Multimed Tools Appl
DOI 10.1007/s11042-013-1496-7

A novel specific image scenes detection method

Yuxiang Xie · Xiao-Ping Zhang · Xidao Luan · Li Liu · Xin Zhang

Abstract Automatic image scene detection is a crucial step for various tasks in computer vision. Current scene detection methods are often computationally expensive for use in real-time image classification. In this paper, a novel and efficient scene detection method based on local invariant features is presented. First, the SIFT feature detector and descriptor has been utilized to extract local image features since the SIFT descriptor has been proved to be an excellent local method that yields high quality features. However, the SIFT descriptor has been shown to produce high dimensional and redundant local features, which can create processing difficulty and computational burden in the successive classification stage. Therefore, two new feature selection strategies are proposed to reduce the number of SIFT keypoints and hence reduce the computational complexity. In both strategies, each image is represented by a single feature vector which assures the efficiency. Finally, a multi-classifier based on a support vector machine is applied to perform the scene detection task. Experimental results show that the proposed method can achieve accurate satisfactory classification results with significantly reduced computational complexity.

Keywords Scene detection · Feature extraction · Local invariant feature · SIFT

1 Introduction

With the rapid development of multimedia and Internet technology, there is an exponentially increasing amount of image data. Since scene is an important semantic feature for nature images, image scene detection becomes important for the labeling and semantic retrieval of images. In this paper, we investigate scene detection for fast and robust image classification

Y. Xie (✉) · L. Liu · X. Zhang
Science and Technology on Information System Engineering Laboratory,
National University of Defense Technology, Changsha 410073, People's Republic of China
e-mail: xyx89@163.com

X.-P. Zhang
Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada
e-mail: xzhang@ee.ryerson.ca

X. Luan
Changsha University, Changsha 410003, People's Republic of China

Published online: 28 May 2013

application. We consider eight images scenes: meetings, mass, beach, etc. By effective feature extraction from images and construct of reasonable classification model, we realize accurate and robust specific scene detection methods and thus lay a foundation for the later semantic image retrieval.

Specific scene detection and analysis of images can be treated as classifying images according to the predefined standard. It needs to apply content analysis technology to analyze and understand the semantic information of images automatically. The first step for image scene detection is to extract and represent image features. The extracted features should represent the semantic content of images sufficiently and have definite robustness and stability in environment changing. So far, most researches on the image analysis are based on low-level global features, such as color, texture, and shape, etc. [17, 18]. However, these low-level image descriptors are usually unintuitive for users. In recent years, local features are introduced to image feature extraction because of their invariance to image rotation, illumination and scaling changes [12, 20]. Meanwhile, considering the advantages of global features and local features respectively, some researchers have also proposed the feature fusion methods to extract image feature for effective scene detection [6, 14, 19]. As to the construct of classifier, some statistical machine learning methods like Bayesian theory, support vector machine, and graph model are applied to semantic classification and scene detection early or late. For example, in [4], a Bayesian hierarchical model for natural scenes categories learning is proposed. In this method, it represents the image of a scene by a collection of local regions, denoted as codewords obtained by unsupervised learning. It finally realizes satisfactory categorization performances on a large set of 13 categories of complex scene based on Bayesian hierarchical model. In [21], a method for semantic classification of images is proposed by using the combination image features of texture, edge, color histogram and support vector machine. So far, the methods mentioned above are all looking on an image as a whole to fulfill image scene detection. To better represent the semantic content of an image, some researchers use an intermediate semantic representation such as bag-of-words, or object-based models before classifying. In [3], a new learning technique, which extends Multiple-Instance Learning (MIL) and its application to the problem of region-based image categorization, is proposed. In this method, it extends multiple instances learning framework to DD-SVM framework, and proposes a block-based image scene detection method by description of images with block sets. Li et al. [10] develop an ALIP (Automatic Linguistic Indexing of Pictures) system, which obtains specific image concepts by training the color and texture features of the blocks of images and using two-dimensional Hidden Markov Model. Murphy et al. [13] proposes the joint recognition method of scene and objects by using graph model to correlate block features with objects. In recent years, with the further studies on BOVW (Bag of Visual Words) [22] and PLSA (Probabilistic Latent Semantic Analysis) [1, 5], some researchers also focus on improving them [7, 9, 23], using visual words to represent image semantics and then adopting statistical model to realize scene classification.

In summary, there are generally two aspects in image scene detection: one is the extraction and representation of image features, the other is the construct of image classifier. As to the former, many methods have already been proposed such as global feature representation, local feature representation, and intermediate semantic feature extraction, et al. In recent years, especially with the introducing of visual information such as local invariant feature, and the continuous improvement of statistical machine learning methods such as support vector machine, it becomes possible for automatic image scene detection. Although there have already been many work focused on image scene classification, fast and accurate scene detection remains to be a challenging problem. Most of the current work pays much attention to the precision of image scene detection, but cares little about the efficiency. However, for many image scene detection applications, we need a better trade-off between precision and efficiency.

In this paper, according to the requirement of eight specific scene detection applications, we propose an automatic image scene detection method based on local invariant features. First, considering that all these eight specific scenes have obvious local features, which are invariant to image rotation, scaling, illumination changes, we extract the local features of images by local keypoints detection. Then calculate the SIFT feature descriptors. Since there are always hundreds and thousands of keypoints for a complicated scene image, the computational complexity of feature extraction is usually high. To overcome this difficulty, we propose two feature extraction strategies to help reducing the computational complexity. One is to select those keypoints whose scale sizes are above a certain threshold, and then average the features of them. The other is to sort all the keypoints by their scale sizes, select the top *n* keypoints and then average the features of the selected keypoints. In this way, each image can be represented by one feature. Finally, we construct a multi-classifier of image scenes, which is based on support vector machine and has the specificity of small sample and nonlinear.

