Multimed Tools Appl (2014) 70:605–624
DOI 10.1007/s11042-012-1107-z

Evaluation of local features and classifiers in BOW model for image classification

Yanyun Qu · Shaojie Wu · Han Liu · Yi Xie · Hanzi Wang

Published online: 26 May 2012

Abstract Bag-of-word (BOW) is used in many state-of-the-art methods of image classification, and it is especially suitable for multi-class classification. Many kinds of local features and classifiers are applicable for the BOW model. However, it is unclear which kind of local feature is the most distinctive and meanwhile robust, and which classifier can optimize classification performance. In this paper, we discuss the implementation choices in the BOW model. Further, we evaluate the influences of local features and classifiers on object and texture recognition methods in the framework of the BOW model. To evaluate the implementation choices, we use two popular datasets: the Xerox7 dataset and the UIUCTex dataset. Extensive experiments are carried out to compare the performance of different detectors, descriptors and classifiers in term of classification accuracy on the object category dataset and the texture dataset. We find that the combinational detector which combines the MSER detector with the Hessian-Laplacian detector is efficient to find discriminative regions. We also find that the SIFT descriptor performs better than the other descriptors for image classification, and that the SVM classifier with the EMD kernel is superior to other classifiers. More than that, we propose an EMD spatial kernel to encode the spatial information of local features. The EMD spatial kernel is implemented on the Xerox7 dataset, the 4-class VOC2006 dataset and the 4-class Caltech101 dataset. The experimental results show that the proposed kernel outperforms the EMD kernel which does not consider the spatial information in image classification.

Keywords Bag of words · MSER · Hessian-Laplace · SIFT · EMD spatial kernel · Keypoint detector · Keypoint descriptor

Y. Qu · S. Wu · H. Liu · Y. Xie
Department of Computer Science, Xiamen University, Xiamen, China

Y. Qu
e-mail: yyqu@xmu.edu.cn

H. Wang (✉)
Center for Pattern Analysis and Machine Intelligence, Xiamen University, Xiamen, China
e-mail: hanzi.wang@ieee.org

1 Introduction

The Bag-of-words (BOW) model is widely used in visual object categorization [6] and image retrieval [8], because it is suitable to represent the multiple classes of objects in a unified framework. The core of the BOW model is that an image is represented by the visual words in a visual dictionary. In detail, a keypoint detector is firstly used to find interest regions, each of which is represented by a keypoint descriptor. After that, the local features are quantized and an image is represented by a histogram of word frequency. In the following content, we refer to both the keypoint detector and descriptor as "local features". In this paper, we focus on the two implementation choices: local feature and classifier, which greatly affect the performance of a method in image classification. We evaluate several popular keypoint detectors, keypoint descriptors and classifiers to see which local feature is more distinctive and meanwhile more robust for image classification, and which classifier can achieve better classification performance. We want to answer the question how the performance of a recognition method based on the BOW model is affected by the choice of both local features and classifiers.

Many studies have discussed the influence of the components of the BOW model on the accuracy of visual object categorization. Sampling strategy is the first key issue in the BOW model. A common strategy is to use a sophisticated multiscale keypoint operator, such as DOG [16], Harris-Laplacian [18], Harris-Affine [18], Hessian-Laplacian [18], Hessian-Affine [18], MSER [17], etc. Lazebnik et al. [11] combined the Harris-Laplace detector with the Laplacian detector to sample interest regions. The Harris-Laplace detector detects key points and the Laplacian detector detects key blobs which are complementary to some degree. Eric et al. [21] discussed the sampling strategy and demonstrated that the number of sampling patches is important in visual categorization: the more patches, the better performance for image classification.

A keypoint descriptor plays a second important role in the BOW model. There are many popular keypoint descriptors such as SIFT [16], SURF [1], GLOH [19], GIH [14], shape context [2] and steerable filter [9], etc. The BOW model quantizes these local features and represents an image by a histogram of word frequency. However, the BOW model neglects the spatial information among the local features. Therefore, many researchers pay attention to utilizing spatial information in the BOW model. Especially, in the field of image retrieval, many methods are proposed to utilize the spatial information in the BOW model. Wu et al. [27] proposed the bundle features which used the geometric rank of local features to represent the spatial information. Cao et al. [4] proposed the spatial BOW method to improve the retrieval results. Zhang et al. [28] proposed to model the spatial context of the local features in a group to avoid any single local feature instability or noises, which is superior to the bundle features in term of retrieval precision. However, these methods are more applicable to image retrieval instead of image classification, because they only use the spatial information and local features to measure the similarity between each pair of images, and they do not represent an image by a feature vector which is required by a classifier. Therefore, we discuss how to introduce the spatial information to the BOW model for object classification.

The design of a classifier is the third important factor in the BOW model. Lazbnike et al. [12] used the maximum entropy classifier for visual object categorization. Moosmann et al. [20] designed a random clustering forest for the visual object categorization. Liu et al. [15] designed a classifier based on the boosting algorithm to select the category-specific words which are composed of the "visual bits".

The codebook construction is the fourth important factor. Csurka et al. [6] grouped the local features of the training images by the K-means algorithm where the cluster centers were regarded as the visual words. Winn et al. [26] clustered the responses of the filter-bank where the compact visual dictionary was learned by pair-wise merging. Farquhar et al. [7]

used the Gaussian Mixture Model to model the density function of the key points for each class of images. Perronnin [22] built adaptive vocabularies which combined universal vocabularies with specific vocabularies. Larlus et al. [10] proposed the Gaussian Mixture-Multimodal LDA [3] model to generate the visual codes. Moreover, the construction of visual dictionary is often related to classifiers, and the design of a classifier is based on the visual word representation [15, 20].

There are two work [19, 29] closely related to our work. Mikolajczyk et al. [19] evaluated the influence of local descriptors in image matching and object (or scene) recognition. They showed that GLOH and SIFT are superior to other keypoint descriptors. However, they evaluated the keypoint desciptors only for image matching, not for image classification. It is unclear whether or not a descriptor can achieve the same performance in image matching as in image classification. Furthermore, we want to know if a keypoint descriptor designed for image matching under some constrained conditions is also discriminative for image classification. For example, GIH [14] is a deformation invariant descriptor for image matching, and we want to know what performance it achieves in image classification. Zhang et al. [29] did a comprehensive study on local features for texture and object classification. Based on the extensive experiments, they pointed out that the descriptor combining SIFT with SPIN and RIFT was the most discriminative for image classification. Moreover, they mainly analyzed the performance of SVM with different kernels and designed a SVM classifier with the Earth Mover's Distance (EMD) kernel. However, there are three limits in Zhang's method: 1) they only evaluated the performance of the SVM classifiers with different kernels, but they did not compare its performances with the performances of the other types of classifiers; 2) they neglected the MSER detector [17], which is widely used in computer vision; 3) the spatial information of the local features was not mentioned. In this paper, we evaluate the performance of various keypoint detectors, keypoint descriptors and classifiers. We also discuss how these components affect the classification performance in the BOW framework. In detail, we evaluate seven different types of classifiers: the SVM classifier with the Radius Basis Function (RBF), the SVM classifier with the χ^2 kernel, the SVM classifier with the EMD kernel, the Adaboost classifier, the random forest classifier, the maximum entropy classifier and the Naïve Bayes classifier. Furthermore, we propose the EMD spatial kernel, which encodes the spatial information.

