Computers and Electrical Engineering 38 (2012) 917–925

Contents lists available at SciVerse ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Scene categorization based on integrated feature description and local weighted feature mapping ☆

Fengcai Li [a], Guanghua Gu [a,b,*], Chengru Wang [a]

[a] School of Information Science and Engineering, YanShan University, Qinhuangdao 130300, China
[b] Institute of Information Science, Beijing Jiaotong University, Beijing, China

ARTICLE INFO

Article history:
Received 14 August 2011
Received in revised form 25 February 2012
Accepted 27 February 2012
Available online 29 March 2012

ABSTRACT

Local and global features are considerably important features in computer vision and play an important role in scene categorization task. In this paper, an integrated feature description for scene categorization is constructed. First, we extract a type of extended contextual features for scene images that contain the local gradient information and more comprehensive local structural information. Mapping the local features by using improved LLC (Local-constrained Linear Coding) scheme to form the original image representation; Secondly, a set of global features named 'gist' are extracted that provide a statistical summary of the spatial layout properties of the scene; Then, the contextual features and 'gist' features are weighted combined based on their contribution for the integrated feature description, and each image is represented by using LLC scheme. Finally, we perform the scene categorization by libSVM with the HIK (Histogram Intersection Kernel) function. The proposed method achieves a satisfactory average accuracy rate 87.60% on a set of 15-scene categories.

1. Introduction

Scene categorization is a fundamental problem in image understanding. It is a challenging task in computer vision and widely applied in many domains, e.g. image retrieval, video surveillance, medical browser, travel navigation. Designing automatic techniques for improving the scene categorization performance has attracted considerable attentions in recent years.

Scene categorization generally contains four stages: image feature extraction, image representation, classifier training and image categorization. Especially, the first three stages can greatly influence the final categorization performance. Traditional strategies mainly paid more attentions to the global information of the images, e.g. color [1], texture [2], edge response [3], gradient [4] etc. For scene categorization, Olive and Torralba [5–7] proposed a formal approach to build the 'gist' of the scene from global features and provided a statistical summary of the spatial layout properties (naturalness, openness, expansion, depth, roughness, complexity, ruggedness, symmetry) of the scene. The low dimensional global features are based on configurations of spatial scales and estimated without invoking segmentation or grouping operations. These global features may be sufficient for separating scenes with significant differences in the global properties (e.g. living room vs. forest, bedroom vs. highway). However, for the scenes with similar global characteristics (e.g. living room vs. bedroom, tall building vs. street), the global features may be poorly discriminative. Therefore, many methods of local image features extraction have been proposed and show fine classification performance [8–11]. Qin et al. [12] constructed the contextual features based on the local SIFT (Scale-Invariant Feature Transform) features to obtain the new descriptors of the features. The contextual

☆ Reviews processed and approved for publication by Editor-in-Chief Dr. Manu Malek.

* Corresponding author at: School of Information Science and Engineering, YanShan University, Qinhuangdao 130300, China.
E-mail address: guguanghua@ysu.edu.cn (G. Gu).

0045-7906/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compeleceng.2012.02.017

features provide local structural information and the method is called CVW (Contextual Visual Words) in this paper. Though the contextual feature contains local spatial information, it has shortcomings in some degree; on one hand, it lacks global information; the other, the contextual information may be incomplete due to that it only considers the 4-neighbor structural information. Moreover, it is impossible to meet the computation cost and storage requirement with the scale level increasing.

The representation of scene images is the second important step for scene categorization. As one of state-of-the-art techniques, the BoFs (bag-of-features) [13–15] model has been extremely popular in scene categorization and receives extensive considerations in characterizing the images. However, this scheme discards the spatial structural information, which severely limits the descriptive power of the image representation. In order to overcome the shortcoming, Lazebnik et al. [16] proposed an extension of the BoFs model named SPM (spatial pyramid matching). The strategy alleviates the loss of spatial structural information and becomes one of the most successful approaches. Recently, researches focus on the scene categorization using statistical models and probability models [17–19]. These approaches have improved the scene categorization accuracy effectively. However, it is difficult to meet the requirement of computational complexity and storage space.

For the third step, the crucial core of classifier designing is the selection of kernel function. It is better to select suitable kernel functions for different data sets. In scene categorization task, histograms are used in almost every aspect from feature descriptors to image representation. HIK (Histogram Intersection Kernel) [20–23] has been shown to be suitable for comparing the similarity of two histograms in machine learning tasks. It has been further shown to be a conditionally positive definite kernel when the histograms only contain non-negative values, which makes HIK suitable for SVM (Support Vector Machine) classification. Experiments have shown that HIK achieves higher accuracies in SVM classification than linear and RBF (radius-based function) kernel in different fields [23].