The rest of the paper is organized as follows. First, the problem formulation of specific image scene detection is further introduced in Section 2. Section 3 describes the implementation of the specific image scenes detection algorithm. In Section 4, the experimental results are given and discussed. Section 5 concludes the paper with direction for future work.

2 Problem formulations

The specific image scene detection is to detect and recognize the specific scene included in an image by machine learning of computers. Supposing the image database is defined as: $I=\{C_i\}_{i=1}{}^{K}$, where C_i denotes the *i-th* scene category in the image database, K is the number of categories. $x_i \in R^N$ represents the image feature vector. The goal of the specific image scene detection is to classify the unlabelled image by its feature x_i to some certain category C_i.

In this paper, we detect eight specific image scenes, namely beach, highway, meeting, rocket launch, indoor, mass, mountain and tall building as shown in Fig. 1. The scene images are selected according to the following three standards, namely usability, observability and feasibility. Where, usability means whether the selected scene can be used for further image retrieval and analysis. Observability means whether the selected scene is easy to be distinguished by its visual information. Feasibility means whether the selected scene can be detected by machine learning methods, namely the complexity and ambiguity of the scene.

On one hand, the above image scenes are common in that they all have obvious local features. Lowe [11] first proposed efficient SIFT (Scale Invariant Feature Transform) local invariant feature. Since then, it is widely used in various applications especially in image matching. This method has been proved to have excellent invariant features in image rotation, scaling, affine, and differing viewpoint. These local features are significant in our specific image scene detection.

On the other hand, specific image scene detection can also be regarded as a multi-classification problem. For pattern classification, support vector machine (SVM) has a good generalization performance without domain knowledge of the problems, so we select SVM as the image classifier. Meanwhile, considering the advantages of support vector machine in solving problems such as small samples, non-linear and high dimensional pattern

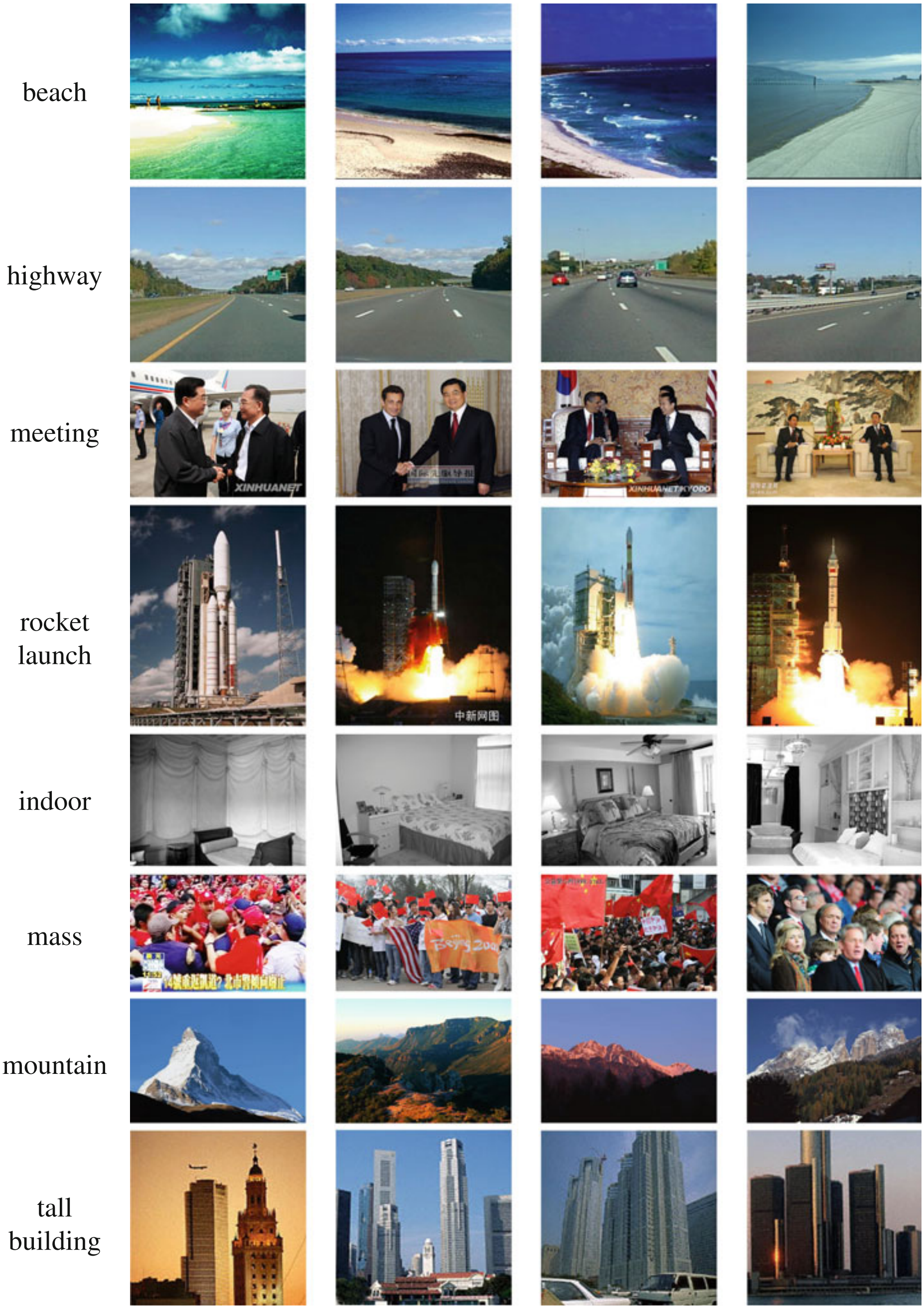


Fig. 1 Samples of eight specific image scenes

recognition, we will take full advantage of it. Support vector machine is based on the statistical theory of Vapnik-Chervonenkis (VC) dimension and the structural risk minimization. It makes the best trade-off between the model complexity (namely the learning precision of specific training samples) and learning capacity (namely the capacity of recognizing any samples correctly) according to the finite sample information to achieve the best generalization performance. Though SVM is a typical two-class classifier, it is also suitable to solve multi-classification problems such as specific image scene detection [20].