This paper extends our previous work [5] by proposing a novel spatial kernel to improve the classification performance, adding additional empirical results and making an in-depth analysis of our approach's performance. The rest of this paper is organized as follows. In section 2, we evaluate the performances of the keypoint detectors, the keypoint descriptors and the classifiers. And then we propose the EMD spatial kernel in section 3. We evaluate the EMD spatial kernel in term of object classification on the Xerox7 datasets, the 4-class VOC2006 dataset and the 4-class Caltech101 dataset in section 4. In section 5, we draw conclusions.

2 Empirical evaluations

2.1 Experimental setup

We use two types of datasets: the Xerox7 dataset, which is an object category dataset; and the UIUCTex dataset, which is a texture dataset. These two standard datasets are popular and have been used to evaluate the performance in visual object categorization. The Xerox7 dataset contains 7 object classes: faces (792), bikes (125), buildings (150), cars (201), trees (150), books (142), phones (216). It is a challenging dataset because the images are captured in real world and in different viewpoints which cause the variance of appearance for the same

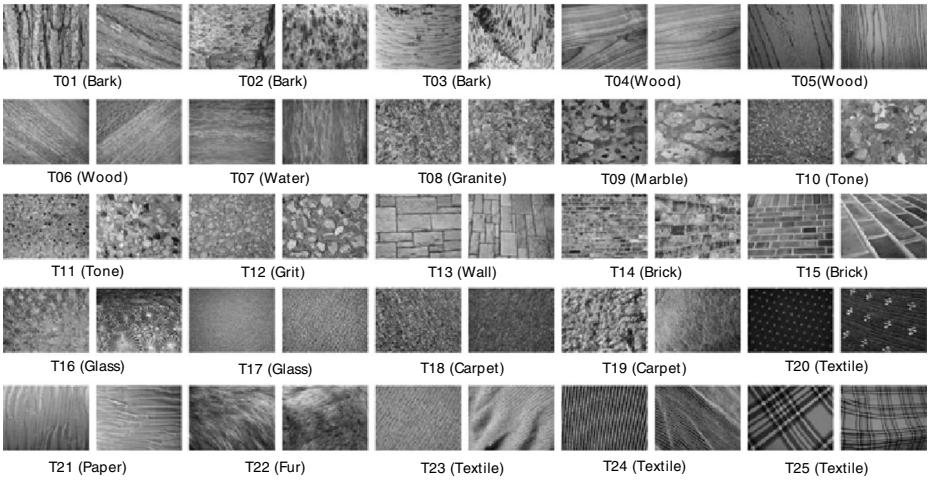
class of objects. The UIUCTex dataset contains 25 texture classes and each class contains 40 images. Figure 1 shows some examples in two datasets.

In order to evaluate the local features and the classifiers, we use the classification accuracy as a criterion. For the image representation of histogram distribution, we use the K-means algorithm to cluster local features and get a dictionary with 1,000 visual words. Thus an image can be represented as a histogram distribution with 1,000 dimensions. Moreover, each image class is split randomly into two separate sets: one set is for training and the other set is for testing. In the evaluation, we run 5 random trails and report the averaged results. In order to classify multiple classes, we take the one-against-other strategy for the binary classifier, such as SVM and Adaboost.

2.2 Empirical evaluation of keypoint detectors

We evaluate the six popular keypoint detectors in term of classification accuracy on the Xerox7 dataset and the UIUCTex dataset. The competing detectors are respectively

a) Xerox7



b) UIUCTex

Fig. 1 Some examples of each class in two datasets: **a**) Xerox7. **b**) UIUCTex

Difference of Gaussian (DOG) [16], Harris-Affine (HarA) [18], Harris-Laplacian (HarL) [18], Hessian-Affine (HesA) [18], Hessian-Laplacian (HesL) [18], and MSER [17]. DOG, HarL and HesL detect small circle blobs, while HarA, HesA, and MSER detect ellipse blobs. In Fig. 2, we show some results obtained by the six detectors. In the experiments, we use SIFT as the local feature descriptor and use SVM with RBF kernel as the classifier. The quantitative results in Tables 1 and 2 show that the Hessian-Laplacian detector has the superior performance compared with other competing detectors and that the MSER detector can find the discriminative regions. The averaged number of the detected interest points used in the experiments is given in Table 3. The regions detected by MSER are larger than those detected by other detectors, so the number of the keypoints detected by MSER is the smallest. Considering the trade-off between representation ability and computational complexity, we combine the Hessian-Laplacian detector with the MSER detector which we call as the HesLM detector. That is, we use both of the detectors to detect the key points, so the image is sampled densely and described by dense description. Figure 5a and b show the dense sampling and dense description. As shown in the last column of Tables 1 and 2, the HesLM detector has achieved a promising result in the classification accuracy on both datasets. The classification accuracy obtained by HesLM is similar to those obtained by Hessian-Laplacian and Hessian-Affine. HesLM spend much less running time than HesA. And HesLM achieves the highest classification accuracy in the four object classes of the Xerox7 dataset and in the 11 texture classes of the UIUCTex dataset.

2.3 The evaluation of keypoint descriptors

We evaluate seven state-of-the-art descriptors of local features in term of classification accuracy: i.e. SIFT [16], GLOH [19], SURF [1], Shape Context (SHAC) [2], steerable filter (STEF) [9], SPIN [11] and GIH [14].

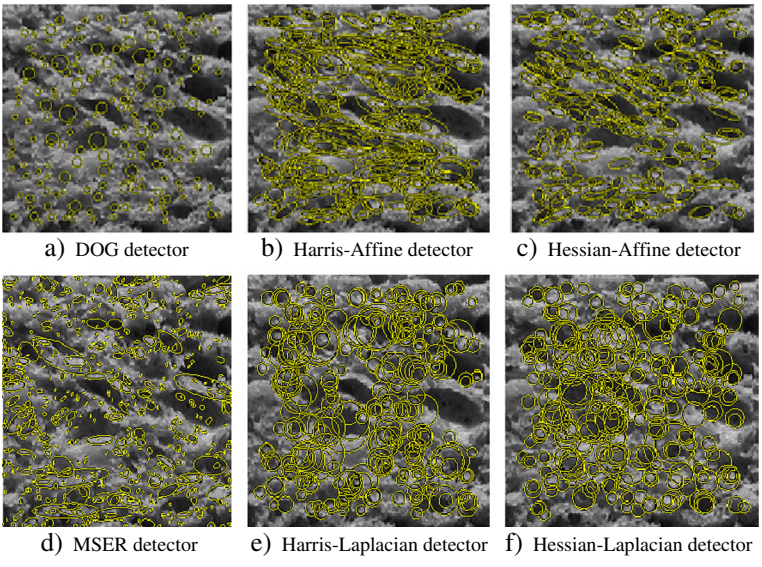


Fig. 2 The interest points detected by the six detectors. **a**) DOG detector. **b**) Harris-Affine detector. **c**) Hessian-Affine detector. **d**) MSER detector. **e**) Harris-Laplacian detector. **f**) Hessian-Laplacian detector

