directly derived from localized multi-class maximum-margin learning. In multi-class learning, the 1 versus 1 voting strategy is employed. Thus, we can get the number of votes each candidate visual word has, and then the number of votes is used for probability estimation. Assume that we assign the probabilities to H visual words that have the most number of votes, $vote_1, vote_2, \cdots, vote_H$. The probability

for the visual word corresponding to $vote_j$ is calculated as:

$$p_j = \frac{vote_j}{\sum_{i=1}^{H} vote_i}. \tag{8}$$

9. Experimental results

In this section, the performance of the proposed scene classification method is evaluated on two scene datasets, which have been widely used in previous research [14, 16, 17, 33, 34].

9.1. Datasets and Evaluation Method

SCENE-8 Dataset: 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). The average size of each image is 256×256. The color version of these images has been used. For this dataset, we fuse SIFT feature with color feature; and SIFT feature with color and LBP feature.

SCENE-15 Dataset: consists of 4485 images from 15 categories: bedroom (216 samples), suburb (241 samples), industrial (311 samples), kitchen (210 samples), living room (289 samples), coast (360 samples), forest (328 samples), highway (260 samples), inside city (308 samples), mountain (374 samples), open country (410 samples), street (292 samples), tall building (356 samples), office (215 samples), store (315 samples). This dataset is an extension of Dataset 1. Several indoor scene types (e.g., office, bedroom, living room and kitchen) have been included here. Since portion of this dataset is composed of gray images, the gray version of the images in this dataset are used for evaluation; and the local binary pattern feature are extracted and fused with SIFT feature. Figure 8 depicts the image samples from this dataset. We can observe the variety in the content, scale and view angle of the images. For this dataset, as only gray version of the images is available, we only consider to fuse SIFT feature with LBP feature.

In the experiment, we perform a 10-fold cross-validation to achieve more accurate performance estimation. The average of the accuracy rates across

Figure 8: Image samples from SCENE-15 dataset.

the 10-fold cross-validation is used for comparison. Apart from the experiment that combines several scale levels, we also perform experiments at each scale level separately in order to investigate how scale level affects the performance of the proposed method.

9.2. The influence of scale level to the performance of the proposed method.

Table 1 presents the 10-fold cross-validation results from scale level 1 to 4 using the proposed method to combine SIFT and color features and using SIFT feature only (0-1 binary coding) for SCENE-8 dataset. These results show that the average accuracy results are improved by 5.18%, 4.91%, 2.70% and 0.62% at scale 1, 2, 3 and 4 respectively by fusing SIFT and color features. The results also reveal that the proposed method is more effective at coarser scales. At finer scales, since the image is divided into smaller regions and each image consists of larger number of image patches, even some of these patches are correctly represented by the suitable visual words using the proposed method, the influence of this correction to the final recognition rate becomes smaller. We also obtained result at scale 5, but no obvious improvement is achieved.

Table 1: Accuracy rates (mean (standard deviation)%) of the proposed method to combine SIFT and color features comparing with using SIFT feature only at scale 1, 2, 3 and 4 respectively at SCENE-8 dataset (0-1 binary coding).

	Scale 1	Scale 2	Scale 3	Scale 4
SIFT only	66.63(4.54)	74.21(3.65)	82.68(3.00)	88.15(3.03)
SIFT+color	**71.81(4.16)**	**79.12(3.09)**	**85.38(3.61)**	**88.87(2.80)**