3 The proposed algorithm

3.1 Flowchart of the algorithm

The flowchart of the method is shown in Fig. 2. There are generally three parts in this method, namely image data pre-processing, image feature extraction and scene detection. Data pre-processing includes fault sample elimination, image graying and image size standardization. The second part is feature extraction, focusing on the detection of keypoints, local invariant feature extraction of images and feature extraction strategy. The third part is scene detection and analysis, accomplishing the construct and training of classifier and the recognition of image scene categories. We mainly demonstrate the later two parts of the algorithm.

3.2 Local feature extraction

3.2.1 Local keypoint detection

Currently, the relative effective method of local keypoint detection is based on multi-scale space of gray images. The main idea of the method is as follows: First, we establish the multi-scale description of images. Then, local keypoints are confirmed by searching all the pixels in a certain scale space and find the local extrema. One of the advantages of the method is that some hidden features in a certain scale space can be found in another scale space.

While establishing the multi-scale description of images, we first adopt pyramid method of images [2]. Supposing the multi-scale description of an image $I(x, y)$ can be expressed as $L(x, y, \sigma)$, which can be defined as follows:

$$L(x,y,\sigma) = G(x,y,\sigma)*I(x,y) \tag{1}$$

where

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-\left(x^2+y^2\right)} \Big/ 2\sigma^2 \tag{2}$$

(x, y) represents the pixel position of the image, σ is the scale size of the image. Generally the larger scale size corresponds to the outline feature of an image, whereas the smaller scale size corresponds to the detailed feature of an image. $G(x, y, \sigma)$ is the Gaussian kernel function.

Then, we subtract the neighbor scale space function to get the Gaussian difference scale space called DoG (Difference of Gaussian), which can be denoted as $D(x, y, \sigma)$:

$$D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma))*I(x,y) = L(x,y,k\sigma) - L(x,y,\sigma) \tag{3}$$

where k is the coefficient of the scale size.

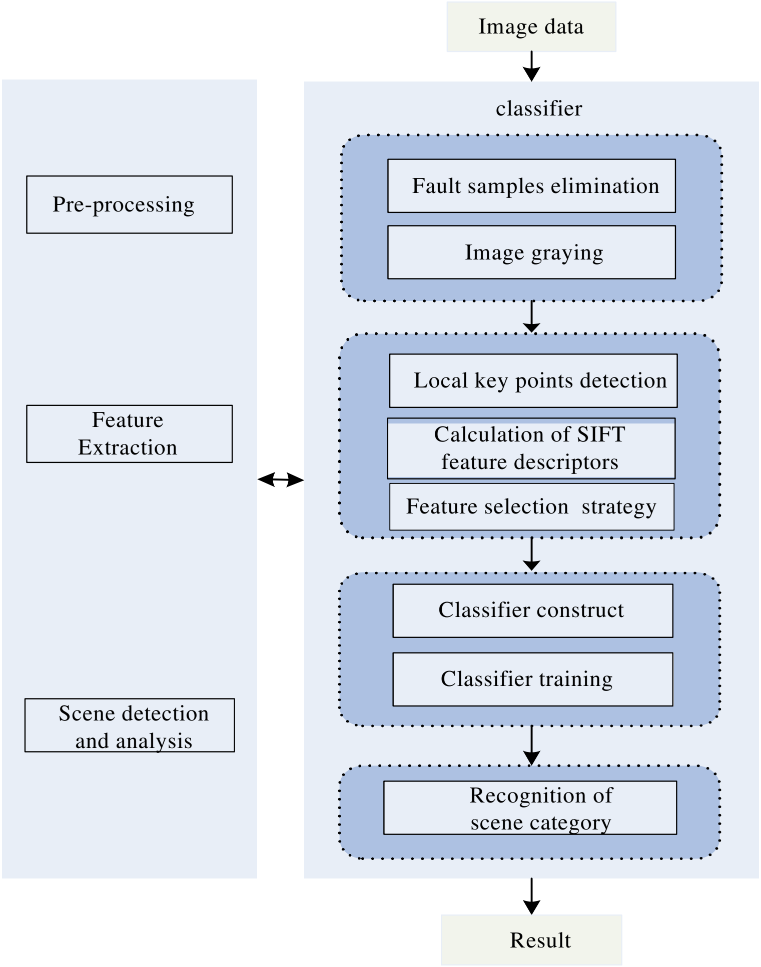


Image data
classifier
Pre-processing
Fault samples elimination
Image graying
Feature Extraction
Local key points detection
Calculation of SIFT feature descriptors
Feature selection strategy
Classifier construct
Classifier training
Scene detection and analysis
Recognition of scene category
Result

Fig. 2 Flowchart of the algorithm

In this scale space, we detect the local extrema as the candidates of local keypoints. The so-called local extrema means the DoG values of its pixel is maximum or minimum compared with other entire neighbor pixel's DoG value. Since DoG value is sensitive to noise and edges, we need a further check to assign it as a local keypoint after detection of the local extrema in the above DoG scale space. To fulfill this task, we first precisely locate the position and scale of the local keypoints by fitting the 3D quadratic function. Then, we discard the local extrema with low contrast and those unstable edge response points to improve the ability of anti-noise.

The approach can be described as follows [12]:

First, use the Taylor expansion of the scale-space function $D(X)$, shifted so that the original is at the sample point:

$$D(X) = D(x_0, y_0, \delta) + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2} X \tag{4}$$

where $X = (x, y, \sigma)^T$ is the offset from the sample point. The derivatives are estimated by taking differences of neighboring sample points.