- SIFT is a very popular descriptor which is a histogram of gradient orientation. An interest region is divided into 4-by-4 grids. In each grid, the gradient angle is quantized into 8 orientations, such that a 128-dimention feature vector is formed.
- GLOH is an extension of SIFT. It computes a histogram of gradient orientation for a log-polar location grid.
- SURF is similar to SIFT in description. It is faster than SIFT due to using the Haar-like template instead of the Gaussian function to compute interest regions. It computes a four-dimension feature vector for a grid: the sum of gradients in x coordinate, the sum of gradients in y coordinate, the sum of gradient absolute values in x coordinate, the sum of gradient absolute values in y coordinate. Each interest region is divided into 4-by-4 grids, and then a 64-dimension vector is formed.
- Shape context is widely used to describe the shape of an object. It computes the histogram of edge points for a polar location grid, and each bin is the number of edge points in the corresponding grid.
- Steerable filter uses the derivates of a patch up to 4th orders which are computed by convolution with Gaussian derivatives. The changes of the orientation of derivatives give the element values of the feature vector.
- SPIN is a two-dimension histogram: the intensity is quantized to 10 bins and the radius is quantized to 5 bins. Each row of SPIN is a normalized histogram for a homocentric circle region.
- GIH is a deformation invariant descriptor. In GIH, an image is regarded as a two-dimension surface embedded in 3D space. The method uses the geodesic distance instead of the Euclidean distance. It is a 2D joint distribution of geodesic distance and the intensity.

In the test, we use MSER to detect the interest regions, and use SVM with the RBF kernel to classify the images. The results shown in Tables 4 and 5 demonstrate that SIFT is superior to other descriptors in term of the classification accuracy. In contrast, GLOH and shape context is inferior to SIFT for object classification. However, SIFT, GLOH, SURF, shape context and SPIN achieve a similar classification accuracy for texture classifications.

Furthermore, we also compare the local features with the textons on the UIUCTex dataset in order to address whether the local features are superior to the textons for texture recognition. We represent a texture image based on the textons according to Varma's work [24, 25]. In detail, we use MR8 [24] which are the filter banks to filter the images, and then we cluster the filter responses and obtain the textons. In our experiments, 10 cluster centers are computed by using the K-means method for each class, and all the cluster centers generated for all the classes form the texton set. At last a texture image is represented by the distribution of the textons. The SVM classifier with the χ^2 kernel is adopted to classify the texture classes. As shown in Table 5, the last column is the results based on the textons. The classification accuracies based on the textons generated by MR8 are inferior to those based on the local features except for GIH. In other words, the image patch features are superior to the filter response features. Our experimental results are consistent with Varma's conclusion [25].

2.4 Evaluation of the classifiers

We evaluate the seven classifiers in this section: i.e., SVM with the RBF kernel, SVM with the χ^2 kernel, SVM with the EMD kernel, Adaboost, the random forest classifier, the maximum entropy classifier, and the Naïve Bayes classifier. The first five classifiers belong to discriminative models. SVM is one of the most efficient tools for binary classification. The kernel selection in classifiers greatly affects the classification performance. In this paper,

Table 1 The evaluation of the seven different detectors in term of classification accuracy on the Xerox7 dataset

	DOG	HarL	HarA	HesL	HesA	MSER	HesLM
Face	97.2	98.1	97.5	98.7	**98.8**	96.1	98.2
Building	50.4	52	49.3	64.5	55.7	45.1	**65.6**
Tree	84.5	73.1	77.3	86.1	**89.3**	79.2	85.9
Phone	82	80	80.7	88.1	86.5	87.4	**88.7**
Car	63.2	62	54	71.2	67	62	**74.6**
Bike	87.1	87.4	85.8	90.6	**92.9**	86.5	92.6
Book	69	62.3	65.6	77.2	80.6	69.6	**80.8**
Average	83.5	82.2	81.4	88.1	86.8	82.7	**88.8**

Numbers in bold shows the results which achieve the best performance among the competing methods

Table 2 The evaluation of the seven different detectors in term of classification accuracy on the UIUCTex dataset

	DoG	HarL	HarA	HesL	HesA	MSER	HesLM
T01(bark)	98	**100**	**100**	**100**	**100**	**100**	**100**
T02(bark)	92	94	89	93	95	**99**	93
T03(bark)	95	94	96	98	**100**	99	99
T04(wood)	**100**	83	65	**100**	96	93	**100**
T05(wood)	98	**100**	96	**100**	**100**	91	98
T06(wood)	97	92	94	99	99	99	**100**
T07(water)	**100**	96	95	**100**	**100**	**100**	**100**
T08(granite)	92	**98**	92	94	96	90	96
T09(marble)	**97**	93	88	97	96	91	95
T10(tone)	94	97	97	97	97	**99**	97
T11(tone)	89	**100**	95	96	**100**	94	91
T12(grit)	99	**100**	**100**	98	95	95	96
T13(wall)	98	95	91	96	**99**	95	96
T14(brick)	98	96	87	94	**100**	96	**100**
T15(brick)	**100**	**100**	**100**	**100**	**100**	99	**100**
T16(glass)	**100**	**100**	99	**100**	**100**	**100**	**100**
T17(glass)	**100**	**100**	**100**	**100**	95	95	99
T18(carpet)	97	98	**100**	98	**100**	**100**	99
T19(carpet)	88	**98**	91	91	91	88	92
T20(textile)	96	**100**	**100**	**100**	99	99	97
T21(paper)	84	61	45	87	**94**	82	88
T22(fur)	97	87	71	**100**	99	92	**100**
T23(textile)	99	**100**	89	**100**	**100**	**100**	**100**
T24(textile)	**100**	**100**	**100**	**100**	**100**	**100**	**100**
T25(textile)	**100**	**100**	99	**100**	**100**	99	**100**
Average	96.5	95.3	91.2	97.5	**98**	95.8	97.6

Numbers in bold shows the results which achieve the best performance among the competing methods

Table 3 The mean number of the interest points sampled by different methods on the Xerox7 dataset and the UIUCTex dataset

	DOG	HarL	HarA	HesL	HesA	MSER
Xerox7	378	327	319	948	683	159
UIUCTex	1491	1304	1247	3174	2569	921

we focus on the kernels based on the exponential function $k(x,y) = \exp\left(-\frac{D(x,y)}{A}\right)$, where $D(x, y)$ is the distance between the vectors **x** and **y**. We compare the following three kernels.

- The RBF kernel uses the Euclidean distance to measure the distance between two vectors, $D(x,y) = \|x-y\|^2$.
- The χ^2 kernel defines $D(x, y)$ as the χ^2 distance, where $D(x,y) = \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-y_i)^2}{x_i+y_i}$.
- The EMD kernel defines $D(x, y)$ as the EMD distance, where $D(x,y) = \min_{\{f_{ij}\}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}}$.