9.3. Results after combining several scale levels and comparison with global combination methods

Table 2 depicts the result of the proposed method after combining visual words belonging to different scale levels (The number of scale levels for combination is determined by cross-validation in the training set. Five scale levels are combined in this paper), the result that uses multiple-kernel learning (MKL) to combine SIFT and color BOW models for SCENE-8 dataset (SIFT and LBP BOW model for SCENE-15 dataset) and the results of using SIFT feature only. We used the source code provided by Bach et al. [35] to do the multiple-kernel learning. We also illustrate the results when the proposed localized maximum-margin learning is used to select the best representative visual word and create the 0-1 coding based BOW feature vector. For SCENE-8 dataset, there are 2688 images for testing in total during the 10-fold cross-validation. We can observe that comparing with using SIFT feature only, the proposed method improves the average accuracy rate by 2.19%, which means that about 59 images misclassified by the SIFT feature only method are correctly classified by the proposed method. Comparing with MKL-based global combination method, the proposed method improves the average accuracy rate by 1.92%, which means that about 52 images misclassified by the global combination method are correctly classified by the proposed method. For SCENE-15 dataset, there are total 4485 images for testing. Comparing with using SIFT feature only, the proposed method improves the average accuracy rate by 2.23%, which means that about 100 images misclassified by the SIFT feature only method are correctly classified by the proposed method. Comparing with using MKL-based global combination method, the proposed method improves the average accuracy rate by 1.55%, which means that about 70 images misclassified by the global combination based method are correctly classified by the proposed method. We also can observe from Table 2 that comparing with 0-1 coding, the proposed method in which soft

Table 2: Accuracy rates (mean (standard deviation)%) of the proposed method (SIFT+color features for SCENE-8 dataset; SIFT+LBP features for SCENE-15 dataset) and the global combination method based on MKL and the method using SIFT feature only.

	Proposed method	Proposed method but with 0-1 coding	Global combination based on MKL	SIFT feature only
SCENE-8 (SIFT+color)	**92.49** (1.99)	**91.35 (2.26)**	90.57(2.44)	90.30 (2.54)
SCENE-15 (SIFT+LBP)	**87.39 (2.48)**	**86.01 (2.67)**	85.84(2.90)	85.16 (1.62)

multiple visual words assignment are employed improves the accuracy rate by 1.14% and 1.38% for SCENE-8 dataset and SCENE-15 dataset respectively. Finally, for SCENE-8 dataset, we further combine the LBP feature with SIFT and color feature using the proposed method with linear kernel. The average accuracy rate is 91.90 (2.31)%, which is not as good as the result when SIFT and color feature are combined but still improved the accuracy rate by 1.60% comparing with using SIFT feature only.

Figure 9 shows the samples of correctly classified images. Let us look at some details about how the proposed method corrects the errors in local region. Figure 10 depicts an 'Open country' image which is wrongly classified as 'Coast' using SIFT feature only but correctly classified using the proposed method. Figure 10(b) depicts the patch samples that form the visual word selected to represent the red regions in the given image using SIFT feature only and using the proposed method to fuse color information respectively. Some regions of the grass land of the given image are wrongly represented by the visual word which represents the region of sea water or sand of 'Coast' due to the similarity in the SIFT feature without incorporating color. However, after using the proposed method to combine color locally, these grass land regions are correctly coded by the visual word which represents the grass land. Figure 11 depicts a 'Mountain' Image which is wrongly classified as 'Forest' without using the proposed method but correctly classified as 'Mountain' after using the proposed method to combine color information. In the previous case, the lack of color information resulted in some parts of the mountain being wrongly coded by the visual words that represent parts