The location of the extremum $\widehat{X}$ is determined by taking the derivative of this function with respective to X and set it to be zero, giving:

$$\widehat{X} = -\frac{\partial^2 D^{-1}}{\partial X^2}\frac{\partial D}{\partial X} \tag{5}$$

By substituting the extremum $\widehat{X}$ into the scale-space function, we can get the function value at the extremum $D(\widehat{X})$.

$$D\left(\widehat{X}\right) = D(x_0, y_0, \delta) + \frac{1}{2}\frac{\partial D}{\partial X}\widehat{X} \tag{6}$$

Then all extremum with a value of $D(\widehat{X})$ less than a certain threshold are being looked upon as low contrast keypoints and will be discarded.

Second, to further reduce the edge responses of DoG function, a 2×2 Hessian matrix H is computed at the location and scale of the keypoint.

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{7}$$

Then, compute the sum of the eigenvalues from the trace of H and their product from the determinant:

$$Tr(H) = D_{xx} + D_{yy} = \alpha + \beta \tag{8}$$

$$Det(H) = D_{xx}D_{yy} - \left(D_{xy}\right)^2 = \alpha\beta \tag{9}$$

where α is the eigenvalue with the largest magnitude and β is the eigenvalue with the smaller one. Let r be the ratio between them, then,

$$\frac{Tr(H)^2}{Det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r} \tag{10}$$

If the condition below is not satisfied, the keypoint will then be discarded.

$$\frac{Tr(H)^2}{Det(H)} < \frac{(r+1)^2}{r} \tag{11}$$

In this way, we can get the local keypoints for later processing.

3.2.2 Local SIFT feature description

After local keypoint detection, we construct the feature description which is invariant to image scaling, rotation and illumination change. Lowe summarized the current feature detection methods based on invariant technology, and proposed a local feature descriptor method called SIFT, which is based on scale space and has all the above characters [12]. This descriptor is invariant to scale and orientation by using the principle orientation of the neighbor keypoints as the orientation of the keypoint.

The calculation of SIFT descriptors mainly includes two steps, namely the assignment of the principle orientation of local keypoints and the calculation. In the first step, we assign a principle orientation for each local keypoint to ensure the invariance of image rotation. This work is realized by the statistical analysis of the gradient orientation histogram distribution of each keypoint's neighbors. To reduce the calculation complexity, we sample among the neighborhood centered on each local keypoint, and calculate the gradient orientation histogram of the neighborhood. The maximum of the histogram represents the principle orientation of the keypoint's neighborhood. So far, we have finished constructing the local keypoint descriptor which has three types of information, namely location, scale and orientation.

In the second step, we calculate descriptors of local image area, which are invariant to illumination changes and differing viewpoint. First, we rotate the axis to be consistent with the local keypoint's principle orientation or the assigned multi-orientation to ensure the invariance of image rotation. Then, select 8×8 window centered on the keypoint as neighbor pixels. We construct feature descriptors by using 4×4 (all together 16) seeds for each local keypoint to build up the robustness of later image matching. Since each seed is composed of 8 dimensional orientation vectors, each local keypoint generates 16×8=128 dimensional feature descriptors. This is the so-called SIFT feature of local keypoint, which is invariant to image scale and rotation. Finally, we normalize the length information of the feature vector to eliminate the influence of illumination changes.

Figure 3 shows the result of SIFT feature extraction of the image whose scene category is "meeting". Where, (a) is the original image. (b) represents the coordinate of local keypoints. (c) shows the scale of each local keypoint. (d) represents the orientation of each keypoint.

3.2.3 Proposed feature selection strategy

In the former feature extraction step, we have already got the features including the location, scale, orientation and SIFT description information. Since there are often hundreds and thousands of keypoints in an image, we need to select those most representative features for training and speed up the training efficiency. From Fig. 3, we can see that the larger the scale is, the more important the keypoint may be to reflect the key information in a scene [14]. However, there seems to be no such importance to the coordinate and orientation information. Thus, we adopt mainly the scale information for scene detection. To reduce the computational complexity, the simplest and most apparent way is to reduce the number of keypoints. Based on this assumption, we propose two different simple feature extraction strategies as follows.

Strategy one: Select those keypoints whose scale size are above a certain threshold, and average the features of them.

Supposing image I as $I(x, y, \sigma, F)$, where x, y represents the location of the keypoint in an image, σ represents the scale size, and F represents the

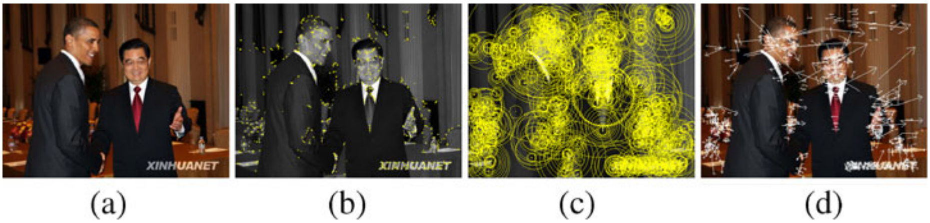


Fig. 3 Diagram of SIFT feature extraction of image whose scene category is "meeting"

description of SIFT features. Define A as a set of features whose scale size is above a certain threshold θ.

$$A = \left\{ I_i \middle| \sigma_{I_i} > \theta \right\} \tag{12}$$

Where I_i represents for the i-th keypoint's SIFT features. Then the average feature can be calculated as follows:

$$\overline{X} = \frac{1}{\|A\|} \sum_{I_i \in A} I_i \tag{13}$$

We use the average feature to represent an image for later training.

Strategy two: Sort all the keypoints by their scale size, select the top n keypoints, filter the later ones and then average the features of the selected keypoints.