The Adaboost algorithm is famous for its successful application in face detection. The classifier is the linear combination of a set of weaker classifiers, and it can be written as $h(x) = sign\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right)$, where $h_t(x)$ is a weaker classifier, and $\alpha_t > 0$. The random forest classifier is an ensemble classifier, which combines a set of decision trees and is a multi-class classifier. The object label is predicted by the decision tree classifiers voting or by using the average confidence value of the tree classifiers. In our case, we take the latter strategy to determine the label.

Among the classifiers mentioned above, the last two classifiers are of the generative models. The maximum entropy algorithm solves a classifier by maximizing the conditional entropy, which can be written as:

$$\{\lambda_k\} = \arg\max\left(-\frac{1}{|T|}\sum_{I\in T}\sum_{c} P(c|I)\log P(c|I)\right), \tag{1}$$

where $|T|$ is the number of the elements in the object set T, c is the number of the object classes, and $P(c|I)$ is the posterior probability, $P(c|I) = \frac{1}{Z}\exp\left(\sum_k \lambda_k f_k(I,c)\right)$.

Table 4 The evaluation of the seven descriptors on the Xerox7 dataset

	SIFT	GLOH	SURF	SHAC	STEF	SPIN	GIH
Face	**96.1**	95.3	95.2	95.6	92.8	94.5	92.3
Building	45.1	**47.5**	36	45.3	31.7	26.9	26.9
Tree	**79.2**	**79.2**	77.1	75.7	64.5	73.3	61.1
Phone	**87.4**	83.5	81.1	85.6	77.2	78.9	48.3
Car	**62**	60.4	56.8	57.8	54.8	53.6	36.8
Bike	86.5	86.1	87.1	**87.7**	75.2	80.3	85.2
Book	69.6	**69.9**	65.6	69	60.8	59.2	41.4
Average	**82.7**	81.8	79.7	81.5	75.3	76.7	68

Numbers in bold shows the results which achieve the best performance among the competing methods

Table 5 The evaluation of the eight descriptors on the UIUCTex dataset

	SIFT	GLOH	SURF	SHAC	STEF	SPIN	GIH	Texton
T01(bark)	98	**100**	99	**100**	**100**	99	29	75
T02(bark)	96	94	91	**100**	96	93	27	76
T03(bark)	93	**97**	90	96	87	94	52	69
T04(wood)	89	**94**	88	88	88	90	23	74
T05(wood)	98	94	**100**	95	92	98	67	85
T06(wood)	98	97	**100**	98	92	98	71	81
T07(water)	**100**	99	98	98	99	**100**	60	97
T08(granite)	96	89	95	88	83	**98**	23	91
T09(marble)	94	91	**98**	97	62	92	10	93
T10(tone)	93	97	96	98	96	**100**	15	97
T11(tone)	84	98	95	91	96	**100**	92	79
T12(grit)	90	**99**	90	97	95	92	15	78
T13(wall)	**100**	93	94	99	94	90	11	80
T14(brick)	**100**	96	91	97	91	97	28	60
T15(brick)	98	97	**100**	99	**100**	97	28	96
T16(glass)	**100**	**100**	**100**	**100**	94	94	79	91
T17(glass)	**100**	97	**100**	**100**	**100**	**100**	80	80
T18(carpet)	98	**100**	99	**100**	97	99	43	88
T19(carpet)	95	86	82	85	81	**96**	22	67
T20(textile)	**100**	99	96	**100**	**100**	**100**	73	89
T21(paper)	**99**	89	90	82	81	86	7	84
T22(fur)	91	80	**97**	77	86	88	56	77
T23(textile)	97	**100**	97	98	99	**100**	57	67
T24(textile)	**100**	**100**	**100**	**100**	**100**	**100**	88	90
T25(textile)	**100**	**100**	**100**	99	99	**100**	41	94
Average	**96.3**	95.4	95.4	95.3	92.3	96	43.9	82.3

Numbers in bold shows the results which achieve the best performance among the competing methods

The Naïve Bayes can be viewed as a maximum a posteriori classifier and the class label is predicted as follows:

$$c(x) = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(w_i|c) \tag{2}$$

where P(c) is the prior probability; $P(w_i|c)$ is the condition probability of the *ith* word given the *cth* object class, and $P(w_i|c) = \frac{1+\sum_{I_j \in c} N(t,i)}{|V|+\sum_{s=1}^{|V|}\sum_{I_j \in c} N(s,i)}$, where *N(t,i)* is the concurrent matrix of words and images.

In the experiments, we represent an image in two ways: the histogram distribution and the signature representation. Considering the computational complexity, we sample interest points randomly along the edges detected by the Canny operator instead of sparse sampling in order to compute the histogram distribution, and then obtain 200 image patches. We take SIFT as the descriptor of the sampling patches, and use the K-means algorithm to construct

the codebook which contains 1,000 words. We represent the image by the histogram of word frequency. At last, we classify the images by using different classifiers while we do not use the SVM classifier with the EMD kernel because it requires that an image is represented by signatures instead of the histogram of word frequency.

For the signature representation, there are two sampling methods: one is to use the HesLM detector; the other is to use the Harris-Laplacian detector combined with the Laplacian detector (HesLL) [11]. We use SIFT to describe the interest regions. After that, we cluster the SIFT feature vectors in each image and form 40 cluster centers which are regarded as the signatures of the image. And the ith image is denoted by $\{(s_{i1}, w_{i1}), (s_{i2}, w_{i2}) \cdots (s_{il}, w_{il})\}$, where s_{ik} is the ith cluster center, l is equal to 40, and w_{ik} is the corresponding weight which is the frequency of the kth cluster center in the ith image. Given two images, in order to compute the distance between two signatures S_i and S_j, where

$$S_i = \{(s_{i1}, w_{i1}), (s_{i1}, w_{i1}) \cdots (s_{il}, w_{il})\}, \quad S_j = \left\{\left(s_{j1}, w_{j1}\right), \left(s_{j1}, w_{j1}\right) \cdots \left(s_{jl}, w_{jl}\right)\right\},$$

we solve the optimal problem which minimizes a cross-bin dissimilarity measure between S_i and S_j,

$$D_{emd}(S_i, S_j) = \frac{\sum_{m=1}^{l}\sum_{n=1}^{l} f_{mn} d\left(s_{im}, s_{jn}\right)}{\sum_{m=1}^{l}\sum_{n=1}^{l} f_{mn}}$$

$$\text{s.t.} \begin{cases} f_{ij} > 0 \\ \sum_{j=1}^{l} f_{ij} \le w_{im} \quad 1 \le m \le l \\ \sum_{i=1}^{l} f_{ij} \le w_{jm} \quad 1 \le m \le l \end{cases} \tag{3}$$

where $d(s_{im}, s_{jn})$ is the Euclidean distance between the mth signature in the ith image and the nth signature in the jth image, f_{mn} is the flow value which is obtained by solving the linear programming problem. More details about the algorithm are given in [13] and [23]. After that, we use the SVM classifier with the EMD kernel to classify the objects. We implement the methods mentioned above on the Xerox7 dataset and the UIUCTex dataset. The classification accuracies are shown in Tables 6 and 7. We find that the SVM classifier with the EMD kernel achieved better accuracy than the other classifiers, and the EMD kernel based on the HesLM

