Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Beyond bag of latent topics: spatial pyramid matching for scene category recognition*

Fu-xiang LU‡[1], Jun HUANG[2]

([1]*School of Information Science & Engineering, Lanzhou University, Lanzhou 730000, China*)

([2]*Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China*)

E-mail: lufux@lzu.edu.cn; huangj@sari.ac.cn

Received Mar. 7, 2015; Revision accepted July 14, 2015; Crosschecked Sept. 21, 2015

Abstract: We propose a heterogeneous, mid-level feature based method for recognizing natural scene categories. The proposed feature introduces spatial information among the latent topics by means of spatial pyramid, while the latent topics are obtained by using probabilistic latent semantic analysis (pLSA) based on the bag-of-words representation. The proposed feature always performs better than standard pLSA because the performance of pLSA is adversely affected in many cases due to the loss of spatial information. By combining various interest point detectors and local region descriptors used in the bag-of-words model, the proposed feature can make further improvement for diverse scene category recognition tasks. We also propose a two-stage framework for multi-class classification. In the first stage, for each of possible detector/descriptor pairs, adaptive boosting classifiers are employed to select the most discriminative topics and further compute posterior probabilities of an unknown image from those selected topics. The second stage uses the prod-max rule to combine information coming from multiple sources and assigns the unknown image to the scene category with the highest 'final' posterior probability. Experimental results on three benchmark scene datasets show that the proposed method exceeds most state-of-the-art methods.

Key words: Scene category recognition, Probabilistic latent semantic analysis, Bag-of-words, Adaptive boosting
doi:10.1631/FITEE.1500070 **Document code:** A **CLC number:** TP391.4

1 Introduction

With the exponential growth on high quality digital images, the need of semantic scene category recognition is becoming increasingly important to support effective image database indexing and retrieval (Zhang *et al.*, 2006). However, scene category recognition is one of the most challenging problems in computer vision, especially in the presence of intra-class variation, clutter, occlusion, and illumination changes. On the one hand, a scene category recognition method must generalize across all possible instances of certain categories; on the other hand, it should not confuse between scenes of different categories that are quite similar.

Thus far, no one has constructed a scene category recognition system which approaches the performance level of a two-year-old child. However, the combination of computer vision and machine learning techniques has recently led to significant progress. We take inspiration from this progress and aim to contribute to content-based image understanding and analysis by establishing a robust method for the representation and subsequent recognition of scene categories.

‡ Corresponding author

* Project supported by the Fundamental Research Funds for the Central Universities, China (No. lzujbky-2013-41), the National Natural Science Foundation of China (No. 61201446), and the Basic Scientific Research Business Expenses of the Central University and Open Project of Key Laboratory for Magnetism and Magnetic Materials of the Ministry of Education, Lanzhou University (No. LZUMMM2015010)

ORCID: Fu-xiang LU, http://orcid.org/0000-0002-5810-7631

Bag-of-words is currently a popular method for scene category recognition (Lazebnik *et al.*, 2006; Zhang *et al.*, 2006; Lu *et al.*, 2011; Wu and Rehg, 2011). Local descriptors, either at interest points or densely sampled, are firstly extracted, and an image is then considered a bag of visual words. Originally, bag-of-words is orderless; i.e., it retains only the frequencies of the individual descriptors, and discards all information about their spatial layouts. Later, some researchers find that incorporating spatial information among visual words can make a significant performance improvement. Such models have proven effective for scene category recognition (Lazebnik *et al.*, 2006). In addition, the co-occurrence visual word (Qi *et al.*, 2014) can further boost the discriminative power of the bag-of-words feature because the traditional single word method describes a smaller supporting region individually and ignores the spatial relationship among adjacent descriptors.

In a different direction, probabilistic latent semantic analysis (pLSA) (Hofmann, 1999) allows for the extraction of a compact, discriminant representation for accurate scene category recognition, which remains competitive with recently proposed approaches (Quelhas *et al.*, 2007). More importantly, pLSA allows to address issues related to synonymy (different visual words may represent the same scene type) and polysemy (the same visual word may represent different scene types in different contexts). Similar to bag-of-words, pLSA ignores spatial relations existing among the topics, but the performance of pLSA is adversely affected in many cases due to the loss of spatial information (Lazebnik *et al.*, 2006; Lu *et al.*, 2009).

Based on pLSA, in this paper we explore the problem of recognizing images by the scene categories they contain in the case of a large number of scene categories. The contributions of this paper are summarized in the following:

1. Spatial position relationships among latent topics are introduced to pLSA by means of spatial pyramid (SP). Crudely speaking, given an image, it is firstly partitioned into hierarchical blocks. In the sequel, the topic histograms for individual image blocks in spatial grid layout are obtained by using pLSA based on the bag-of-words representation. Finally, concatenating all block-specific topic histograms leads to a 'long vector', the pyramid topic histogram, for representing the whole image. Due to the introduction of spatial layout for topics, the pyramid topic histogram always outperforms standard pLSA.

2. By combining various interest point detectors and local region descriptors, significant improvement can be made for real-world image datasets. It is well known that different types of interest point detectors place their specific emphases on different types of regions in an image. For example, the Harris detector succeeds in detecting corners (Harris and Stephens, 1988) and the Kadir-Brady detector succeeds in detecting salient regions over the entire image (Kadir and Brady, 2001). Similarly, different types of region descriptors describe the support region around an interest point from different viewpoints such as shape, edge, and texture. So, the discriminative power of the pyramid topic histogram partly depends on its specific choices about the interest point detector and local region descriptor involved in the bag-of-words representation. Accepting this reality, one has to apply multiple detectors and descriptors on a given image to attack diverse scene categorization tasks in the case of a large number of scene categories. This leads to multiple pyramid topic histograms, where each one considers a particular detector and a particular descriptor. Evidently, the main difference among different pyramid topic histograms lies in what detector/descriptor pair is used in the bag-of-words model.