Supposing $F = \{I_i | i = 1, 2, \ldots, m\}$is a set of features of m keypoints. F' = $\{I_j | j = 1, 2, \ldots, n\}$ is a set of features after sorting by the scale size of all keypoints' features and selecting the top n keypoints. Then the average feature can be calculated as follows:

$$\overline{X} = \frac{1}{\|F'\|} \sum_{I_j \in F'} I_j \tag{14}$$

We also use the average feature to represent an image for later training.

In conclusion, both feature selection strategies focus on the scale information of the keypoints and aim at reducing the computational complexity by discarding the keypoints. The difference is that, the first strategy uses the threshold method to rapidly reduce the number of keypoints, which is efficient but may cause the misdetection of image scenes. However, the second strategy sets the number of the keypoints to be top n, which can assure the precision of detection to some extent. Thus, considering the trade-off between efficiency and precision, we think strategy two more suitable for our applications. The later experimental results prove our analysis here.

3.3 Classification

Based on the above feature extraction, the next work is to train and test specific image scenes samples. Considering the advantages of support vector machine in solving problems such as small samples, non-linear and high dimensional pattern recognition, we use support vector machine to solve the problem of specific image scene detection. The main idea of support vector machine is to seek the optimal hyperplane such that the sum of distance between two class samples and the hyperplane is maximized.

Here we give a brief overview of binary classification with SVMs [20]. Let the separable training set be $T=\{(x_i,y_i)\}_{i=1}^{N}$, where $x_i \in R^N$ is the input feature vector, $y_i \in \{-1, +1\}$ is the label of a class, +1 denotes the positive example, -1 denotes the negative example, and N is the number of training samples. Supposing the discriminating function is $g(x) = w^T x + b$, where w is a weight vector, b is a bias. If the training set is linearly separable, we can easily get the classifier hyperplane by calculating: $w^Tx + b$=0. The goal of the SVM is to find the parameters w_0 and b_0 such that the distance between the hyperplane and the nearest sample point is more than 1.

If it is not linearly separable, it is necessary to map the input training vector into a high-dimensional feature space using a kernel function $K(x, x_i)$, then create the optimal hyperplane and implement the classification in the high-dimensional feature space.

Traditional SVM method can only solve two-class classification problem, while in this paper we need to classify eight different scene categories, that means we need to construct a multi-classifier to fulfil the task. For example, the construct of multi-classifier can be accomplished by combination of $n(n\text{-}1)/2$ number of two-class classifier, where n is the number of the scene categories. During the training process, we focus on feature selection and kernel function selection.

The most commonly used kernel functions in support vector machine are linear kernel function, polynomial kernel function and radial basis kernel function (RBF).

The linear kernel function equals to a linear classifier and can be denoted as:

$$K(x, x_i) = (x, x_i) \tag{15}$$

The polynomial kernel function is a q-order polynomial as follows:

$$K(x, x_i) = [(x, x_i) + 1]^q \tag{16}$$

In the radial basis kernel function, each kernel function corresponds to a support vector. The output weight and the support vector are all decided by the algorithm itself. The function can then be expressed as follows:

$$K(x, x_i) = \exp\left(-\frac{|x - x_i|^2}{2\sigma^2}\right) \tag{17}$$

In the experiments, we investigate these three kernel functions respectively to find the most efficient kernel function for image scene detection.

4 Experimental evaluations

4.1 Image data and experimental setup

To verify the effectiveness of the proposed feature extraction strategies, we select a set of representative images as the experimental image database. There are eight categories of images; they are beach, highway, meeting, rocket launch, indoor, mass, mountain and tall building as we mentioned in Section 2. Some of the scene categories (like indoor) are from the scene dataset provided by Fei-Fei [4]. Some scene categories (like meeting, mass, rocket launch) are added here according to the requirement of our application. Most of these images are randomly selected from the websites. Each category has two hundred images, all are in JPG format. We design the following experiment, that is, under the same experimental environment, do feature extraction for the same image data, then detect and analyse them in classifiers with different kernel function. In the end, prove the performance of the algorithm by the precision of scene detection.

For the training data: We select one hundred images in each class, all together there are eight hundred image samples as training set;

For the testing data: We select the left one hundred images in each class, all together there are eight hundred image samples as testing set;

We adopt the precision of classification as the evaluation criteria, which can be denoted as follows:

$$P = N_{\text{pos}} \Big/ N \tag{18}$$

where P represents the precision of classification, N_{pos} is the correctly classified samples, and N is the total samples in the database.

The experiments are designed to focus on two problems, namely the decision of kernel functions of classifier and the strategy of feature selection.

4.2 Results and discussion

4.2.1 Evaluation of the optimal kernel function

The most commonly used kernel functions in support vector machine are linear kernel function, polynomial kernel function and radial basis kernel function. In this paper, we first compare the efficiency of these three kernel functions and then select the best one as our kernel function.

Table 1 shows the efficiency comparison of these three kernel functions in specific image scene detection under the given parameters.

The experiment shows that linear kernel function has the worst efficiency. Polynomial kernel function is superior in fitting training data sets, but is insufficient in generalization of testing data sets. In summary, radial basis kernel function has a whole performance in both the fitting capacity of training data and the generalization capacity of testing data. Thus, in the following experiment, we adopt RBF kernel function.

4.2.2 Evaluation of the feature extraction strategy

- Experimental results of strategy one

  Strategy one is to select those keypoints whose scale is above a certain threshold to train and test.

  For better comparison of each class of image scene detection using strategy one, we draw a diagram according to the experimental result as shown in Fig. 4. This diagram reflects the detection precision of each image category with different scale size threshold of n ($1 \leq n \leq 10$ and $n \in Z$). From Fig. 4, we can see that all image categories have relatively high detection precision when the scale size is 1. And among all the scale sizes, “mass” images have higher performance than other categories of images. That is because “mass” images have many dense keypoints, which helps it not to be confused with other categories.