Table 6 The evaluation of the seven classifiers using the Xerox7 dataset

	EMD1 HesLM_sig	EMD2 HarLL_sig	χ^2	RBF	Entropy	Bayes	Adaboost	RanTree
			Histogram distribution					
Face	96.8	**99**	96.5	95.7	95.5	90.3	93.4	98.4
Building	**77.5**	55	72	51.5	60.3	57.6	28	16.8
Tree	62.7	91	**93.3**	92.8	90.4	88.8	81.9	92.8
Phone	83	**91**	87	82.8	82.6	77.6	55.7	69.8
Car	**97.9**	78	88	70.2	63.8	77.8	45.4	61.2
Bike	94.4	**100**	85.5	88.7	89.4	89.4	63.2	67.4
Book	**85.3**	83	70.4	62	66.5	71.8	38.3	38.3
Average	**90.1**	89.7	89.1	84.1	84.2	82.9	70.3	76.3

Numbers in bold shows the results which achieve the best performance among the competing methods

Table 7 The evaluation of the seven classifiers on the UIUCTex dataset

	EMD1 HesLM_sig	EMD2 HarLL_sig	χ^2	RBF	Entropy	Bayes	Adaboost	RanTree
	Histogram distribution							
T01(bark)	**100**	**100**	66.3	96	93	68	81	75
T02(bark)	90	**95**	61.3	74	68	79	62	66
T03(bark)	**100**	95	68.8	82	74	63	70	47
T04(wood)	95	**100**	73.8	99	97	99	94	94
T05(wood)	**100**	**100**	76.3	85	88	84	91	87
T06(wood)	**100**	95	82.5	75	80	76	67	61
T07(water)	**100**	**100**	80	94	89	**100**	99	**100**
T08(granite)	**95**	90	81.3	79	83	87	74	85
T09(marble)	**100**	**100**	81.3	88	78	88	78	90
T10(tone)	**100**	**100**	63.8	69	75	47	60	36
T11(tone)	**90**	85	63.8	70	73	46	58	36
T12(grit)	**100**	95	70	62	67	53	47	21
T13(wall)	95	90	61.3	**97**	**97**	94	91	96
T14(brick)	**100**	**100**	75	63	68	32	51	26
T15(brick)	**100**	**100**	96.3	**100**	**100**	**100**	99	**100**
T16(glass)	**100**	**100**	92.5	98	99	98	98	**100**
T17(glass)	**100**	**100**	83.8	92	91	82	88	83
T18(carpet)	**100**	**100**	78.8	90	85	91	90	97
T19(carpet)	**100**	**100**	86.3	39	55	45	43	13
T20(textile)	**100**	**100**	86.3	**100**	98	98	91	**100**
T21(paper)	**95**	**95**	80	89	93	90	82	66
T22(fur)	**100**	**100**	70	94	89	87	83	83
T23(textile)	**100**	**100**	68.8	90	89	71	93	**96**
T24(textile)	**100**	**100**	81.3	99	**100**	**100**	**100**	**100**
T25(textile)	**100**	**100**	78.1	**100**	**100**	**100**	98	94
Average	**98.4**	97.6	76.2	85	85.2	79.1	79.5	74.1

Numbers in bold shows the results which achieve the best performance among the competing methods

detector is superior to that based on the HarLL detector. Moreover, the HesLM signatures achieve the highest classification accuracy in three image sets of the Xerox7 dataset, and in 22 image sets of the UIUCTex dataset.

Furthermore, we evaluate how a sampling method affects the performance of classification. We compare the previous mentioned sparse sampling methods and the random sampling method, and

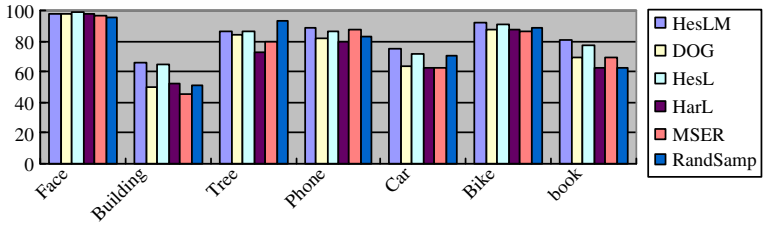


Fig. 3 The comparison of the sampling methods in term of the classification accuracies on the Xerox7 dataset

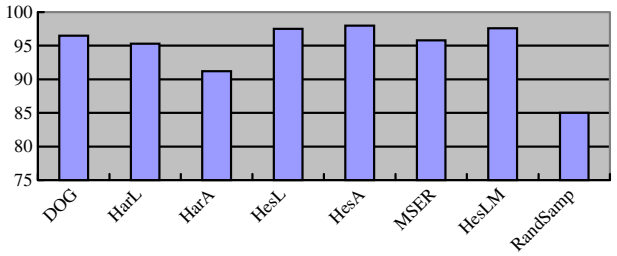


Fig. 4 The comparison of the sampling methods in term of averaged classification accuracy of the 25 texture classes on the UIUCTex dataset

all the experimental data come from Tables 1 and 6 for object classifications and from Tables 2 and 7 for texture classifications. Figure 3 shows the classification accuracies obtained by using the six sampling methods on each object class of the Xerox7 dataset, and Fig. 4 shows the averaged classification accuracies on all the texture classes achieved by the mentioned eight sampling methods. As shown in Figs. 3 and 4, random sampling is inferior to the sparse sampling methods in term of the classification accuracy. Because, for the purpose of simplicity, we only randomly sample 200 points along the edges which is smaller than the number of sampling points generated by the sparse sampling methods. The experimental results demonstrate that the number of the sampling points affects greatly the classification performance.

3 The EMD spatial kernel

For the experiments in section 2, we find that 1) the combinational detector HesLM can detect the discriminative regions; 2) SIFT is superior to the other local feature descriptors in image representation; 3) SVM with the EMD kernel achieves better performance in classification accuracy than the other classifiers. The experimental results also demonstrate that the BOW model gives a unified framework for multi-class image classification. However, it neglects the spatial information of the local features. Considering the relationship between the MSER regions and the Hessian-Laplacian blobs, these two detectors are complementary to some degree. The MSER detector can detect the interest regions which are bigger than the interest blobs detected by the Hessian-Laplacian detector. Moreover, an MSER region may contain some Hessian-Laplacian blobs, which forms the combinational pattern and embeds the spatial relationship. In this section, we propose an EMD spatial kernel to encode the spatial information of the local features, and boost the classification performance further.