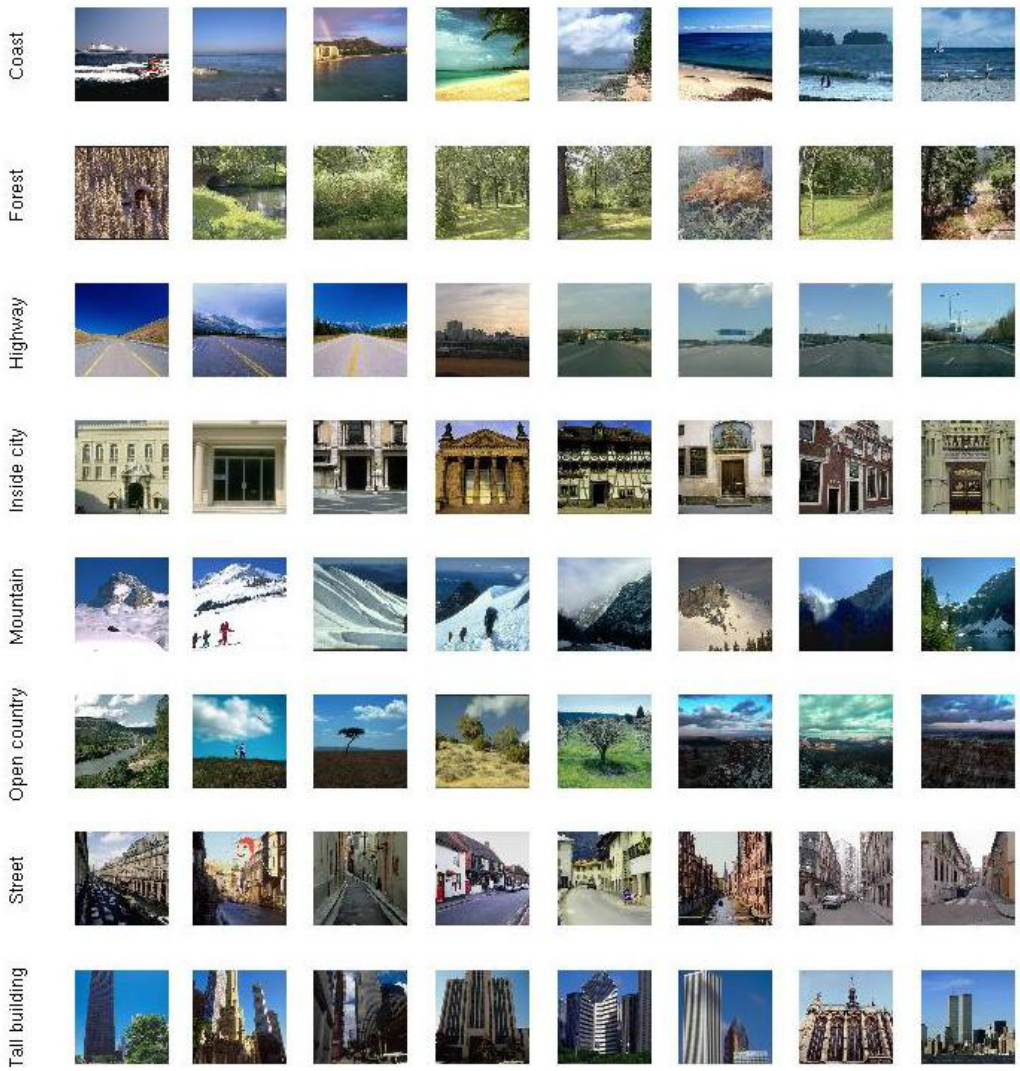

Figure 9: : Samples of correctly classified images.

(a) (b)

Figure 10: : (a) An 'Open country' Image which is wrongly classified as 'Coast' without using the proposed method to utilize color information but correctly classified as 'Open country' after using the proposed method to combine color information. (b) The 1st row shows the patch samples that form the visual word selected to represent the left red rectangle region in (a) (using SIFT only). The 2nd row shows the patch samples that form the visual word selected to represent the right red rectangle region in (a) (using SIFT and color).

Table 3: Accuracy rates (mean (standard deviation)%) of employing different types of kernels during the localized maximum-margin learning for SCENE-8 dataset.

Linear kernel	Polynomial kernel	RBF kernel
92.49 (1.99)	91.83 (2.11)	92.12 (1.82)

of the forest. This error is successfully correctly by the proposed method.

9.4. The impact of different types of kernels during the localized maximum-margin learning

Table 3 depicts the results after employing different types of kernels during the localized maximum-margin learning for SCENE-8 dataset. The linear kernel is defined as equation (5), the polynomial kernel is defined as equation (6) (we set the power $pw = 2.$) and the RBF kernel is defined as equation (7) (the parameter δ is estimated using the average distance between the samples for learning). From Table 3, we can see that the linear kernel has the best performance among three types of kernels.

(a) (b)

Figure 11: : (a) A 'Mountain' Image which is wrongly classified as 'Forest' without using the proposed method to utilize color information but correctly classified as 'Mountain' after using the proposed method to combine color information. (b) The 1st row shows the patch samples that form the visual word selected to represent the left red rectangle region in (a). The 2nd row shows the patch samples that form the visual word selected to represent the right red rectangle region in (a). The third row shows the patch samples that form the visual word selected to represent the left red rectangle region in (a) (using the proposed method to fuse color feature). The forth row shows the patch samples that form the visual word selected to represent the right red rectangle region in (a) (using the proposed method to fuse color feature).