In this paper, we extend the scheme in [12] to form contextual features in more comprehensive range at fewer scales (s = 1, 2, 3). Also, a practical scheme called LLC (Locality-constrained Linear Coding) [15] is improved for feature mapping to form the original image representation. To obtain the global information for the final image representation, a set of global features called 'gist' are combined according to the contribution. One, contextual features contain local gradient and structural information, and 'gist' provides global information; the other, SPM scheme also provides global structural information. The final image representation shows more discriminative for scene categorization. At last, we apply SVM classifier with HIK to recognize the scene images. All of the experiments are performed on a complex set of the 15-category scenes.

The rest of the paper is as follows. In Section 2, we describe our approach in details. Experimental results are presented in Section 3. We then discuss the problems of scene categorization in Section 4, and provide the conclusions in Section 5.

2. Our approach

2.1. Local contextual features

The SIFT descriptors [24] are 128-dimensional feature vectors and highly distinctive because of their invariance to image scaling and rotation, and partial invariance to change in illumination and 3D camera viewpoint. According to the comparative evaluation by Fei-Fei and Perona [25], the dense regular grid works better than the other detectors for scene categorization. In our method, we use a regular grid instead of interesting point detection at multi-scales $s = 1, 2, \ldots, S$. Multi-scales means that the SIFT descriptors are sampled with different configurations. More exactly, $s = 1$ denotes the SIFT descriptors of 16×16 pixel patches computed over a grid with spacing of eight pixels; s = 2 denotes the SIFT descriptors of 32×32 pixel patches computed over a grid with spacing of 16 pixels, and so on. Fig. 1 depicts the overall framework of the multi-scales features sampling process.

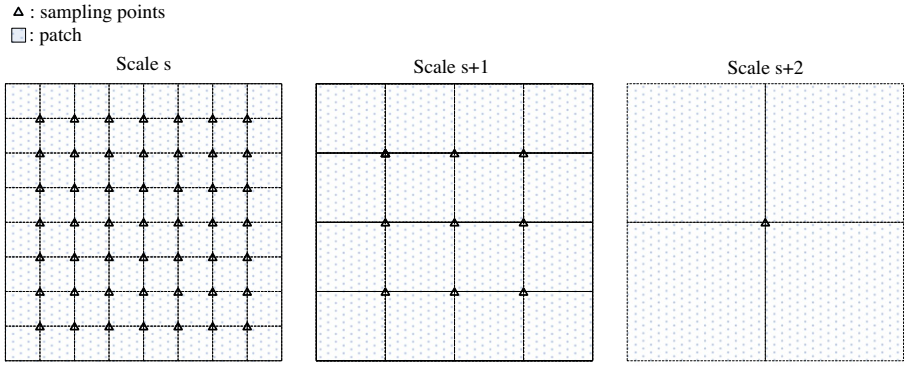

Fig. 1. The overall framework of the multi-scales features sampling process.

The extension of the contextual feature description is illustrated in Fig. 2. Fig. 2(a) shows the range of the former method. Aiming to achieve more comprehensive contextual information in our scheme, here we extend the 4-neigbor of the ROI (region of interest) to 8-neighbor when constructing the contextual features. It is clear in Fig. 2(b) that there are eight raw SIFT features within corresponding spacing pixels away from ROI at each scale. The extended neighbors provide more comprehensive local spatial structural information.

Set $f_1, f_2, \ldots, f_S$ as the 128-dimensional SIFT features at each scale and term them as raw features; $f_{1_c}, f_{2_c}, \ldots, f_{(S-1)_c}$ denote the contextual features at each scale except the last scale S. In details, the contextual features at scale s are constructed in the form as $f_{s_c} = [f_s; w_c \times f_{s+1}; w_n \times f_{s_n}]$. f_s denotes any raw feature of ROI at scale s, f_{s+1} denotes the raw feature at the next scale or the coarser level but has the same centre with f_s, and f_{s_n} are the 8-neighbor of ROI. w_C and w_N are the weight parameters that control the significance of the features from the coarser level and the neighbor regions, respectively, in order to balance the discriminative power and generalization ability of the contextual information. The concept of constructing contextual features is illustrated in Fig. 3. Aiming to improve the processing efficiency and save the storage space, we apply PCA (principle component analysis) to map the high-dimensional contextual features at each scale to low-dimensional space. Each image has multiple BoFs corresponding to the multi-scales. To form multiple vocabularies, the clustering processing is carried out at multi-scales. The details are described in [26].

The extended contextual features contain both local gradient information and more comprehensive local spatial structural information. This type of features are more distinctive for some scenes (e.g. bedroom vs. living room, tall building vs. street) with similar global characteristics but different local spatial structural information.

2.2. Global features extraction

Global feature is one type of the most important features for images. Compared with the local features, the global features can distinguish the scenes with significant difference in the global properties (e.g. coast vs. kitchen). The two types of features can have complementary advantages. The 'gist' features are proposed specifically for scene categorization [5]. In our work, we extract low dimensional global features 'gist' as f_g. The 'gist' descriptor computes the outputs energy of 24 filter banks. These filters are Gabor-like filters turned to eight orientations at four different scales. The square output of each filter is then averaged on a 4×4 grid.