3. A two-stage framework is proposed to perform multi-class classification. Since adaptive boosting (AdaBoost) classifiers (Freund and Schapire, 1997) have shown their promise for text classification and visual recognition tasks, in the first stage they are used to select from the individual pyramid topic histograms the topics that are discriminative against scene categories, and further achieve a posteriori probabilities of an unknown image belonging to each of a fixed number of scene categories. Assuming that the individual pyramid topic histograms are statistically independent, in the second stage decision fusion is performed according to the prod-max rule and the unknown image is assigned to the scene category with the highest 'final' a posteriori probability. It is worth highlighting that the prod-max rule is very robust to violations of its independence assumption.

Extensive experimental results on three benchmark scene datasets show that the proposed pyramid topic histogram performs significantly better than

standard pLSA and that the proposed method exceeds most state-of-the-art methods.

2 Probabilistic latent semantic analysis

In text analysis, probabilistic latent semantic analysis (pLSA) (Hofmann, 1999) is frequently used to discover semantic topics in a corpus using the bag-of-words document representation. pLSA is applied to images by using a visual analogue of a word, formed by vector quantizing the descriptor of the local region around the interest point.

Given a collection of training images $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ with code words from a visual code book $\mathcal{W} = \{w_1, w_2, \ldots, w_V\}$, where N is the number of training images and V the number of distinct code words in the visual code book, the training images can be summarized in a $V \times N$ co-occurrence table $\boldsymbol{N} = (N_{in})$ with entry $N_{in} = s(w_i, d_n)$ representing the number of occurrences of word w_i in image d_n.

pLSA is a statistical model that associates a latent (or hidden) variable $z \in \mathcal{Z} = \{z_1, z_2, \ldots, z_R\}$, where R is the number of topics, with each observation (occurrence of a code word in an image). These latent variables, usually called 'topics', are finally used to build a joint probability model over images and code words, defined as

$$\begin{aligned} p(w_i, d_n) &= \sum_{r=1}^{R} p(w_i, d_n, z_r) \\ &= p(d_n) \sum_{r=1}^{R} p(w_i|z_r) p(z_r|d_n), \end{aligned} \tag{1}$$

where the joint probability $p(w_i, d_n, z_r)$ is assumed to have the form of the graphical model shown in Fig. 1a. Marginalizing over topics z_r in Eq. (1) determines the condition probability $p(w_i|d_n)$:

$$p(w_i|d_n) = \sum_{r=1}^{R} p(z_r|d_n) p(w_i|z_r). \tag{2}$$

pLSA introduces a conditional independent assumption: the occurrence of a code word w_i is independent of the image d_n it belongs to, given a topic z_r. As can be seen from Eq. (2), the conditional probability $p(w_i|d_n)$ is expressed as a convex combination of the topic-specific distributions $p(w_i|z_r)$. This corresponds to matrix decomposition as shown in Fig. 1b.

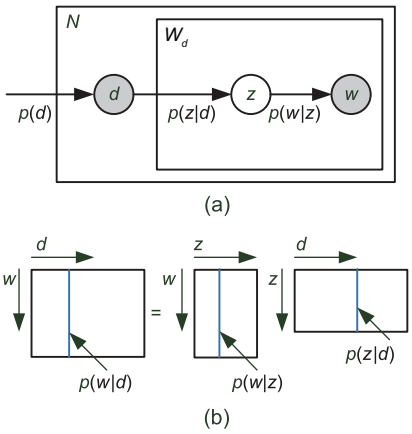

Fig. 1 pLSA model: (a) In the pLSA graphical model representation (plate notation), filled circles indicate observed random variables, unfilled circles indicate unobserved ones, and W_d is the number of words per image; (b) In pLSA, the image-specific code word distribution $p(w|d)$ can be expressed as a convex combination of the topic-specific distributions $p(w|z)$

The parameters of the model are estimated by using the maximum likelihood principle. Specifically, given a set of training images $\mathcal{D}$, the log-likelihood function of model parameters Θ can be expressed by

$$\mathcal{L}(\Theta|\mathcal{D}) = \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} s(w, d) \ln p(z, d, w), \tag{3}$$

where the notation is simplified by dropping the indices from corresponding variables. The optimization is conducted by using the expectation maximization (EM) algorithm. This estimation procedure involves determining the topic-specific distributions $p(w|z)$, which are common to all images, and the mixture coefficients $p(z|d)$, which are specific for each image. Algorithm 1 summarizes the learning procedure of pLSA. Note that $p(d|z)$, not $p(z|d)$, is returned by Algorithm 1. In fact, $p(z|d)$ can be easily obtained without too much mathematics after the convergence of the EM algorithm.

In the testing stage, when observing a novel image d_{test}, the mixing coefficients $p(z|d_{\text{test}})$, which are used to represent the testing image, are computed using the fold-in heuristic. This is achieved by running EM in a similar manner to that used in learning, but now only the coefficients $p(z|d_{\text{test}})$ are updated in each M-step with the learned $p(w|z)$ kept fixed.