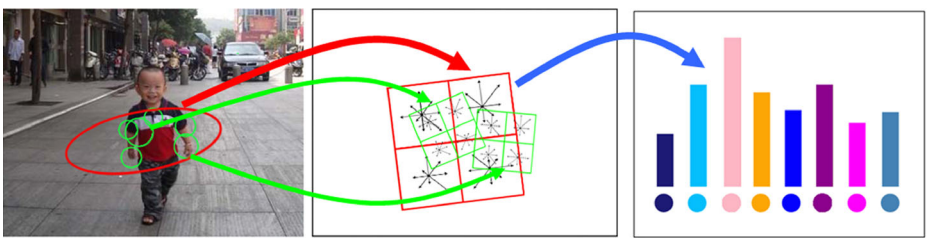


Fig. 5 The image representation of signatures based on the HesLM detector. **a**) The HesLM detector. **b**) The SIFT descriptor. **c**) The signatures

Firstly, we use the HesLM detector to sample image patches which are shown in Fig. 5a. The red circle represents the region detected by MSER, and the green circles represent the regions detected by the Hessian-Laplacian detector.

Secondly, we use the SIFT descriptor to describe the interest regions which are shown in Fig. 5b. Unlike to the methods [27–28] which are used in image retrieval and measure the similarity by using a voting strategy, we need to compute a kernel matrix to record the similarity between each pair of images in the training dataset. We propose to encode the spatial similarity in a kernel matrix. The element of the kernel matrix between the *ith* image and the *jth* image is defined as,

$$K(I_i, I_j) = K_{emd}(I_i, I_j) + K_{space}(I_i, I_j). \tag{4}$$

There are two parts in the kernel matrix. The first part is just like the EMD kernel matrix mentioned above. We compute 40 signatures for each image which are shown in Fig. 2. And then we use these signatures to compute the EMD kernel matrix. The second part encodes the spatial information. We construct a codebook with 200 visual words obtained by using the K-means algorithm for the MSER detector and the Hessian-Laplacian detector respectively. We measure the spatial similarity between each pair of images which is illustrated in Fig. 6. For each MSER region in an image, we split the region into four quadrants according to its major axis and minor axis. And the major orientation of the gradients in the MSER region is assigned to the first quadrant. We firstly match the MSER regions between each pair of images. The MSER regions are matched if they can be represented by the same visual word. And then we count the matching Hessian-Laplacian local features in the matched MSER regions. Two Hessian-Laplacian local features are defined to be matched if they are represented by the same visual word and located in the same quadrant of the MSER region. The similarity is defined as:

$$s_i = tfidf(f_i) * space(f_i), \tag{5}$$

where $tfidf(f_i)$ is to compute the weight of the visual words obtained by MSER according to the term frequency and document frequency, and $space(f_i)$ is the number of the matched Hessian-Laplacian local features in the MSER regions whose visual words are consistent.

Finally, we use the SVM classifier with the EMD spatial kernel to classify the images. When a query image is input, we do the same operations to obtain the signatures and the space information and then classify the query image with the SVM classifier obtained by using the training images. The framework of our approach is shown in Fig. 7.

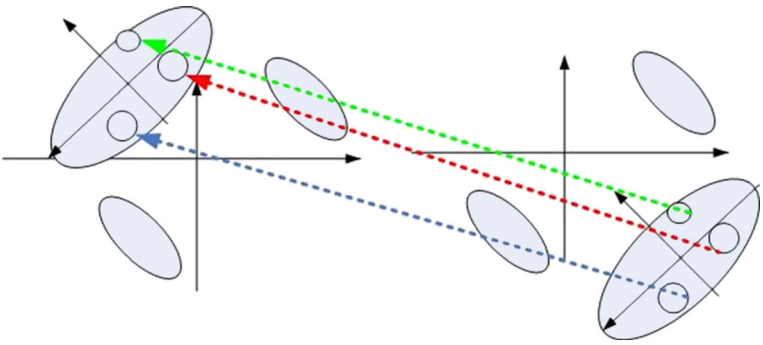


Fig. 6 The spatial matching of the visual words

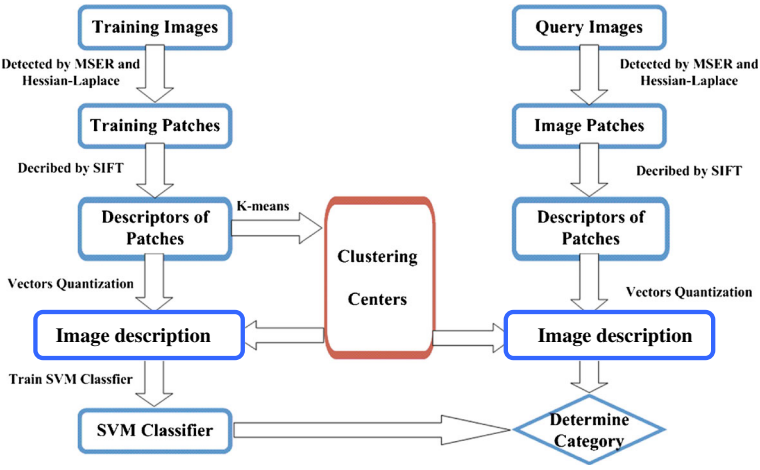


Fig. 7 The framework of the proposed approach

4 Experimental results

We test our approach on the Xerox7 Dataset, the 4-class VOC2006 dataset and the 4-class Caltech101 dataset. The 4-class VOC2006 dataset contains four image sets: bicycles (270), cars (553), motobikes (235), persons (666). The 4-class Caltech101 dataset contains four image sets: faces (435), Buddha (85), grand-pianos (99), and sunflowers (85). Each class is randomly split into two separate sets of images with the same size. One set is for training and the other set is for testing. We firstly compare different image representations. Three detectors and two representation features are considered. The three detectors are the Hessain-Laplacian detector (HesL), the MSER detector, and the HesLM detector. The two representation features are the histogram distribution (HD) and signatures (Sig). There are five kinds of combinations to represent an image:

(1) HesL + HD: Hessain-Laplacian detector + histogram distribution;
(2) MSER + HD: MSER detector + histogram distribution;
(3) HesLM + HD: Hessian Laplacian and MSER detectors + histogram distribution;
(4) HesLM + Sig: Hessian-Laplacian and MSER detectors + signatures;
(5) HesLM + Sig + space: Hessian-Laplacian and MSER detectors + signatures + space information.

For the first three image representation combinations, we construct a dictionary with 1,000 visual words, and thus an image is represented as a feature vector of 1,000 dimensions. We use the SVM classifier with a RBF kernel to classify objects. For the last two image representation combinations, we generate 40 signatures for an image. For the fourth combination, we use the SVM classifier with an EMD kernel. In order to compute the spatial information, we construct a dictionary with 200 visual words for the MSER detector and for the Hessian-Laplacian detector respectively. We use the EMD spatial kernel to classify the objects.

In Fig. 8, we compare the classification accuracy among the five representation methods, and the last group of the bars represents the averaged results obtained by the five methods on the dataset. The experimental results demonstrate that our approach achieves the highest classification accuracy, and the proposed EMD spatial kernel is superior to the single EMD kernel in object categorization. Therefore, our proposed EMD spatial kernel averagely improves the overall classification performance for object categorization.