2.3. Features mapping

Image representation based on BoFs usually contains four steps: features extraction, vocabulary formation, features mapping and image representation. The schemes of vocabulary formation include supervised learning, semi-supervised learning and unsupervised learning. Furthermore, the unsupervised learning methods become the major trends, e.g. K-means clustering and GMM (Gaussian Mixture Model) [17]. Features mapping is to quantize the image features into a set of visual words. Then an image is represented by the distribution of the visual words in the vocabulary. In our experiments, we obtain the initial vocabulary using K-means clustering and further apply KSVD (K-means Singular Value Decomposition) [25] to update vocabulary at each scale.

Lazebnik et al. [16] encoded the features descriptors using VQ (vector quantization) on the condition of the least square fitting:

$$\begin{aligned} &\arg\min_{C} \sum_{i=1}^{N} \|x_i - Bc_i\|^2 \\ &\text{s.t.} \|c_i\|_{\ell^0} = 1, \|c_i\|_{\ell^1} = 1, \quad c_i \geqslant 0, \forall i \end{aligned} \tag{1}$$

where $X = [x_1, x_2, \ldots, x_N] \in R^{D \times N}$ denotes the BoFs of an image. $B = [b_1, b_2, \ldots, b_M] \in R^{D \times M}$ is vocabulary with M words, $C = [c_1, c_2, \ldots, c_N]$ represents the set of codes for X. The constraint condition $\|c_i\|_{\ell^0} = 1, \|c_i\|_{\ell^1} = 1, c_i \geqslant 0$ restricts that there

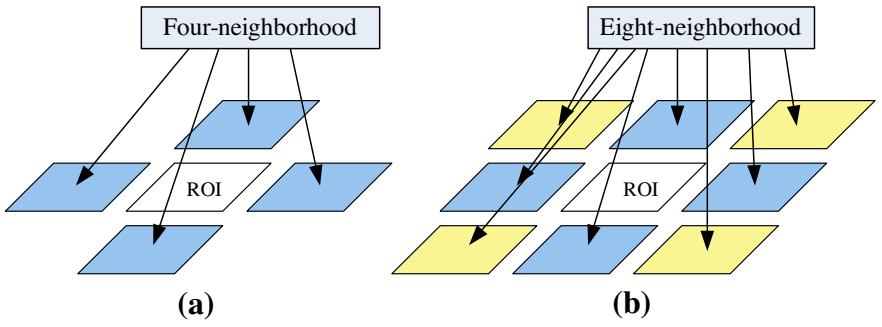

Fig. 2. The extension of our approach against the former method. (a) the 4-neighbor of the ROI in the former method. (b) the 8-neighbor of the ROI in our approach.

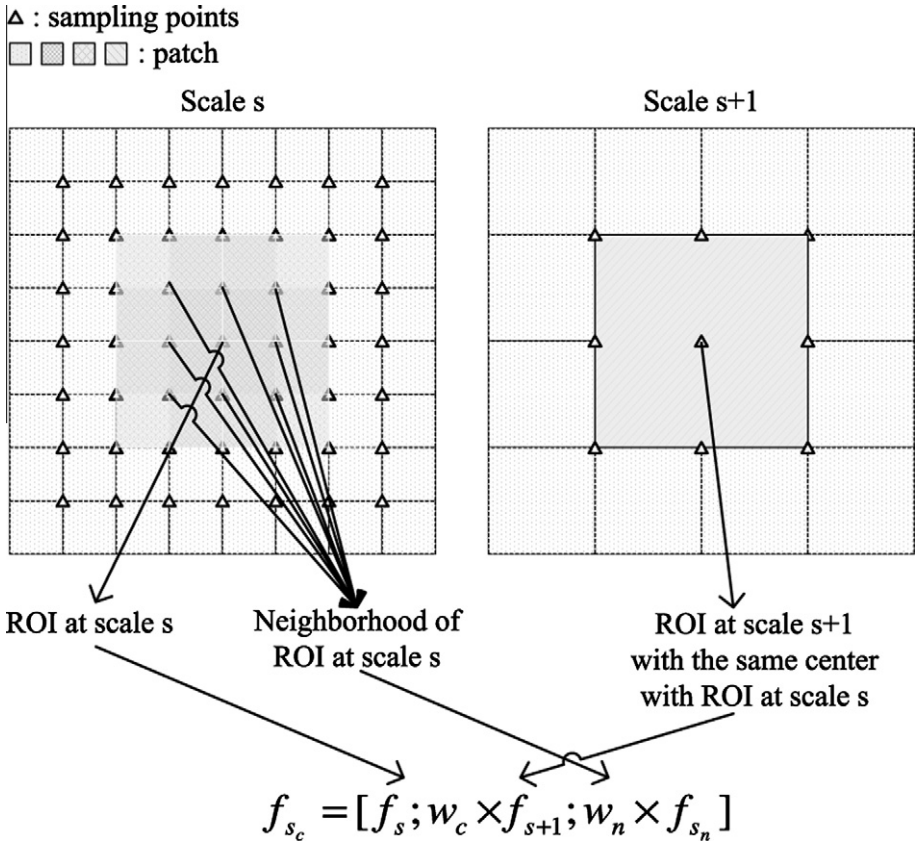

Fig. 3. The schema of the extended contextual features at scale *s*.