9.5. Error analysis

Figure 12 depicts images misclassified by the proposed method. The first row of Figure 12 shows the errors caused by incorrect labeling in the original dataset; the second row of Figure 12 shows the errors caused by the ambiguity of image content; and the third row of Figure 12 shows the errors caused by the weakness of the proposed method. Figure 12(a) depicts an image which was labeled as 'Highway' in the dataset but classified as 'Street' by the proposed method. In our opinion, this image is more likely to be 'Street' instead of 'Highway'. Because, these is only one lane on the road and there is a pedestrian on the road which is very unusual for a 'Highway' scene. In Figure 12(b), an image which may belong to 'Forest' is mislabeled as 'Coast' in the dataset. The proposed method correctly classified it as 'Forest'. In Figure 12(c), an image from 'Street' is mislabeled as 'Highway'. Because, the region for the pedestrians to walk across the road should not exist in a highway scene but this region is very common in a street scene. Using the proposed method, this image is classified as a 'Street' which we think is more appropriate. Figure 12(d) pictures an image which is labeled as 'Open country' in the dataset while is classified as 'Coast' by the proposed method. This is not exactly wrong as the image does include part of the coast on the left hand side. Thus, it may be difficult to determine whether it is an 'Open country' or 'Coast'. Figure 12(e) is an image labeled as 'Forest' in the dataset but it is classified as 'Mountain' by the proposed method. However, the upper part of this image clearly shows the mountain in the distance. Figure 12(f) is an image labeled as 'Open country' in the dataset but it is classified as 'Mountain' by the proposed method. Upon close inspection, it is found that the image does have a mountain in the background. Figure 12(g) depicts a 'Coast' image being misclassified as 'Open country'. The fogging weather condition makes the surface of the sea looks like the surface of the plain in 'Open country'. Without the bridge in the background, it would be difficult to determine the category of this image. Figure 12(h) depicts a 'Mountain' image being misclassified as 'Open country'. Looking at the surface of the mountain in this image, it contains grassland and shrubs which are also quite common in 'Open country' scenes. In order to differentiate this 'Mountain' image from 'Open country' image, the depth and shape information that describe the rising trend of the land may have to be utilized. Figure 12(i) depicts a 'Forest' image being misclassified as 'Mountain'. The back lighting of the sun caused the trees to appear much darker than their natural color, making them resemble closer to the shrubs common in 'Mountain' scenes