Table 1 Efficiency comparison of different kernel functions

	Linear	Polynomial	Radial basis
Training data (%)	69.25	99.75	91.25
Testing data (%)	58.75	62.38	66.00

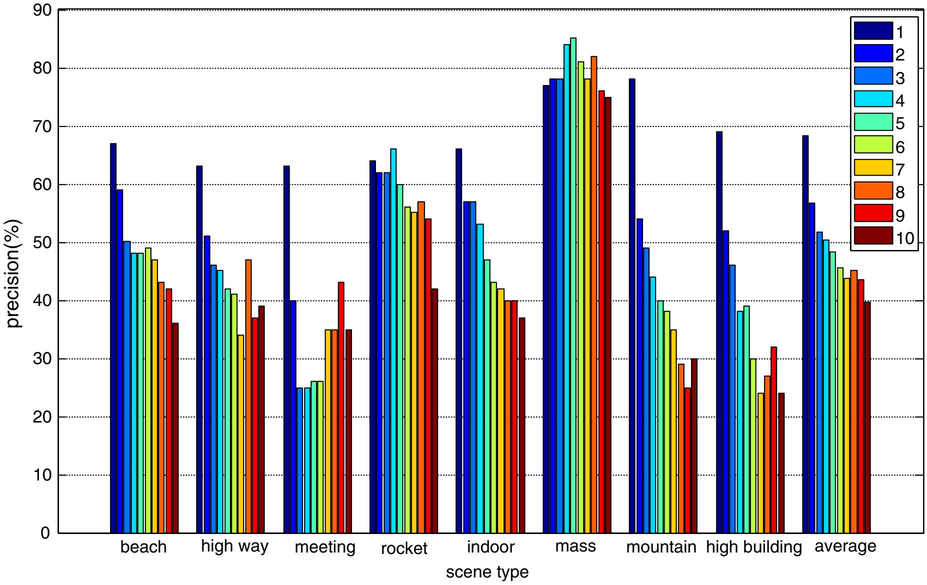


Fig. 4 Detection precision of eight different scenes with different scale size by using strategy one

- Experimental results of strategy two

  Strategy two is to sort all the keypoints by their size of scale, and then select the top *n* keypoints and average them.

  For better comparison of each class of image scene detection using strategy two, we draw a diagram according to the experimental result as shown in Fig. 5. In the legend, different colour represents for different number of keypoints.

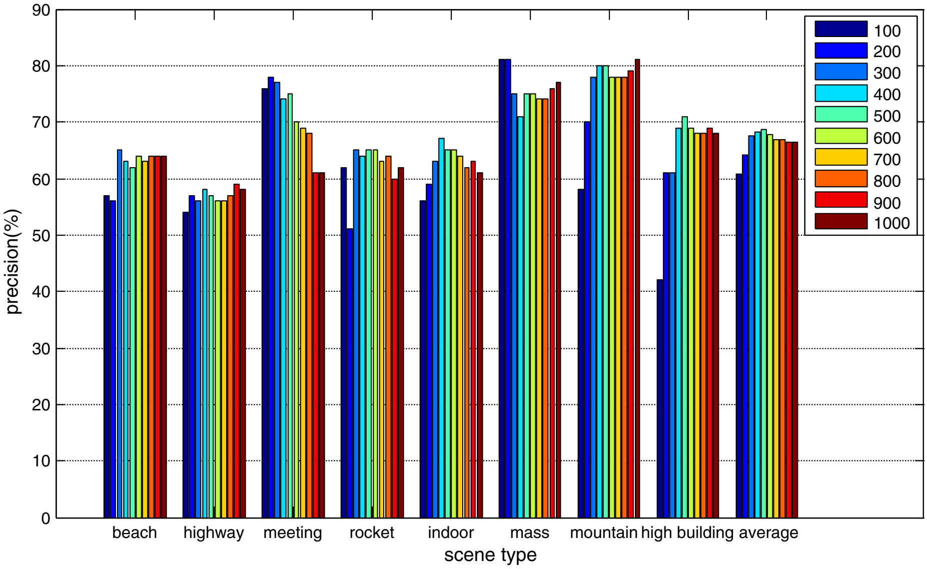


Fig. 5 Detection precision of eight different scenes with different scale size by using strategy two

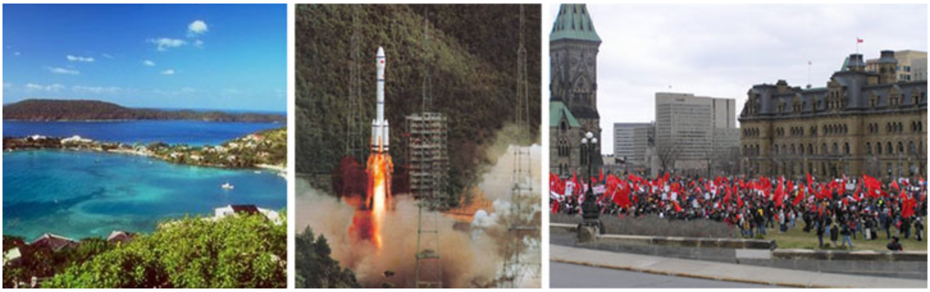


Fig. 6 Images which should have bi-annotation

From the diagram, we find that strategy two is more robust than strategy one, and can achieve relatively higher detection precisions than strategy one.

At the same time, we find that there are some false detection results. There are some reasons for that:

(1) Single annotation

This is one important factor that induces the false scene detection. Sometimes the scene semantic content of an image is mixed. For example, beach and mountain may appear in an image at the same time, but we always assign it to one scene category according to our experience. Of course, this assumption is not in accordance with the characters of machine learning, as shown in Fig. 6.