3 Pyramid topic histogram

Given a certain pair of detector and descriptor, the construction of pyramid topic histogram $\boldsymbol{x}$ from

Algorithm 1 Learning procedure of pLSA

Require: $\boldsymbol{N} = (s(d,w))$: co-occurrence table; R: number of topics; T: maximum number of iterations; ϵ: a small positive constant

1: Initialize $p(z)$, $p(d|z)$, and $p(w|z)$ to $p^0(z)$, $p^0(d|z)$, and $p^0(w|z)$, respectively

2: $t = 0$

3: **repeat**

4: (E-step) Compute

$$p^t(z|d,w) = \frac{p^t(z)p^t(d|z)p^t(w|z)}{\sum_z p^t(z)p^t(d|z)p^t(w|z)} \tag{4}$$

5: (M-step) Evaluate the following three equations:

$$p^{t+1}(w|z) = \frac{\sum_d s(d,w)p^t(z|d,w)}{\sum_{w,d} s(d,w)p^t(z|d,w)} \tag{5}$$

$$p^{t+1}(d|z) = \frac{\sum_w s(d,w)p^t(z|d,w)}{\sum_{w,d} s(d,w)p^t(z|d,w)} \tag{6}$$

$$p^{t+1}(z) = \frac{\sum_{d,w} s(d,w)p^t(z|d,w)}{\sum_{d,w} s(d,w)} \tag{7}$$

6: Evaluate the log-likelihood function

$$\begin{aligned}\mathcal{L}^{t+1} &= \sum_{d,w} s(d,w) \ln p^{t+1}(z,d,w) \\ &= \frac{s(d,w)p^t(z|d,w)}{\sum_{d,w} s(d,w)}\end{aligned} \tag{8}$$

and log-likelihood difference $\Delta\mathcal{L}^{t+1} = \mathcal{L}^{t+1} - \mathcal{L}^t$

7: $t = t + 1$

8: **until** $\Delta\mathcal{L} < \epsilon$ or $t > T$

9: **return** $p(z)$, $p(d|z)$, $p(w|z)$

an image d involves several main steps (Fig. 2). In brief, a spatial pyramid is first generated from the image, and then interest points (or keypoints) in the blocks of the pyramid are automatically detected. Third, local descriptors are computed over the image regions associated with these points. Fourth, all descriptors are quantized into code words, and all occurrences of each specific code word of the code book are counted to build the bag-of-words representation for the blocks, i.e., co-occurrence tables. The top part of Fig. 2 highlights interest point detection, region description, and vector quantization involved in the bag-of-words computation for a given block. Finally, the topic histograms are computed over the blocks based on pLSA using the EM algorithm and further concatenated to form a long vector (i.e., pyramid topic histogram) to represent the image. In the following, we describe each step in more detail.

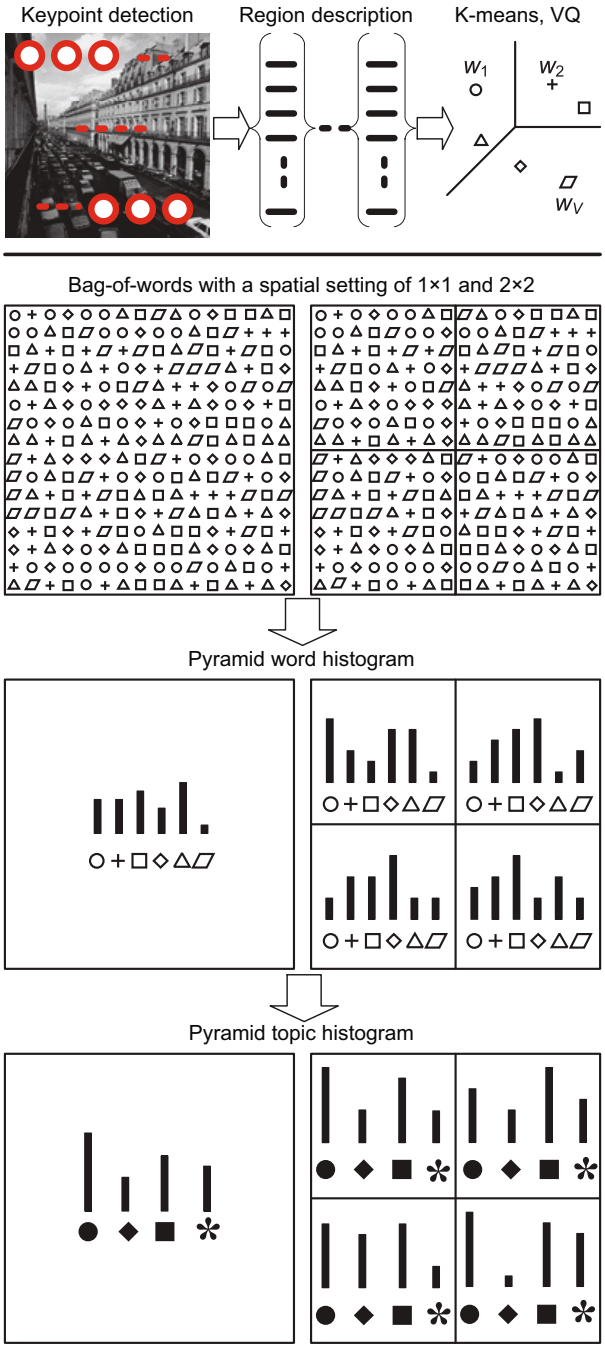

Fig. 2 The main steps of computing the pyramid topic histogram for an image

3.1 Spatial pyramid generation

A spatial pyramid of an image is a collection of decreasing resolution images arranged in the shape of a pyramid. As can be seen from Fig. 3, the base of the pyramid contains a high-resolution representation of the image being processed; the apex contains a low-resolution approximation. As we move up the

pyramid, both size and resolution decrease. Suppose base level D is of size $M \times M$, apex level 1 is of size $(M/2^{D-1}) \times (M/2^{D-1})$, and general level ℓ is of size $(M/2^{D-\ell}) \times (M/2^{D-\ell})$. Then, we divide each level, e.g., level ℓ, into $2^{\ell-1} \times 2^{\ell-1}$ non-overlapping blocks with the same size, which makes a total of $2^{2(\ell-1)}$ blocks at level ℓ. We can easily see from Fig. 3 that all blocks at the different levels of the pyramid contain the same number of pixels.