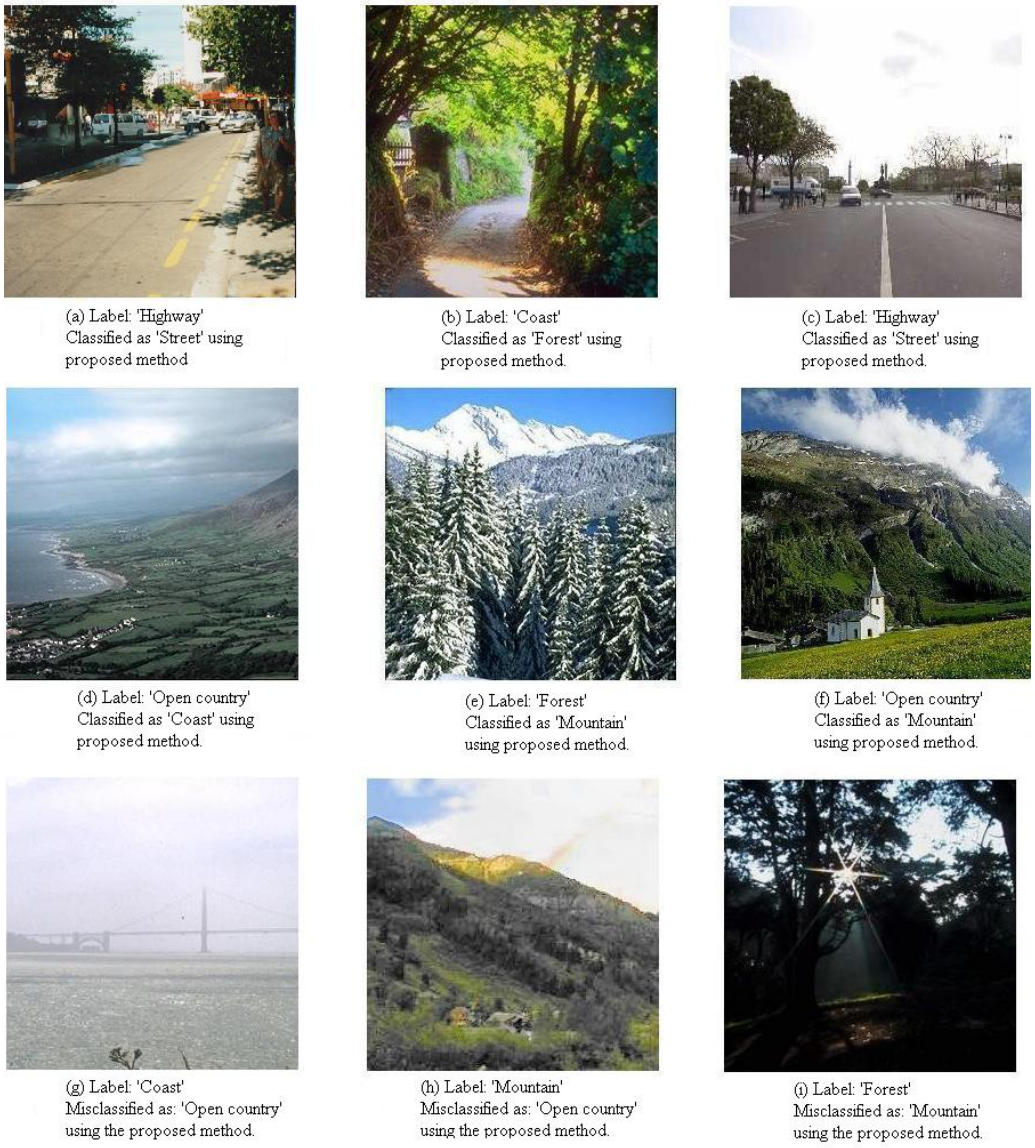

Figure 12: Samples of misclassified images.

than 'Forest'.

9.6. Experimental result of the Caching strategy

In our experiment, after 300 images have been trained, more than 80% paired models can be directly read from the cache in which no further optimization is needed. For testing images in our experiment, more than 95% paired model can be directly read from the cache, i.e., the computational cost is reduced by more than 95% comparing with no caching strategy.

9.7. Comparison with other representative scene classification methods

Finally, the proposed method is compared with other representative scene classification methods, i.e., the 'gist' feature based method [4, 8], the pLSA model [16] based method, the method based on spatial pyramid matching (SPM) [17] and the method based on the mid-level feature method [36]. Since the experiments for the comparison of different scene methods in [14, 33] are solely based on a single training set and testing set split, we believe the performance estimation is not as reliable as the estimation based on 10-fold cross validation. Thus, the same 10-fold cross-validation is employed to estimate the performance of the 'gist', pLSA and SPM based methods. For the mid-level feature based method [36] which has reported the result of the average accuracy rate based on 10 times random training and test set split, we just report their result. The implementation of the 'gist' feature method, pLSA method and the SPM method for comparison is as follows:

'Gist': We implemented the 'gist' feature based on the code provided by Oliva and Torralba (`http://people.csail.mit.edu/torralba/code/spatialenvelope/`). Using the parameters in [8], we used four scale levels (i.e., 1:256, 1:128, 1:64, 1:32) and four orientations (i.e., 0, 45, 90, 135^o) for the 'gist' feature. The SVM classifier with linear kernel was used for classification.

pLSA: The implementation of the pLSA model based method is based on the pLSA code included in the International Conference on Computer Vision) ICCV 2005 short course Recognizing and Learning Object Categories (`http://people.csail.mit.edu/torralba/shortCourseRLOC/index.html`) by Fei-Fei Li, Rob Fergus and Antonio Torralba. We extract dense SIFT feature from patches with scales 8, 16, 24 and 32 pixels (corresponding to support regions with radius 4, 8, 12 and 16