In Fig. 6, the first image is false detected by our method as "mountain" because it is annotated as "beach" by manual. The second image is false detected as "mountain" while its manual annotation is "rocket". The third image is false detected as "tall building" while its manual annotation is "mass".

(2) Ambiguity of image's semantic itself

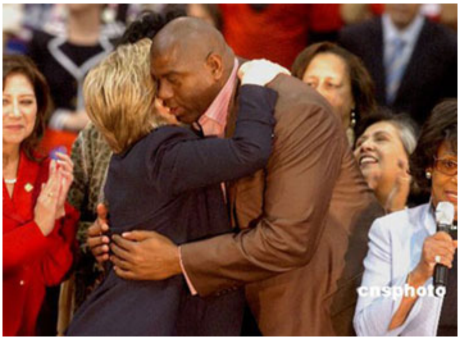
Some images have both the features of scene A and scene B, which makes it difficult to distinguish even if by manual annotation, as shown in Fig. 7.

(3) Similar keypoints distribution between images

For example, the image of a church going straight to the sky has similar keypoint distribution with the image of rocket. Another example, the image of peaceful sea has the similar keypoint distribution with the image of high way, as shown in Fig. 8.

By the above experiment, we find that from the view of detection precision, strategy two has an average performance among different scene categories, while

Fig. 7 Ambiguity between "meeting" and "mass"



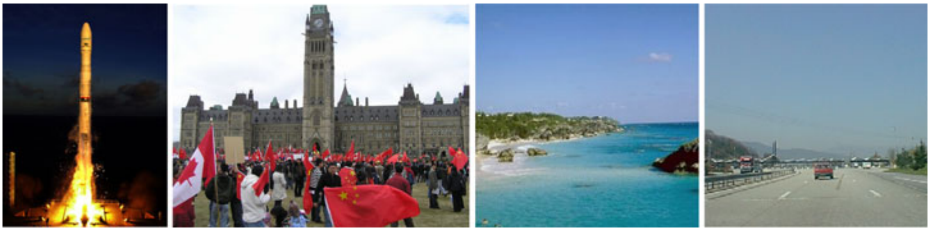


Fig. 8 Similar keypoints distribution results in the confusing of scene detection

strategy one seems to have more unstable performance. But whatever strategy we choose, the detection precision of "mass" images can achieve excellent performance. Thus, from the application's consideration, we eventually choose strategy two as the ultimate strategy of our scene detection.

4.2.3 Overall results and comparative evaluation

According to the requirement of our special application, we need to realize relatively fast and accurate image scene detection for eight specific scene categories. To evaluate the effectiveness of the proposed algorithm, we mainly focus on the average detection precision. The average scene detection precisions for different methods are shown in Table 2.

From Table 2, we can see that our method has similar detection precision compared with other methods. As to the efficiency of scene detection, since seldom literatures have mentioned about that, it is difficult for us to make a reasonable comparison. However, our goal is to make a better trade-off between precision and efficiency for scene detection to satisfy the users' requirements, we can get the estimated computational complexity by analyse the main steps in each method. Most of current scene detection methods are based on the feature extraction methods like visual bag-of-words, PLSA or some other feature fusion models. Although these methods can sometimes get high detection precision, the computational complexity of them are always high accordingly, which makes them not practical for fast scene detection. For example, in [4], its image feature detection and representation method is mainly based on the modified Latent Dirichlet Allocation (LDA) model, which has a high complexity in time of $O\left(NK\overline{n(d)}\left[\overline{n(d)}+M\right]\right)$ [8], where N is the number of the document, K is the number of topics, $\overline{n(d)}$ is the mean of the documents' length, and M is the size of the document. In [15], there are four steps in the representation of an image in all, namely interest point detection, descriptor computation, quantization and vocabulary model construction, and PLSA modelling. The first two steps have similar computational complexity to the feature extraction of our method, but as mentioned to the later two steps, it is well known that they are both time-consuming, which also has a high

Table 2 Comparison of the average detection precision

Method	Average detection precision (%)
[4]	65.2
[15]	66.5
[16]	72.7
Our method	66.7

time complexity like LDA [8]. In [16], it uses an intermediate space based on the low dimensional semantic “theme” image representation and weak supervision to realize image scene detection. The steps in this method, such as bag of features extraction and theme modelling, are also time-consuming like [4] and [15]. However, the method we proposed here is simple and has much lower computational complexity. It uses the average local feature as the representative of an image, thus it can meet the users’ needs and make a better trade-off between efficiency and precision for image scene detection.

In our method, to promise the relatively high scene detection precision, we adopted local invariant feature like SIFT for its invariance to image scaling, rotation and illumination changes. At the same time, to accelerate to scene detection speed, we reduced the number of the detected local keypoints and proposed two feature extraction strategies to obtain image features. After using the suitable kernel function to train the images, we finally got the relatively fast and accurate detection results which satisfied the actual requirement for specific image scene detection.

5 Conclusions

In this paper, according to the requirements of our fast and robust scene detection application, we focus on the problem of automatic detection of eight specific scenes including meeting, mass, beach, etc. First, we extract local features of images by local keypoints detection, and then calculate the SIFT feature descriptors. Then we propose two feature extraction strategies to reduce the computational complexity. Finally, considering the advantages of support vector machines in solving problems such as small samples, non-linear and high dimensional pattern recognition, we construct multi-classifier based on a support vector machine and select feasible features for training to achieve acceptable detection results. We also design the corresponding experiment focusing on solving two problems, namely, the decision of kernel function of classifier and the feature selection strategy. The experimental results show that the new method can achieve relatively accurate and robust specific scene detection results by using radial basis kernel function for the classifier and using the feature extraction strategy of selection the top *n* keypoints by the scale size order. Future work includes the solving multi-annotation problem while maintaining the high detection precision.