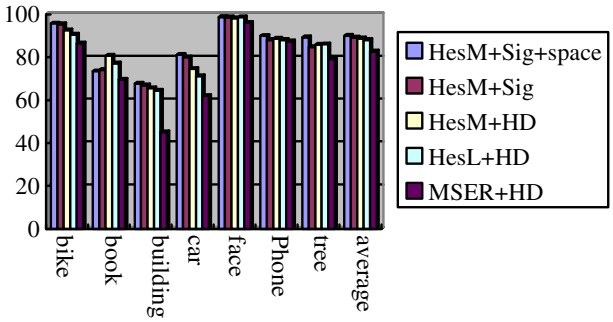


Fig. 8 The performance comparison of the five representation methods on the Xerox7 dataset

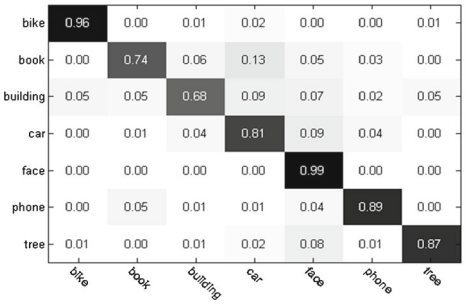
We also use the confusion matrix to evaluate our approach as follows,

$$M_{ij} = \frac{\left|\left\{I_k \in C_j \cap h(I_k) = i\right\}\right|}{\left|C_j\right|} \tag{6}$$

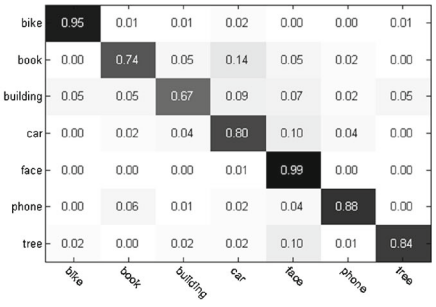
where $i,j \in \{1,\cdots,N_c\}$, N_c represents the number of classification, C_j denotes the images which belong to the jth class, and $h(I_k)$ denotes the predicted class for the image I_k. The values of the diagonal elements in the confusion matrix are marked by the black color and they represent the classification accuracy for each category.

We compare the EMD spatial kernel with the single EMD kernel. As shown in Figs. 9 and 10, the proposed EMD spatial kernel is superior to the single EMD kernel for almost all the classes of objects in term of classification accuracy except for the Face set and the Bicycle set, which demonstrates that our approach is more effective in object categorization.

Moreover, we compare the explicit spatial coding in (5) (see Fig. 11) and our implicit spatial coding (4) (see Figs. 9a and 10) in term of classification accuracy. As we know, (5) is mostly used in image retrieval. We design a classification method based on the retrieved results. In detail, we firstly use each image in a class to retrieve the images in the dataset. Secondly, we use a k-NN method to select the most similar images. Thirdly, in the retrieved images, we count the images belonging to each class, and the label of a query image is determined by the class of images whose number is the largest in the retrieved images. Figure 11 shows how the number of the remaining images affects the classification accuracies on the Xerox7 dataset (see the left subfigure) and on the 4-class VOC2006 dataset (see the right subfigure). Next, we analyze the average value of the classification accuracies obtained by using the two different coding approaches. On the Xerox7



a) The confusion matrix of the EMD spatial kernel



b) The confusion matrix of the EMD kernel

Fig. 9 The comparison of the EMD spatial kernel and the EMD kernel in term of confusion matrix on the Xerox7 dataset. **a**) The confusion matrix of the EMD spatial kernel. **b**) The confusion matrix of the EMD kernel

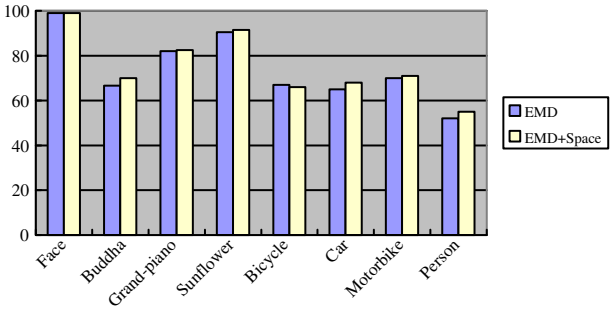


Fig. 10 The performance comparison of the EMD spatial kernel and the single EMD kernel on the 4-class Caltech101 dataset and the 4-class VOC2006 dataset

dataset, the highest averaged classification accuracy is 59 % when we use 330 retrieved images. On the 4-class VOC2006 dataset, the averaged classification accuracy is 43 % when we use 150 retrieved images. These two results are much smaller than the averaged classification accuracies of our proposed method which is 90 % on the Xerox7 dataset (obtained from Fig. 9a) and 63 % on the 4-class VOC2006 dataset (obtained from Fig. 10). Thus, we can conclude that the implicit spatial coding is superior to the explicit spatial coding.

5 Conclusions

In this paper, we analyze the influence of local features and the classifiers used in the BOW model on image classification accuracy. We evaluate different local features and classifiers on the Xerox7 dataset and the UIUCTex dataset. We find that the combination of the Hessian-Laplacian detector and the MSER detector can obtain more discriminative regions, and the SVM classifier with the EMD kernel can achieve the highest classification accuracy. We also propose an EMD spatial kernel which encodes the spatial information. We have evaluated our approach on the Xerox7 dataset, the 4-class VOC2006 dataset and the 4-class Caltech101 dataset. The experimental results demonstrate that (1) the EMD kernel with signature representation achieves higher classification accuracy than other kernels with histogram distribution representation; and (2) the EMD spatial kernel is superior to the single EMD kernel in term of classification accuracy.

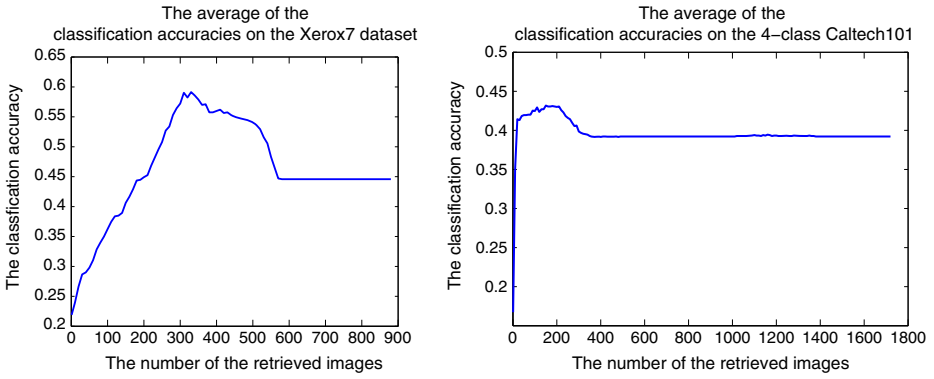


Fig. 11 The averaged classification accuracies obtained by the explicit spatial coding based on the image retrieval method. The left is the obtained classification accuracy on the Xerox7 dataset, and the right is the classification accuracy on the 4-class VOC2006 dataset

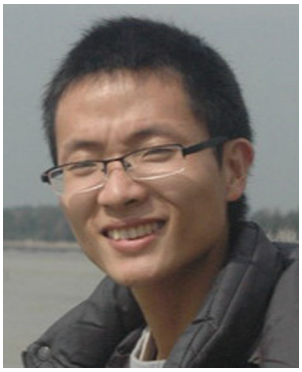
Acknowledgments The authors would like to thank the reviewers for their valuable comments, which greatly helped to improve the quality of the paper. The research work was supported by the Fundamental Research Funds for the Central Universities (2010121067), National Defense Basic Scientific Research program of China under Grant (B1420110155), National Natural Science Foundation of China (61170179), the Special Research Fund for the Doctoral Program of Higher Education of China under Project (20110121110033), and Xiamen Science & Technology Planning Project Fund (3502Z20116005) of China.