Table 4: Comparison with other representative algorithms (in average accuracy rate (standard deviation)%). (1) proposed method (SIFT+color for SCENE-8 dataset, SIFT+LBP for SCENE-15 dataset); (2) spatial pyramid matching with SIFT feature only; (3) probability latent semantic analysis (pLSA, color-SIFT); (4) gist feature including color information; (5) middle-level feature (SIFT only)

	(1)	(2)	(3)	(4)	(5)
SCENE-8	**92.49 (1.99)**	88.19 (3.46)	84.78 (1.93)(color) 78.80 (3.29) (gray)	80.48 (3.94)	-
SCENE-15	**87.39 (2.48)**	83.30 (1.62)	53.81 (1.36)	67.85 (2.89)	85.6(0.2)

pixels in [16]) at the sampling points with 10 pixels interval. As in [16], we set the size of vocabulary $V = 1500$ (30 images from each category (totally 240 images on SCENE-8 dataset, and 450 images on SCENE-15 dataset) were randomly selected from training set to create the vocabulary), the number of topics $Z = 25$ (100 images from each category (totally 800 images on SCENE-8 dataset and 1500 images on SCENE-15 dataset) were randomly selected from training set to create the topics) and the number of neighbors $K = 10$ for nearest neighbor classifier.

SPM: The implementation of the spatial pyramid matching based method is based on the LIBPMK toolkit [37]. The parameters setting is the same as the settings in [17]. The SIFT features are extracted from 16×16 patches with spacing of 8 pixels. The size of vocabulary is $M = 200$ and the levels of spatial pyramid take $L = 0, 1, 2, 3$.

The comparison results are given in Table 4, from which we can observe the superiority of the proposed method to the methods using single features and the methods which combine multiple features in both datasets.

10. Conclusion

In this paper, we have proposed a localized maximum-margin learning method to fuse multiple features locally in the bag-of-visual-word model forming procedure. After selecting K-nearest neighbor visual words based on dominant feature using Euclidean distance measurement, the other features of the samples which form the K-nearest neighbor visual words are

retrieved to train a classifier using maximum-margin learning. Then, the classification result of the feature of a given image region is used to select the best representative visual words (or to estimate the probabilities of the visual words used to represent the local image region). Comparing with the global feature combination method, the virtue of the proposed method is that it is capable of determining different fusion strategy according to different local feature property. The experimental results show the effectiveness of the proposed method. Meanwhile, in order to reduce the computation cost of the proposed method, we have also introduced a caching strategy that stores parameters between visual words after maximum-margin learning, so that, if the calculation between these visual words is needed again, these parameters can be retrieved directly without having to solve the quadratic programming problem again. In future, we will consider introducing depth and shape information to the method in order to further enhance its performance.

Acknowledgment

The authors would like to thank Antonio Torralba, Fei-Fei Li, Rob Fergus and Lazebnik for providing their data sets, and the 'gist' feature and pLSA source codes. We also would like to thank Francis Bach for providing the multiple kernel learning source code. We also are very grateful to John J. Lee for providing the spatial pyramid matching toolkit and Kwan Wing Keung from the computer center of the University of Hong Kong for providing high performance computing support.

References

[1] J. Z. Wang, L. Jia, G. Wiederhold, Simplicity: semantics-sensitive integrated matching for picture libraries, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (9) (2001) 947–963, 0162-8828.

[2] E. Chang, G. Kingshy, G. Sychay, W. Gang, Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines, IEEE Transactions on Circuits and Systems for Video Technology 13 (1) (2003) 26–38, 1558-2205.

[3] A. Vailaya, M. Figueiredo, A. Jain, H. J. Zhang, Content-based hierarchical classification of vacation images, in: M. Figueiredo (Ed.),

IEEE International Conference on Multimedia Computing and Systems, Vol. 1, 1999, pp. 518–523.

[4] C. Siagian, L. Itti, Gist: A mobile robotics application of context-based vision in outdoor environment, in: L. Itti (Ed.), 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 3, 2005, pp. 1063–1069.

[5] R. Manduchi, A. Castano, A. Talukder, L. Matthies, Obstacle detection and terrain classification for autonomous off-road navigation, Autonomous Robots 18 (1) (2005) 81–102, 10.1023/B:AURO.0000047286.62481.1d.