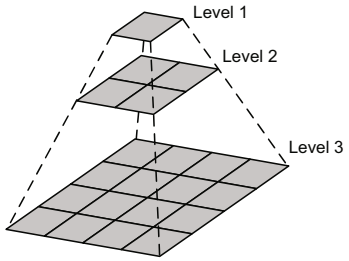

Fig. 3 Spatial pyramid of depth $D = 3$

3.2 Interest point detection

For each of blocks of the pyramid, we extract interest points such as Lowe's difference-of-Gaussian (DoG) points (Lowe, 2004), Harris corner points (Harris and Stephens, 1988), maximally stable extremal regions (MSER) (Matas *et al.*, 2004), and Kadir-Brady salient regions (Kadir and Brady, 2001). These interest point detectors are invariant to different types of geometric and photometric transforms and still discriminative enough to establish correct correspondence (Mikolajczyk and Schmid, 2004). An empirical evaluation of interest point detectors for scene category recognition showed that evenly sampled grid points (sometimes called 'dense points') spaced at a fixed number of pixels, which can be helpful in describing textureless regions such as the sky, are superior to other types of interest point detectors (Li and Perona, 2005).

3.3 Region description

In the description stage, each of the local regions around the detected interest point location is converted into a more compact and stable (invariant) descriptor that can be matched against other descriptors. The descriptors should be distinctive and at the same time robust to changes in viewing conditions. To date, many descriptors have been proposed. For example, scale invariant feature transform (SIFT) (Lowe, 2004), census transform histogram (CENTRIST) (Wu and Rehg, 2011), and self-similarity (SSIM) (Shechtman and Irani, 2007) are frequently used descriptors for scene category recognition. These local region descriptors have been shown to be very effective for scene category recognition (Lu *et al.*, 2011; Wu and Rehg, 2011).

3.4 Vector quantization and bag-of-words representation

Applying the previous two steps just described to each of the blocks of the pyramid for an image, we obtain a set of local region descriptors, which are then vector quantized into visual code words according to the nearest neighbor (NN) rule. The distinct visual code words constituting the code book are obtained by clustering local descriptors from a subset of training images, either using K-means, hierarchical clustering, or randomized k-d trees. The hope here is that the cluster centers (i.e., code words) are meaningful and representative common sub-patterns, such as spots, flat regions, edges, edge ends, and corners. For the distance measure, either the Euclidean distance or histogram intersection distance (HID) can be adopted in the clustering. Despite the fact that the HID for histogram descriptors is more effective in supervised learning tasks (Wu, 2012), the Euclidean distance is more computationally efficient. K-means is used for a similar reason. Another point to mention is that the length of the code book depends on the specific task and needs to be determined using cross validation (CV) in most cases.

In the sequel, all occurrences of each specific code word in the blocks, i.e., the co-occurrence tables, are counted to build the bag-of-words representation for the blocks. Specifically, given a collection of training images $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, we can obtain $\sum_{\ell=1}^{D} 4^{\ell-1}$ co-occurrence tables $\{\boldsymbol{N}_{\ell c} | \ell = 1, 2, \ldots, D;\ c = 1, 2, \ldots, 4^{\ell-1}\}$ with $\boldsymbol{N}_{\ell c}$ denoting the co-occurrence table which summarizes the training images residing in the cth block at level ℓ in the spatial pyramid layout.

3.5 Topic histogram computation

Fitting pLSA on the co-occurrence tables $\{\boldsymbol{N}_{\ell c}\}$, we obtain the block-specific mixing coefficients (i.e., the topic histogram) $p_{\ell c}(z_r|d)$ $(r = 1, 2, \ldots, R)$ for image d, where R (the number of

distinct topics) is assumed to be fixed for different blocks despite their positions. In the end, the feature $\boldsymbol{x}$ representing the image d that we will use is defined by concatenating the topic histograms from all the blocks in the corresponding pyramid:

$$\boldsymbol{x} = (p_{\ell c}(z_r|d)), \tag{9}$$

where $r = 1, 2, \ldots, R$, $\ell = 1, 2, \ldots, D$, and $c = 1, 2, \ldots, 4^{\ell-1}$. This type of feature is referred to as the pyramid topic histogram. Algorithm 2 gives the pseudo-code for computing the pyramid topic histogram from the training images. Note that $\mathcal{D}_{\text{voc}}$ (Line 2, Algorithm 2), a subset of $\mathcal{D}_{\text{train}}$, is used to construct the visual code book.

Because the discriminative power of the pyramid topic histogram depends on its specific choices about the interest point detector and local region descriptor employed in the model of bag-of-words, various types of interest points can be detected and each of interest points can also be represented by various types of local region descriptors. For example, we can detect DoG points, Harris corner points, and MSER, and each of the detected points can be described by SIFT, SSIM, CENTRIST, and histogram of gradients (HOG). This leads to multiple pyramid topic histograms, where each one considers a particular interest point detector and a particular region descriptor. Evidently, the main difference among different pyramid topic histograms lies in what detector/descriptor pair is selected in the model of bag-of-words. As expected, a combination of multiple detectors and descriptors usually achieves better results than even the most discriminative individual detector and descriptor. Obviously, the proposed feature has high flexibility such that other detectors and descriptors can be easily added. Due to a combination of multiple detectors and descriptors, the proposed feature is capable of handling diverse scene categorization tasks.