is only one non-zero element in each code c_i and the single non-zero element is 1 by searching the nearest neighbors. So this mapping scheme easily losses quantization information. In order to improve the performance of codes, Yang et al. [14] proposed to map the descriptors by using a sparse regularization term via relaxing the restrictive cardinality constraint in VQ:

$$\begin{aligned} &\arg\min_C \sum_{i=1}^{N} \|x_i - Bc_i\|^2 + \lambda\|c_i\| \\ &\text{s.t.} \|b_m\| \leqslant 1, \quad \forall m = 1, 2, \ldots, M \end{aligned} \tag{2}$$

Each descriptor is expressed by the linear combination of all the words in $B = [b_1, b_2, \ldots, b_M]$, $C = [c_1, c_2, \ldots, c_N]$ denotes the corresponding coefficients of linear combination. Compared with VQ, this method greatly reduces the quantization error but unfortunately it also losses the correlation between codes. Wang et al. [15] further improve the mapping distinction by LLC. They use local-constraint instead of sparse-constraint to insure the codes more distinctive than SC (sparse coding) and guarantee the codes sparseness with less quantization error simultaneously.

$$\begin{aligned} &\min_C \sum_{i=1}^{N} \|x_i - Bc_i\|^2 + \lambda\|d_i \odot c_i\|^2 \\ &\text{s.t.} 1^T c_i = 1, \quad \forall i \end{aligned} \tag{3}$$

$$d_i = \exp\left(\frac{\text{dist}(x_i, B)}{\sigma}\right) \tag{4}$$

Table 1
The accuracy rates before and after updating the weights.

Features	Before adapting the weights (%)	Weights	After adapting the weights (%)
Combination	86.83	–	87.60
Ignore 'gist'	85.07	$w_c = 0.55$	–
Ignore c_SIFT	71.60	$w_g = 0.45$	–

The constrained condition $1^T c_i = 1$ follows the shift-invariant requirements of the LLC. $\mathrm{dist}(x_i, b_j)$ in $\mathrm{dist}(x_i, B) = [\mathrm{dist}(x_i, b_1), \mathrm{dist}(x_i, b_2), \ldots, \mathrm{dist}(x_i, b_M)]$ denotes the Euclidean distance of descriptor x_i and word b_j. σ is used for adjusting the weight decay speed for the locality adaptor. The technology of searching K nearest neighbors of x_i as the local bases B_i in practice not only satisfies the requirements of mapping, but also greatly reduces the computation complexity.

As we all known, for image representation, different pooling functions construct different image statistics. Experimental results show that the 'max-pooling' is better than 'sum-pooling' and other alternative pooling methods when forming the sub-region representation. The pooled feature is more robust by 'max-pooling' and empirically justified by many algorithms applied to image categorization. However, the single maximum pooling method may be so absolute that the pooled features can lose amounts of significant information. In this paper, we extend the 'max-pooling' to 'T max-pooling', that is, we take the weighted linear combination of the first T maximums instead of the maximum. Denote the pooled features of the region as a matrix $C = [c_1, c_2, \ldots, c_K]$, and K is the number of the local descriptors in the region. Each column of C corresponds to the responses of all the local descriptors to the item in vocabulary. Then elements of each row are arranged in descending order. We select the first T columns as the input of our method. The i-th element of the region representation $Z = [z_1, z_2, \ldots, z_M]$ is formed as:

$$z_i = \sum_{j=1}^{T} |C_{ij}| * 2^{T-j+1}, \quad i = 1, 2, \ldots, M \tag{5}$$

Normalization:

$$\bar{z}_i = \frac{z_i}{\sum_{j=1}^{T} 2^{T-j+1}}, i = 1, 2, \ldots, M \tag{6}$$

Based on the extended contextual features and improved LLC technology, each image representation is generated as $R_c = [\bar{Z}_1; \bar{Z}_2; \ldots; \bar{Z}_J]$, where J denotes the number of all sub-regions in all pyramid levels. Then we combine the contextual features and global features to form the final image representation $R = [w_c R_c; w_g f_g]$, where f_g denotes the 'gist' features. w_c and w_g are the weighting parameters that control the significance of the contextual features and global features. The weighting parameters are defined as follows:

$$w_c = \frac{r_c}{r_c + r_g} \tag{7}$$

$$w_g = \frac{r_g}{r_c + r_g} \tag{8}$$

We initially carry out two sets of experiments to determine the weighting parameters w_c and w_g. One set only takes the contextual SIFT features ignoring 'gist' to get the categorization accuracy rate r_c; the other takes 'gist' features ignoring contextual features, categorization accuracy rate is r_g. The detailed results of the parameters evaluation are exhibited in Table 1.