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

[7] A. Torralba, K. P. Murphy, W. T. Freeman, Contextual models for object detection using boosted random fields, in: Adv. in Neural Information Processing Systems 17 (NIPS), MIT Press, 2005, pp. 1401–1408.

[8] C. Siagian, L. Itti, Rapid biologically-inspired scene classification using features shared with visual attention, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2) (2007) 300–312, 0162-8828.

[9] J. Luo, A. Savakis, Indoor vs outdoor classification of consumer photographs using low-level and semantic features, in: A. Savakis (Ed.), 2001 International Conference on Image Processing, Vol. 2, 2001, pp. 745–748.

[10] J. Vogel, B. Schiele, A semantic typicality measure for natural scene categorization, in: 2004 DAGM, Springer-Verlag, 2004, pp. 195–203.

[11] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from google's image search, in: L. Fei-Fei (Ed.), Tenth IEEE International Conference on Computer Vision, Vol. 2, 2005, pp. 1816–1823.

[12] S. Agarwal, A. Awan, D. Roth, Learning to detect objects in images via a sparse, part-based representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (11) (2004) 1475–1490, 0162-8828.

[13] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Ninth IEEE International Conference on Computer Vision, Vol. 2, 2003, pp. 1470–1477.

[14] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 2005, pp. 524–531.

[15] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuytelaars, L. Van Gool, Modeling scenes with local descriptors and latent aspects, in: Tenth IEEE International Conference on Computer Vision, Vol. 1, 2005, pp. 883–890.

[16] A. Bosch, A. Zisserman, X. Munoz, Scene classification via plsa, in: ECCV 2006, pp. 517–530.

[17] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, Vol. 2, 2006, pp. 2169–2178.

[18] J. Qin, N. H. C. Yung, Scene categorization via contextual visual words, Pattern Recognition 43 (5) (2010) 1874–1888.

[19] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Machine Learning 42 (1) (2001) 177–196.

[20] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022, 944937.

[21] M. Varma, D. Ray, Learning the discriminative power-invariance trade-off, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1–8.

[22] E. Horster, R. Lienhart, Fusing local image descriptors for large-scale image retrieval, in: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp. 1–8.

[23] A. Bosch, A. Zisserman, X. Munoz, Image classification using rois and multiple kernel learning, IJCV 2008.

[24] P. Gehler, S. Nowozin, On feature combination for multiclass object classification, in: Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 221–228.

[25] D. G. Lowe, Object recognition from local scale-invariant features, in: The Proceedings of the Seventh IEEE International Conference on Computer Vision, Vol. 2, 1999, pp. 1150–1157.

[26] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, Pattern Analysis and Machine Intelligence, IEEE Transactions on 27 (10) (2005) 1615–1630.

[27] T. Ojala, Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[28] X. Wang, T. X. Han, S. Yan, An hog-lbp human detector with partial occlusion handling, in: Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 32–39.

[29] J. Qin, N. H. C. Yung, Scene categorization with multi-scale category-specific visual words, Optical Engineering.

[30] S. Haykin, Neural Networks A comprehensive Foundation (Second edition), Prentice Hall, 1998.

[31] V. Vapnik, The Nature of Statistical Learning Theory., Springer-Verlab, 1995.

[32] C.-c. Chang, C.-j. Lin, Libsvm: a library for support vector machines, citeSeerX - Scientific Literature Digital Library and Search Engine [http://citeseerx.ist.psu.edu/oai2] (United States) ER (2001).

[33] A. Bosch, A. Zisserman, X. Muoz, Scene classification using a hybrid generative/discriminative approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (4) (2008) 712–727.

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

[35] F. R. Bach, G. R. G. Lanckriet, M. I. Jordan, Multiple kernel learning, conic duality, and the smo algorithm (2004).

[36] Y. L. Boureau, F. Bach, Y. LeCun, J. Ponce, Learning mid-level features for recognition, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, 2010, pp. 2559–2566.

[37] J. J. Lee, Libpmk: A pyramid match toolkit, Tech. Rep. MIT-CSAIL-TR-2008-17, MIT Computer Science and Artificial Intelligence Laboratory (2008).