Here the latent topics may be regarded as object categories (for example, houses and streets) so that any scene image or block containing instances of several objects is modeled as a mixture of latent topics. Thus, the pyramid topic histogram is a type of mid-level feature, which can partly bridge the semantic gap between low-level content (color, shape, and texture) and high-level concepts (scene categories, events, etc.).

Algorithm 2 Computing the pyramid topic histograms of the training images

Require: $\mathcal{D}_{\text{train}} = \{d_1, d_2, \ldots, d_N\}$: training set; V: length of the code book; R: number of topics; D: depth of spatial pyramid

1: $\mathcal{S} \leftarrow$ null
2: **for all** $d_l \in \mathcal{D}_{\text{voc}} \subseteq \mathcal{D}_{\text{train}}$ **do**
3: $\quad \mathcal{F}_l \leftarrow f(d_l)$, where $f(\cdot)$ denotes the keypoint detector
4: $\quad \mathcal{S} \leftarrow \mathcal{S} \cup u(\mathcal{F}_l; d_l)$, where $u(\cdot)$ denotes local descriptor generation
5: **end for**
6: $\mathcal{W} \leftarrow \text{Kmeans}(\mathcal{S}, V)$
7: **for** $n = 1$ to N **do**
8: $\quad \mathcal{F}_n \leftarrow f(d_n)$
9: $\quad \mathcal{U}_n \leftarrow u(\mathcal{F}_n; d_n)$
10: $\quad \mathcal{Q}_n \leftarrow \text{VQ}(\mathcal{U}_n, \mathcal{W})$, where each descriptor in $\mathcal{U}_n$ is finally represented by the corresponding code word index
11: **end for**
12: **for** $\ell = 1$ to D **do**
13: $\quad$ **for** $c = 1$ to $2^{2(\ell-1)}$ **do**
14: $\quad\quad \boldsymbol{N}_{\ell c} \leftarrow$ null
15: $\quad\quad$ **for** $n = 1$ to N **do**
16: $\quad\quad\quad \boldsymbol{N}_{\ell c} \leftarrow \boldsymbol{N}_{\ell c} + \text{hist}_{\ell c}(\mathcal{Q}_n)$, where $\text{hist}_{\ell c}(\mathcal{Q}_n)$ denotes the code word histogram computed from the cth block at level ℓ of P_n, where P_n is the spatial pyramid of image d_n
17: $\quad\quad$ **end for**
18: $\quad\quad$ Given $\boldsymbol{N}_{\ell c}$, for all (i, r, n), apply Algorithm 1 to compute $p_{\ell c}(w_i|z_r)$ and $p_{\ell c}(z_r|d_n)$
19: $\quad$ **end for**
20: **end for**
21: $\boldsymbol{x}_n \leftarrow$ null
22: **for** $n = 1$ to N **do**
23: $\quad$ **for** $\ell = 1$ to D **do**
24: $\quad\quad$ **for** $c = 1$ to $2^{2(\ell-1)}$ **do**
25: $\quad\quad\quad \boldsymbol{x}_n = \text{cat}(\boldsymbol{x}_n, (p_{\ell c}(z_1|d_n), p_{\ell c}(z_2|d_n), \ldots, p_{\ell c}(z_R|d_n)))$, where $\text{cat}(\cdot)$ is the concatenation function
26: $\quad\quad$ **end for**
27: $\quad$ **end for**
28: **end for**
29: **return** $\boldsymbol{x}_n$, $n = 1, 2, \ldots, N$

4 Two-stage classification

Similar to the distributed fusion schemes II–IV of Hu *et al.* (2005), we propose a two-stage framework to perform multi-class classification in this section. In the first stage Adaboost is used to select from the individual pyramid topic histograms the topics that are discriminative against scene categories, and

further compute posterior probabilities of a novel image belonging to each of a fixed number of categories. Then, the prod-max rule is employed to fuse multiple information cues obtained in the first stage.

4.1 First stage: AdaBoost classification

Boosting is a general approach for improving the performance of a given set of weak classifiers and is one of the most powerful techniques. At the heart of a boosting method lies the so-called base classifier, which is a weak classifier. A series of classifiers is then designed iteratively, employing each time the base classifier but using a different subset of the training set, according to an iteratively computed distribution, or a different weighting over the samples of the training set. At each iteration, the computed weighting distribution gives emphasis to the 'hardest' (incorrectly classified) samples. The final, strong classifier is obtained as a weighted average of the weak classifiers. For every specific pyramid topic histogram, AdaBoost (the most popular one among boosting methods) is used to select the discriminative topics and further compute posterior probabilities of an unknown image belonging to each of a fixed number of scene categories. In a two-class case, we take decision stumps as weak classifiers. A decision stump $h(\boldsymbol{x}, k, p, \theta)$ is a threshold function over the kth dimension of feature vector $\boldsymbol{x}$, where θ is a threshold and p is a polarity indicating the direction of the inequality:

$$h(\boldsymbol{x}, k, p, \theta) = \begin{cases} 1, & px_k < p\theta, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Algorithm 3 shows the learning algorithm of AdaBoost for two-class classification. For every specific pyramid topic histogram, multi-class classification is done with the 'one-versus-the-rest' rule: a classifier is learned to separate each class from the rest, and a testing image is assigned the label of the classifier with the highest response.