2.4. Classifier Training

For scene categorization, SVM is the preferred classifier as a powerful technology in machine learning. In this procedure, kernel selection determines the categorization performance. HIK and SVM are shown to be very effective in dealing with histograms, which have achieved higher category accuracy [23].

The HIK is defined as

$$k_{\mathrm{HI}}(x, y) = \sum_{j=1}^{D} \min(x_j, y_j) \tag{9}$$

HIK has been designed to be a positive definite kernel on non-negative real-valued vectors not limited for non-negative integers. Image representation with histograms formed in this paper just meets the requirement, so we choose SVM with HIK to complete the training and testing.

3. Experimental results

3.1. Datasets and setup

Our test dataset provided by Lazebnik et al. [16] contains 15 natural scene categories with 4485 images: highway (260 images), inside city (308 images), tall building (356 images), street (292 images), suburb (241 images), forest (328 images), coast (360 images), mountain (374 images), open country (410 images), bedroom (216 images), kitchen (210 images), living room (289 images), office (215 images), industrial (311 images) and store (315 images). The first 13 categories were from Li and Perona [27] and the first eight were original collected by Oliva and Torrala [5]. There are $210 \sim 410$ images in each category with size about 300×250. Gray version of the images is used in our experiments.

In our experiments, 100 images are chosen randomly from each category and divided to two separate image sets, i.e. 50 for training and 50 for testing, respectively. We extract SIFT descriptors at three scales i.e. $S = 3$, and construct contextual features at scales 1, 2, 3. Ninety five percent information capacity is reserved when using PCA to reduce the dimension. K-means algorithm and KSVD are applied to generate category vocabulary at each scale. We implement the features mapping based on our improved LLC but set *pyramid level* = [1,2]. The 512-dimensional global features 'gist' are extracted using the code provided by Oliva and Torralba [5].

3.2. Results

To determine the weighted parameters when combining the contextual features c_SIFT and global features 'gist', two individual experiments are carried out. Table 1 shows the average accuracy rates over 10 trails before and after updating the weights. c_SIFT denotes the extended contextual SIFT features. The result $r_c = 85.07\%$ in second row is the categorization rate when taking only the contextual features c_SIFT. When taking only the global features 'gist', the rate $r_g = 71.60\%$. The combination of them without weights shows the result 86.83% in the first row. Our approach obtains the average performance 87.60% after combining the contextual features and global features 'gist' linearly according their weights, which are obtained via Eqs. (7) and (8) shown in the third column. Experiments results in the Table 1 show that the combination without weights outperforms the two single features. And the rate of weighted linear combination is higher than the combination without weights. The detailed category accuracy confusion matrix over 10 trails with $T = 3$ using the weighted linear combination for 15 scene categories is shown in Fig. 4.

Note that the vocabulary size is a key factor of the system performance for the pattern recognition based on the vocabulary model. We evaluate our approach with different vocabulary sizes and different number of maximums when representing the sub-regions. Set k_sc1 to denote the size of each category vocabulary at scale 1, k_sc2 the size of each category vocabulary at scale 2 and T is the number of maximums. Table 2 shows the results when the vocabulary size at scale 2 is equal to that at scale 1, that is $k\_sc1 = k\_sc2 = \{40, 60, 80, 100, 200\}$ and $T = \{1, 2, 3, 5, 10\}$. Results on Table 3 correspond to the situation when the vocabulary size at scale 2 is half of that at scale 1, that is $k\_sc1 = \{200, 100, 80, 60, 40\}$, $k\_sc2 = k\_sc1/2 = \{20, 30, 40, 50, 100\}$ and $T = \{1, 2, 3, 5, 10\}$. As shown in Table 2, the performance increases gradually when the vocabulary size grows from 40 to 100 in each column. The best result appears when $k\_sc1 = k\_sc2 = 100$. Then the performance decreases when the vocabulary size continues to increase. Comparing the results in each row, our approach provides the highest rates when $T = 3$. It is clear in Table 3 that each categorization rate is higher than that in Table 2 on the corresponding position. The method proposed in this paper obtains the best performance 87.60% when $k\_sc1 = 100$,