References

1. Bay H, Tuytelaars T, Van Gool L (2006) SURF: Speeded Up Robust Features. In proceeding of European Conference on Computer Vision
2. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intel 24:509–522
3. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
4. Cao Y, Wang C, Li Z, Zhang L, Zhang L (2010) Spatial-bag-of-features. In proceeding of Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp 3352–3359
5. Cheng YY, Qu YY, Huang JX, Fang TZ, Lu S, Xie Y (2010) Optimal operations for visual categorization. In proceeding of 2nd International Conference on Internet Multimedia Computing and Service, pp 73–76
6. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In proceeding of ECCV Workshop on Statistical Learning in Computer Vision
7. Farquhar J, Szedmak S, Meng H, Shawe-Taylor J (2005) Improving "bag-of-keypoints" image categorisation. In Technical report, University of Southampton
8. Fergus R, Fei-Fei L, Perona P, Zisserman A (2005) Learning object categories from Google's image search. In proceeding of Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 2, pp 1816–1823
9. Freeman WT, Adelson EH (1991) The design and use of steerable filters. IEEE Trans Pattern Anal Mach Intel 13:891–906
10. Larlus D, Jurie F (2006) Latent mixture vocabularies for object categorization. In proceeding of British Machine Vision Conference
11. Lazebnik S, Schmid C, Ponce J (2005) A sparse texture representation using local affine regions. IEEE Trans Pattern Anal Mach Intel 27:1265–1278
12. Lazebnik S, Schmid C, Ponce J (2005) A maximum entropy framework for part-based texture and object recognition. In proceeding of Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 1, pp 832–838
13. Levina E, Bickel P (2001) The earth mover's distance is the mallows distance: some insights from statistics. In proceeding of Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, vol. 2, pp 251–256
14. Ling HB, Jacobs DW (2005) Deformation invariant image matching. In proceeding of Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 2, pp 1466–1473
15. Liu Y, Rong J, Sukthankar R, Jurie F (2008) Unifying discriminative visual codebook generation with classifier training for object category recognition. In proceeding of Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp 1–8
16. Lowe DG (1999) Object recognition from local scale-invariant features. In proceeding of computer vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 2, pp 1150–1157
17. Matas J, Chum O, Urban M, Pajdla T (2002) Robust wide-baseline stereo from maximally stable extremal regions. In proceeding of British Machine Vision Conference
18. Mikolajczyk K, Schmid C (2004) Scale & affine invariant interest point detectors. Int J Comput Vis 60:63–86
19. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intel 27:1615–1630
20. Moosmann F, Triggs B, Jurie F (2006) Randomized clustering forests for building fast and discriminative visual vocabularies. In proceeding of Neural Information Processing Systems
21. Nowak E, Jurie F, Triggs B (2006) Sampling strategies for bag-of-features image classification. In proceeding of European Conference on Computer Vision
22. Perronnin F, Dance C, Csurka G, Bressan M (2006) Adopted vocabularies for generic visual categorization. In proceeding of European Conference on Computer Vision
23. Rothganger F, Lazebnik S, Schmid C, Ponce J (2006) 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. Int J Comput Vis 66:231–259
24. Varma M, Zisserman A (2002) Classifying images of materials: achieving viewpoint and illumination independence. In proceeding of European Conference on Computer Vision, pp 255–271

25. Varma M, Zisserman A (2003) Texture classification: are filter banks necessary? In proceeding of Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 2, pp II-691-8
26. Winn J, Criminisi A, Minka T (2005) Object categorization by learned universal visual dictionary. In proceeding of Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 2, pp 1800–1807
27. Wu Z, Ke W, Isard M, Sun J (2009) Bundling features for large scale partial-duplicate web image search. In proceeding of Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp 25–32
28. Zhang S, Huang Q, Hua G, Jiang S, Gao W, Tian Q (2010) Building contextual visual vocabulary for large-scale image applications. In proceeding of Proceedings of the international conference on Multimedia, Firenze, Italy, pp 501–510
29. Zhang J, Marsza M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: a comprehensive study. Int J Comput Vis 73:213–238

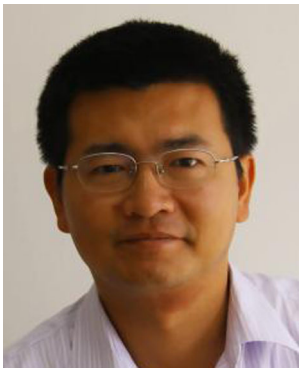
Yanyun Qu received her B.Sc. and M.Sc. degrees in Computational Mathematics from Xiamen University and Fudan University, China, in 1995 and 1998, respectively, and received her Ph.D. degrees in Automatic Control from Xi'an Jiaotong University, China, in 2006. She joined the faculty of Department of Computer Science in Xiamen University since 1998. She was appointed as a lecturer from 2000 to 2007 and was appointed as an associate professor since 2007. She is a member of IEEE and a member of ACM. Her current research interests include pattern recognition, computer vision, image/video processing, machine learning, etc.



Shaojie Wu born in 1988, is currently a graduate at Xiamen University. He was awarded a B.Sc. in Computer Science and Technology Department, Xiamen University in 2010. His research interests mainly lie in the areas of pattern recognition, machine learning, computer vision and related areas.

Han Liu born in 1985, is currently a graduate at Xiamen University. He was awarded a B.Sc. in Mathematical Science, Xiamen University in 2009. His research interests mainly lie in the areas of computer vision and artificial intelligence.

Yi Xie received her B.Sc. and M.Sc. degrees from Xi'an Jiaotong University and received her Ph.D. degree from The Hong Kong Polytechnic University in 2008. Currently, she is an assistant professor with Department of Computer Science in Xiamen University. Her current research interests include image/video processing, system modeling, system simulation and network protocol analysis.



Hanzi Wang is a Distinguished Professor at Xiamen University, China and an Adjunct Professor at the University of Adelaide, Australia. He was a Senior Research Fellow (2008–2010) at the University of Adelaide, Australia; an Assistant Research Scientist (2007–2008) and a Postdoctoral Fellow (2006–2007) at the Johns Hopkins University; and a Research Fellow at Monash University, Australia (2004–2006). He received the Ph.D degree in Computer Vision from Monash University. He was awarded the Douglas Lampard Electrical Engineering Research Prize and Medal for the best PhD thesis in the Department. His research interests are concentrated on computer vision and pattern recognition including visual tracking, robust statistics, object detection, video segmentation, model fitting, optical flow calculation, 3D structure from motion, image segmentation and related fields.