4.2 Second stage: decision fusion

Prod-max (Lu *et al.*, 2011) is used to combine multiple cues derived from various pyramid topic histograms. For M pyramid topic histograms and K classes, let $\{H_{mk}(\boldsymbol{x}^m)\}$ be AdaBoost classifiers learned from the available training subset by using Algorithm 3. In the testing phase, given an unknown image d_{test} with pyramid topic histograms, $\{\boldsymbol{x}^m\}$, an $M \times K$ matrix $\boldsymbol{T} = (t_{mk})$ can be obtained from the AdaBoost classifiers $\{H_{mk}(\boldsymbol{x}^m)\}$, where the (m, k)th element of $\boldsymbol{T}$, i.e., t_{mk}, denotes the probability of d_{test} belonging to class ω_k based on the mth pyramid topic histogram, i.e.,

$$t_{mk} = H_{mk}(\boldsymbol{x}^m).$$

The final decision function of prod-max for image d_{test} is of the following form:

$$y = f(\boldsymbol{T}) = \underset{k \in \{1,2,\ldots,K\}}{\arg\max} \, [y]_k, \tag{12}$$

where the kth component of vector $\boldsymbol{y}$, i.e., $[y]_k$, is defined as

$$[y]_k = \prod_{m=1}^{M} t_{mk}. \tag{13}$$

Note that to avoid symbol confusion, in Eqs. (12) and (13) the kth component of $\boldsymbol{y}$ is written as $[y]_k$, not usually y_k, because y_k has been used to denote the label of image d_k in Algorithm 3. As seen from Eq. (13), for the prod-max rule, the score of the image d_{test} belonging to a given class ω_k is determined by the product of the scores obtained from the individual pyramid topic histograms.

Algorithm 3 AdaBoost for scene classification

Require: $\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$: training set; T: number of weak classifiers

1: Initialize weights to $w_{0,n} \leftarrow 1/N$, $n = 1, 2, \ldots, N$

2: **for** $t = 1, 2, \ldots, T$ **do**

3: Normalize the weights of training images such that $\sum_{n=1}^{N} w_{t,n} = 1$

4: Learn weak classifier $h_t(\boldsymbol{x}) = h(\boldsymbol{x}, k_t, p_t, \theta_t)$ by minimizing the weighted error:

$$\epsilon_t \leftarrow \min_{k,p,\theta} \sum_{n=1}^{N} w_{t,n} |h(\boldsymbol{x}_n, k, p, \theta) - y_n|,$$

where k_t, p_t, and θ_t are the minimizers of ϵ_t

5: Update the weights: $w_{t+1,n} \leftarrow w_{t,n} \beta_t^{1-e_n}$, where $e_n = 0$ if $\boldsymbol{x}_n$ is classified correctly, and $e_n = 1$ otherwise. Here, $\beta_t = \epsilon_t / (1 - \epsilon_t)$

6: **end for**

7: The strong classifier is defined as

$$H(\boldsymbol{x}) \leftarrow \frac{\sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x})}{\sum_{t=1}^{T} \alpha_t}, \tag{11}$$

where $\alpha_t = \log \dfrac{1}{\beta_t}$

The reason for employing the prod-max rule to combine multiple cues can be described from a statistical point of view. Assume that the individual pyramid topic histograms $\boldsymbol{x}^m$ $(m = 1, 2, \ldots, M)$ are statistically independent. Under this assumption, we can write

$$p(\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^M|\omega_k) = \prod_{m=1}^{M} p(\boldsymbol{x}^m|\omega_k). \tag{14}$$

Posterior probabilities $p(\omega_k|\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^M)$ $(k = 1, 2, \ldots, K)$ are now stated as

$$\begin{aligned} p(\omega_k|\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^M) &\propto p(\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^M|\omega_k)p(\omega_k) \\ &\propto \prod_{m=1}^{M} p(\boldsymbol{x}^m|\omega_k)p(\omega_k) \\ &\propto \frac{\prod_{m=1}^{M} p(\omega_k|\boldsymbol{x}^m)}{p^{M-1}(\omega_k)}. \end{aligned} \tag{15}$$

Furthermore, if the a priori probabilities are equal, that is, $p(\omega_k) = 1/K\ \forall k$, Eq. (15) becomes

$$p(\omega_k|\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^M) \propto \prod_{m=1}^{M} p(\omega_k|\boldsymbol{x}^m). \tag{16}$$

This leads to Eq. (13). Although there is no guarantee about the statistical independence Eq. (14) for most cases, the prod-max rule can be very robust to violations of its independence assumption, and it has been reported to perform well for many real-world datasets (Lu *et al.*, 2011).

5 Experiments and results

5.1 Datasets and setup

We apply the proposed approach to three benchmark scene datasets proposed by Oliva and Torralba (2001), Li and Perona (2005), and Lazebnik *et al.* (2006), respectively. We will refer to them as OT, LP, and LSP, respectively. Among these three datasets, LSP is the most challenging. It is composed of 15 scene categories where each category has 200 to 400 images. LSP includes 4485 images in total, and the average image size is 300×250 pixels. LSP is one of the most complete scene category datasets used in the literature. Fig. 4 shows some example images from the LSP dataset. OT and LP consist of 8 and 13 categories, respectively. All experiments in this study are performed in gray scale, even when color images are available. All experiments are repeated five times with different randomly selected training and testing images, and the average of per-category recognition accuracies is recorded for each run. The final results are reported as the mean and standard deviation of the results from the individual runs.