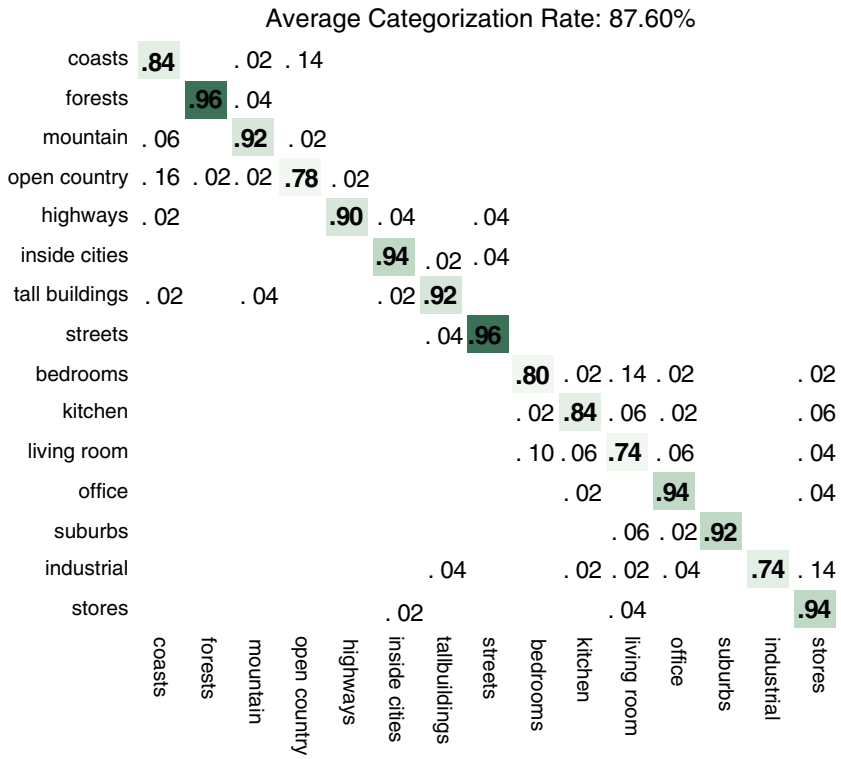

Fig. 4. The confusion matrix of 15 scene categories.

Table 2
The performance when the vocabulary size at scale 2 is equal to that at scale 1.

k_sc1	k_s c2	$T = 1$	$T = 2$	$T = 3$	$T = 5$	$T = 10$
40	40	84.25	84.50	84.76	84.33	84.20
60	60	84.53	84.68	85.00	84.32	84.15
80	80	85.28	85.36	85.45	85.25	85.10
100	100	85.54	85.60	85.73	85.50	85.20
200	200	85.07	85.20	85.47	85.45	85.38

Table 3
The performance when the vocabulary size at scale 2 is half of that at scale 1.

k_sc1	k_sc2	$T = 1$	$T = 2$	$T = 3$	$T = 5$	$T = 10$
40	20	83.60	84.00	84.36	84.03	83.57
60	30	84.67	85.00	85.32	85.20	84.78
80	40	85.45	85.72	86.40	86.15	85.92
100	50	85.96	86.72	87.60	86.40	86.33
200	100	85.20	85.49	86.00	85.78	85.45

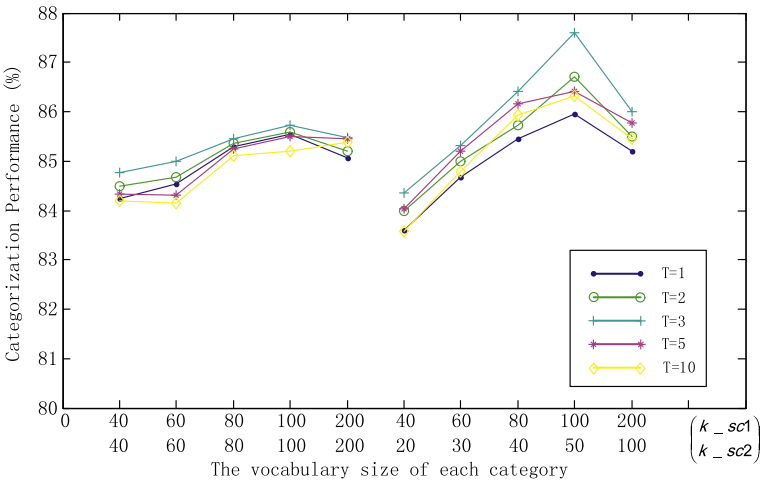

Fig. 5. The relationships of the categorization performance and parameters.

$k\_sc2 = 50$ and $T = 3$. To show the relationship between the categorization performance and parameters, the performance curves are shown in Fig. 5.

The experimental results illustrate the higher performance of our approach on scene categorization. To validate the superiority of our technology, Table 4 shows the performance comparison between our approach and other methods. From Table 4, the result in [5] is 69.35%, which is just based on the global features without local information. Following [14] their own baseline, the linear ScSPM (SPM based on sparse coding) algorithm achieves accuracy of 80.28%. Lazebnik et al. introduced SPM in [16] to incorporate the spatial information with histogram representation and reported the classification rate of 81.40% using SVM with nonlinear histogram intersection kernel. In [12], Qin et al. implemented scene categorization via CVW and provided average rate 85.16%. HG (Hierarchical Gaussianization) representation in [17] achieved a higher performance of 85.2% in accuracy by nearest centroid classifier. GG (Global Gaussian) approach was proposed in [19] based on SURF (Speeded Up Robust Features) descriptors; it incorporated the local and global information and showed the superior performance 86.1%. Our representation incorporates the local and global features, local structural and global spatial information and achieves the best scene categorization 87.60%.