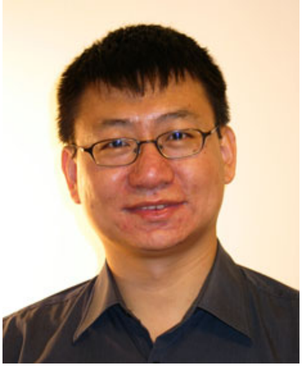
Acknowledgments This work has been supported by the National Natural Science Foundation of China under contract No. 61201337 and by Changsha Municipal Science and Technology Project under contract No. K1205045-11. The authors are grateful to the anonymous reviewers for valuable comments.

References

1. Bosch A, Zisserman A, Munoz X (2006) Scene classification via pLSA. Computer Vision–ECCV 2006:517–530
2. Burt P, Adelson E (1983) The Laplacian pyramid as a compact image code. IEEE Trans Commun 31(4):532–540
3. Chen Y, Wang JZ (2004) Image categorization by learning and reasoning with regions. J Mach Learn Res 5:913–939
4. Fei-Fei L, Perona P (2005) A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005. IEEE, pp 524–531
5. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42(1):177–196
6. Horster E, Lienhart R (2007) Fusing local image descriptors for large-scale image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR’07, 2007. IEEE, pp 1–8

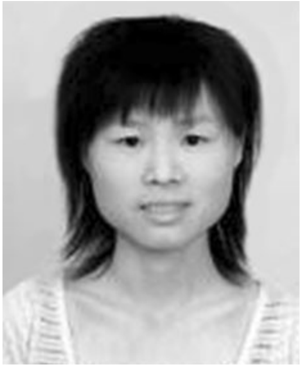
7. Jin B, Hu W, Wang H (2012) Image classification based on pLSA fusing spatial relationships between Topics. Signal Process Lett, IEEE 19(3):151–154
8. Keller M, Bengio S (2004) Theme-topic mixture model for document representation. Learning Methods for Text Understanding and Mining
9. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006. IEEE, pp 2169–2178
10. Li J, Wang JZ (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans Pattern Anal Mach Intell 25(9):1075–1088
11. Lowe DG (1999) Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999. IEEE, pp 1150–1157
12. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
13. Murphy K, Torralba A, Freeman W (2003) Using the forest to see the trees: a graphical model relating features, objects and scenes. Adv Neural Inf Process Syst 16
14. Qin J, Yung NHC (2012) Feature fusion within local region using localized maximum-margin learning for scene categorization. Pattern Recognit 45(4):1671–1683. doi:10.1016/j.patcog.2011.09.027
15. Quelhas P, Monay F, Odobez JM, Gatica-Perez D, Tuytelaars T (2007) A thousand words in a scene. IEEE Trans Pattern Anal Mach Intell 29(9):1575–1589
16. Rasiwasia N, Vasconcelos N (2008) Scene classification with low-dimensional semantic spaces and weak supervision. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008, 2008. IEEE, pp 1–6
17. Szummer M, Picard RW (1998) Indoor-outdoor image classification. In: IEEE International Workshop on Content-Based Access of Image and Video Database, 1998. IEEE, pp 42–51
18. Vailaya A, Figueiredo MAT, Jain AK, Zhang HJ (2001) Image classification for content-based indexing. IEEE Trans Image Process 10(1):117–130
19. Varma M, Ray D (2007) Learning the discriminative power-invariance trade-off. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, 2007. IEEE, pp 1–8
20. Wallraven C, Caputo B, Graf A (2003) Recognition with local features: the kernel recipe. In: Ninth IEEE International Conference on Computer Vision 2003. IEEE, pp 257–264
21. Wan HL, Chowdhury M (2003) Image semantic classification by using SVM. J Softw 14(11):1891–1899
22. Yang J, Jiang YG, Hauptmann AG, Ngo CW (2007) Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the International Workshop on Workshop on multimedia information retrieval, 2007. ACM, pp 197–206
23. Zeng P, Wen J, Wu L (2007) Scene classification using low-level feature and intermediate feature. In: Proceedings of SPIE Fifth International Symposium on Multispectral Image Processing and Pattern Recognition, MIPPR2007, 2007. pp 1–7

Yuxiang Xie is currently an associate professor in School of Information System and Management, National University of Defense Technology. She received her B.S., M.S. and Ph.D degrees from National University of Defense Technology in 1998, 2001 and 2004 respectively, all in Systems Engineering. Her current research interests include image and video analysis, classification and retrieval.



Xiao-Ping Zhang received the B.S. and Ph.D. degrees from Tsinghua University, in 1992 and 1996, respectively, all in electronic engineering. Since Fall 2000, he has been with the Department of Electrical and Computer Engineering, Ryerson University, where he is now Professor and Director of Communication and Signal Processing Applications Laboratory (CASPAL). His research interests include signal processing for communications, multimedia retrieval and video content analysis, computational intelligence, and applications in bioinformatics, finance and marketing. He is a registered Professional Engineer in Ontario, Canada, a Senior Member of IEEE and a member of Beta Gamma Sigma Honor Society.

Xidao Luan is an associate professor in Changsha University. He received his B.S. degree in Applied Mathematics in 1998, M.S. and Ph.D degrees in System Engineering in 2005, 2009 respectively, all from National University of Defense Technology. His current research interest is multimedia information processing and retrieval.



Li Liu is a Lecturer in School of Information System and Management, National University of Defense Technology. She received the B.S., M.S. and Ph.D degrees in Electrical Engineering from the National University of Defense Technology, Changsha, China, in 2003, 2005, 2012, respectively. She was a Visiting Student at the University of Waterloo, Ontario, Canada from 2010 to 2012. Her current research interests include computer vision, texture analysis, pattern recognition and image processing.

Xin Zhang received his B.S. degree in System Engineering from National University of Defense Technology in 2011. He is currently pursuing the M.S. degree. His research interests include image and video analysis, classification and retrieval.