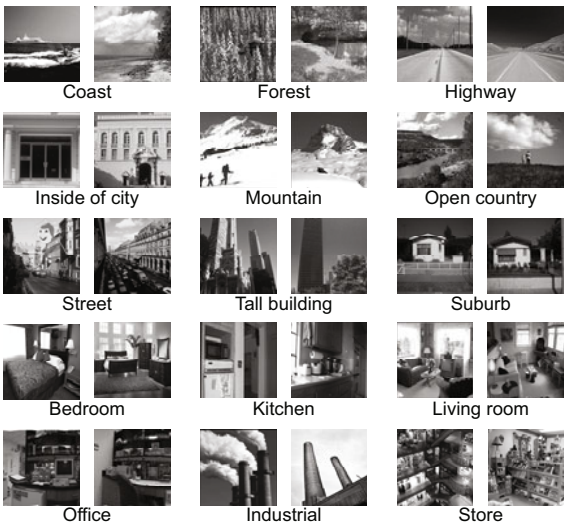

Fig. 4 Example images from the LSP dataset

To evaluate the recognition performance, we stick to the methodologies defined by the designers of the corresponding datasets. Specifically, for all three datasets, 100 images per class are randomly selected for training and the remainder for testing in each run.

5.2 Implementation details

The values of the parameters for pyramid topic histograms over the OT, LP, and LSP datasets are listed in Table 1. In short, six types of descriptors are computed at dense points with a spacing of σ^m pixels for all datasets. At each point, the descriptors are computed over a $\Delta^m \times \Delta^m$ pixel patch. The sizes of all visual code books, V^m, are empirically set according to the comparative evaluation of Lazebnik *et al.* (2006). To incorporate spatial information, spatial pyramids of depth $D^m = 3$ are used except for HOG where $D^m = 4$. The number of topics over each block, R^m, is determined by cross validation, where the step size is set to 5. Note that for a certain detector/descriptor pair, even though we in this study restrict ourselves to using topic histograms that have the same bins (topics) over all blocks, nothing would prevent us from using topic histograms that have different bins over different blocks.

Table 1 The values of the parameters for mPHOTO over the OT, LP, and LSP datasets

m	Detector	Descriptor	R^m	D^m	σ^m	Δ^m	V^m
1	Grid	PATCH	25	3	8	16	200
2		SSIM	40				
3		SIFT	30				
4		CENTRIST	35				
5		HOG	25	4			
6	Grid	CT	20	3	1	3	256

5.3 Experimental results

5.3.1 Comparison between pyramid topic histogram and pLSA

Based on the grid detector and SIFT descriptor, Table 2 shows the results of the pyramid topic histogram over the OT, LP, and LSP datasets as the depth of spatial pyramid increases from 1 to 3. If $D = 1$, the pyramid topic histogram reduces to a standard pLSA. As seen from Table 2, recognition accuracies improve dramatically as we go from $D = 1$ to a multi-level setup for all datasets. For example, if $D = 3$, the pyramid topic histogram achieves recognition accuracies of 84.6%, 80.1%, and 75.4%, which are much higher than the 81.4%, 71.8%, and 65.8% obtained with standard pLSA over the OT, LP, and LSP datasets, respectively. Especially, about 10% improvement is obtained over the LSP dataset.

Table 2 The average of per-category recognition rates over OT, LP, and LSP obtained using the pyramid topic histogram based on the grid detector and SIFT descriptor

D	Recognition rate (%)		
	OT	LP	LSP
1	81.4 ± 0.2	71.8 ± 0.8	65.8 ± 0.8
2	82.3 ± 0.7	75.0 ± 1.1	70.2 ± 0.1
3	84.6 ± 0.1	80.1 ± 0.9	75.4 ± 0.1

5.3.2 Results based on a combination of multiple detectors and descriptors

Table 3 shows the results of pyramid topic histograms based on a combination of six detector/descriptor pairs. For comparison, the results of using each of the individual pyramid topic histograms are also listed. As seen, the pyramid topic histogram with SIFT descriptors performs best for all datasets except that the pyramid topic histogram with SSIM descriptors obtains slightly better results for LSP. For all datasets, heterogeneous pyramid topic histograms combined by the prod-max rule significantly improve the recognition accuracies as compared with a certain pyramid topic histogram. For example, for LSP, the result of the pyramid topic histogram with the most discriminative pair of detector and descriptor (i.e., pyramid topic histogram with the grid point detector and SIFT descriptor) is 75.4%, which is inferior to the prod-max result of 83.7% by a large margin. Similarly, for OT, the recognition accuracy of heterogeneous pyramid topic histograms with a combination of multiple detectors and descriptors is 88.8%, which is much better than 84.6% of the most discriminative pyramid topic histogram.

Table 3 The average of per-category recognition accuracies obtained using each of the individual pyramid topic histograms and their combination for OT, LP, and LSP

Channal number	Recognition accuracy (%)		
	OT	LP	LSP
1	77.2 ± 0.6	69.1 ± 0.5	61.2 ± 1.5
2	83.4 ± 0.8	74.3 ± 0.6	69.2 ± 0.6
3	84.6 ± 0.1	80.1 ± 0.9	75.4 ± 0.6
4	81.0 ± 0.3	77.7 ± 1.0	75.5 ± 0.7
5	80.8 ± 0.6	74.9 ± 0.3	68.1 ± 1.3
6	80.5 ± 0.5	77.7 ± 0.9	75.5 ± 0.4
prod-max	$\mathbf{88.8} \pm 0.2$	$\mathbf{86.7} \pm 0.2$	$\mathbf{83.7} \pm 0.5$

Table 4 shows the resulting confusion matrices for OT, LP, and LSP from one run of using our proposed method. Analyzing the confusion matrices, especially for one among the 15 scene categories, we observe that confusion often occurs between the indoor categories such as bedroom and living room. This can be explained by the fact that there exist similar components and similar configuration in most indoor categories. Mistakes are also made between some natural categories such as forest and mountain, reflecting their sharing of similar textures and the presence of forest in some mountain images.