4. Discussions

We analyze the effectiveness for two aspects: the significance of the combination of local and global features; the significance of the weights. In Table 1, the performance of the combination without weights outperforms the other single representation, which proves the significance of the combination. In addition, after updating the weights, the weighted linear

Table 4
The performance comparison of our approach and other alternative methods.

Approaches	Categorization rate (%)
'gist' [5]	69.35 ± 1.34
ScSPM [14]	80.28 ± 0.93
SPM [16]	81.40 ± 0.50
CVW [12]	85.16 ± 1.62
HG [17]	85.2
GG [19]	86.1
Our approach	87.60 ± 0.35

combination outperforms the representation without weights further, which proves the second issue. As details shown in Fig. 4, the classification performance of our method is satisfactory. There are nine scene categories accuracy rates above 90% (such as forest, mountain, highway, inside city, tall building, street, office, suburb and store). Even the indoor scene categories also are distinguished effectively by our approach. Especially for the kitchen category, the categorization rate achieves 84%.

Also, we discuss the relationship between the performance of our approach and the parameters. One is about the vocabulary size; the other is about the selection of maximum number. Observing the performance in each row of Table 2 and Table 3, the best result in each row is obtained when the number of maximums $T = 3$. In other words, the performance will degrade when T is smaller or larger. Comparing the results of two columns when $T = 1$ and $T = 3$, the latter are higher than the former. The phenomenon proves that the weighted combination of T maximums makes image representation more discriminative and powerful than that only using the single maximum. Furthermore, the larger value of T will weaken the discrimination of our representation.

Compared the results in the same position of Table 2 and Table 3, it is obvious that most of the results in Table 3 are better than that in Table 2. As known, the number of features at scale 2 is much smaller than that at scale 1, so it is reasonable that the vocabulary size at scale 2 is appropriately shorter than that at scale 1. Therefore, we obtain the perfect performance when the vocabulary size at scale 2 is half of that at scale 1 in Table 3. The best result is obtained when $k\_sc1 = 100$, $k\_sc2 = 50$, and $T = 3$. It indicates that the size of vocabulary largely affects the performance as shown in Fig. 5.

In our experiments, the initial representations are formed based on the extended contextual features by the improved LLC technology, then the final image representations are constructed using the weighted linear combination of initial representation and global features. Our approach achieves the best performance. The other alternative algorithms in Table 4 just include part of the information. Relatively speaking, our method incorporates the local and global features, local structural and global spatial information, which guarantees the robustness and discrimination of the representation and show the satisfied effectiveness.

5. Conclusions

For scene categorization, this paper has extended the contextual features and formed the initial image representation using the improved LLC technology. The extended contextual features provide the local property of the ROI and richer contextual property from the coarser and neighborhood regions. Our strategy improved the robustness and distinctiveness of the feature descriptors. The improved features coding method LLC is more accurately based on the weighted linear combination of multiple codes. The initial image representation of contextual features formed by applying the improved LLC technology reduced the ambiguities and errors greatly. Additionally, we also have explored the influence of the maximum number and codebook size for the categorization performance. Then construct the final image representation based on the weighted linear combination of the initial image representation and the global feature based on their contribution to the integrated image representation. The proposed method has powerful distinctiveness because it provides more comprehensive information, local and global features, local structural and global spatial information. The more comprehensive image representation further enhances the categorization performance and the experimental results have illustrated effectiveness of our approach on scene categorization. The comparison of our method and other state-of-the-art technologies further has revealed the superiority of our approach.

However, the computational complexity and storage space still be difficult issues to solve and can not achieve a desired level. Additionally, the performances of some scene categories are still poor, e.g. industrial scenes, because it contains both outdoor and indoor scenes. Reducing the computational complexity and storage space, and improving the performance of the scene categorization are the two aspects in our future work.

Acknowledgements

We would like to thank the anonymous reviewers for valuable comments. This work was partly supported by Natural Science Foundation of China (No. 61025013 and No. 61172129), Sino-Singapore JRP (No. 2010DFA11010), Fundamental

Research Funds for the Central Universities (No. 2009JBZ006), and Program for Plan of Science and Technology of Qinhuangdao City (201101A084).

References

[1] Szummer M, Picard R. Indoor-outdoor image classification [C]. In: IEEE international Workshop on Content-Based Access of Images and Video Databases, January 13, 1998. p. 42–51.
[2] Chang E, Goh K, Sychay G, et al. CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines [J]. IEEE Trans Circ Syst Vid 2003;13(1):26–38.
[3] Vailaya A, Jain A, Zhang HJ. On image classification: city images vs. landscapes [C]. In: IEEE Workshop on Content-Based Access of Images and Video Libraries, Santa Barbara, CA, USA, June 21, 1998. p. 3–8.
[4] Villamizar M, Scandaliaris J, Sanfeliu A, Andrade-Cetto J. Combining color-based invariant gradient detector with HoG descriptors for robust image detection in scenes under cast shadows [C]. In: IEEE International Conference on Robotics and Automation, Kobe, Japó, May 12–17, 2009. p. 1997–2002.
[5] Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope [J]. Int J Comput Version 2001;42(3):145–75.
[6] Oliva A. Gist of Scene [C]. In: The Encyclopedia of Neurobiology of Attention. San Diego: Elsevier; 2005. p. 251–6.
[7] Oliva A, Torralba A. Building the gist of a scene: the role of global image features in recognition [J]. Prog Brain Res Vis Percept 2006;155:23–36.
[8] Safayani Mehran, Shalmani Mohammad Taghi Manzuri. Three-dimensional modular discriminant analysis (3DMDA): a new feature extraction approach for face recognition [J]. Comput Electr Eng 2011;37(5):811–23.
[9] Willamowski J, Arregui D, Csurka G, Dance CR. Categorizing nine visual classes using local appearance descriptors [C]. In: International Conference on Pattern Recognition Workshop on Learning for Adaptable Visual Systems, Cambrige, UK, Aug. 22, 2004.
[10] Grauman K, Darrell T. Pyramid match kernels: Discriminative classification with sets of image features [C]. In: Proceedings of IEEE International Conference on Computer Vision, vol. 2, Beijing, China, October 21, 2005. p. 1458–1465.
[11] Fei-Fei L, Perona P. A Bayesian hierarchical model for learning natural scene categories [C]. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, San Diego, CA, USA, June 20–25, 2005. p. 524–531.
[12] Qin J, Yung NHC. Scene categorization via contextual visual words [J]. Pattern Recogn 2010;43(5):1874–88.
[13] Gemert JC, Geusebroek JM, Veenman CJ, Smeulders AW. Kernel Codebooks for Scene Categorization [C]. In: Proceedings of the 10th European Conference on Computer Vision: Part III, vol. 5304. Marseille, France Springer-Verlag, January 22–25, 2008. p. 696–709.
[14] Yang J, Yu K, Gong Y, Huang TS. Linear spatial pyramid matching using sparse coding for image classification [C]. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, June 20-25, 2009. p. 1794–1801.
[15] Wang J, Yang K, Yu F. Locality-constrained Linear Coding for image classification [C]. In: IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, June 13–18, 2010. p. 3360–67.
[16] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories [C]. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, New York, NY, USA, June 17–22, 2006. p. 2169–2178.
[17] Zhou X, Cui N, Li Z, Liang F, Huang TS. Hierarchical Gaussianization for Image Classification [C]. In: IEEE International Conference on Computer Vision, Kyoto, Japan, September 29, 2009–October 2, 2009. p. 1971–77.
[18] Zhou X, Zhuang XD, Tang H, Johnson MH, Huang TS. A novel gaussianized vector representation for natural scene categorization [C]. In: International Conference on Pattern Recognition, Tampa, FL, USA, December 8–11, 2008. p. 1–4.
[19] Hideki Nakayama, Tatsuya Harada, Yasuo Kuniyoshi. Global Gaussian approach for scene categorization using information geometry [C]. In: IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, June 13–18, 2010. p. 2336–43.
[20] Swain MJ, Ballard DH. Color indexing [J]. Int J Comput Vision 1991;7(1):11–32.
[21] Maji S, Berg AC, Malik J. Classification using intersection kernel support vector machines is efficient [C]. In: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, June 23–28, 2008. p. 1–8.
[22] Maji S, Berg AC. Max-margin additive classifiers for detection [C]. In: IEEE International Conference on Computer Vision, Kyoto, Japan, July 13, 2009. p. 40–7.
[23] Jianxin Wu. A fast dual method for HIK SVM learning[C]. In: European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010. p. 552–65.
[24] Lowe DavidG. Distinctive image features from scale-invariant keypoints [J]. Int J Comput Vision 2004;60(2):91–110.
[25] Aharon M, Elad M, Bruckstein A. K-SVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representation [J]. IEEE Trans Signal Process 2006;54(11):4311–22.
[26] Gu Guanghua, Li Fengcai, Zhao Yao, Zhu Zhenfeng. Scene Classification Based on Spatial Pyramid Representation by Superpixel Lattices and Contextual Visual Features [J]. Opt Eng 2012;51(1):017201.
[27] Fei-Fei L, Perona P. A bayesian hierarchical model for learning natural scene categories [C]. In: IEEE Society Conference on Computer Vision and Pattern Recognition, vol. 2, San Diego, CA, USA, June 20–25, 2005. p. 524–531.

Fengcai Li received her B.Sc degree from Yanshan University in 2009 and now is M.Sc student of Circuit and System, Yanshan University, China. Her research interests include digital image processing and pattern recognition.

Guanghua Gu received his B.Sc and M.Sc degrees from Yanshan University, China, in 2001 and 2004, respectively. He is currently a Ph.D candidate of Beijing Jiaotong University. He is working in Yanshan University, China. His research interests include image classification, image recognition, and image analysis.

Chengru Wang is a Professor of Yanshan University, China. His current research interests include image processing and image analysis.