Table 5 compares our method with the state-of-the-art methods for the benchmark datasets. For the OT dataset, the proposed method obtains a recognition accuracy of 88.8%, which is much higher than the result of 83.7% achieved with GIST (Oliva and Torralba, 2001). For the LP dataset, the variant of latent Dirichlet allocation (LDA) obtains an

Table 4 Confusion matrices demonstrating recognition accuracies (%) for OT (a), LP (b), and LSP (c)

	coa	for	hig	ins	mou	ope	str	tal
coa	85.4	0.4	3.5	0.0	2.3	8.5	0.0	0.0
for	0.0	96.9	0.0	0.0	2.6	0.4	0.0	0.0
hig	2.5	0.0	88.1	5.0	1.9	1.9	0.6	0.0
ins	0.0	0.0	0.0	88.9	0.0	0.0	4.3	6.7
mou	1.1	1.8	0.7	0.0	91.6	3.6	0.0	1.1
ope	8.7	4.5	3.9	0.0	5.5	76.1	1.0	0.3
str	0.0	0.0	2.6	4.2	0.5	0.0	90.1	2.6
tal	0.0	0.4	0.0	5.1	0.4	0.0	0.8	93.4

(a)

	bed	sub	kit	liv	coa	for	hig	ins	mou	ope	str	tal	off
bed	66.4	0.0	3.4	28.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7
sub	0.0	99.3	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0
kit	5.5	0.0	80.0	8.2	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.9	4.5
liv	5.8	0.0	5.8	81.5	0.0	0.5	0.0	1.6	0.5	0.0	1.1	0.5	2.6
coa	0.0	0.0	0.0	0.0	87.3	0.8	0.8	0.0	0.8	10.0	0.4	0.0	0.0
for	0.0	0.0	0.0	0.0	0.0	94.3	0.0	0.0	4.4	1.3	0.0	0.0	0.0
hig	0.6	0.0	0.0	0.0	1.9	0.0	89.4	0.6	0.6	5.6	1.3	0.0	0.0
ins	1.0	0.0	0.0	0.5	0.0	0.0	0.0	87.0	0.0	0.0	4.8	6.3	0.5
mou	0.4	0.0	0.0	0.0	1.8	2.9	0.4	0.0	88.3	5.5	0.4	0.4	0.0
ope	0.0	0.0	0.0	0.3	8.7	4.8	1.3	0.0	2.6	81.3	0.6	0.3	0.0
str	0.0	0.0	0.0	1.0	0.0	0.0	2.1	3.1	0.5	0.0	90.6	2.6	0.0
tal	0.0	0.0	0.8	1.2	0.0	0.4	0.0	4.3	1.6	0.0	1.2	90.6	0.0
off	0.0	0.0	1.7	2.6	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	94.8

(b)

	bed	sub	ind	kit	liv	coa	for	hig	ins	mou	ope	str	tal	off	sto
bed	63.8	0.0	0.0	6.0	25.9	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.9	2.6
sub	0.0	98.6	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.7	0.0
ind	0.5	0.0	65.9	3.8	0.9	0.5	0.0	0.0	3.8	0.5	0.0	1.9	4.7	0.0	17.5
kit	2.7	0.0	0.0	80.0	4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	4.5	7.3
liv	8.5	0.0	0.5	2.6	78.3	0.0	0.0	0.0	1.6	0.5	0.0	0.0	0.5	4.2	3.2
coa	0.0	0.0	0.4	0.0	0.0	86.2	0.4	1.5	0.4	0.8	10.4	0.0	0.0	0.0	0.0
for	0.0	0.0	0.0	0.0	0.0	0.0	96.1	0.0	0.0	3.1	0.4	0.0	0.0	0.0	0.4
hig	0.0	0.0	0.0	0.0	0.0	1.9	0.0	86.9	0.6	2.5	4.4	0.6	0.0	0.0	3.1
ins	0.0	0.0	1.9	1.0	1.0	0.5	0.0	0.0	84.6	0.0	0.0	4.3	3.8	1.0	1.9
mou	0.4	0.0	0.0	0.0	0.0	1.5	3.6	0.4	0.0	90.5	2.2	0.0	0.7	0.0	0.7
ope	0.3	0.0	0.3	0.3	0.3	11.6	3.9	2.3	0.0	4.8	75.2	0.3	0.3	0.0	0.3
str	0.0	0.0	1.0	0.5	0.5	0.0	0.0	3.1	3.1	0.5	0.5	87.5	2.1	0.0	1.0
tal	0.0	0.0	0.0	0.0	2.0	0.4	0.0	0.0	5.5	0.8	0.0	0.8	87.5	0.0	3.1
off	0.0	0.0	0.0	3.5	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	95.7	0.0
sto	0.0	0.0	0.9	5.1	3.3	0.0	0.5	0.5	4.2	0.9	0.0	1.4	1.9	0.0	81.4

(c)

accuracy of 65.2% (Li and Perona, 2005). Classification accuracy for the bag-of-visterms representation is 66.5% (Quelhas *et al.*, 2007). The latent topic histogram based on dense points and the SIFT descriptor with a spatial setting of 1×1, 2×2, and 4×4 gives 75.0% accuracy (Lu *et al.*, 2009). The proposed method achieves the highest accuracy, 86.7%. For LSP, the standard pLSA achieves a classification rate of 65.9% (Lazebnik *et al.*, 2006), and spatial pyramid matching of strong features with a vocabulary of 200 visual words yields an accuracy of 81.1% (Lazebnik *et al.*, 2006). Spatial pyramid matching on the semantic manifold (SPMSM) (Kwitt *et al.*, 2012) represents an image based on the semantic probability simplex, which is augmented with a rough encoding of spatial information, and obtains an accuracy

of 82.3%. The linear distance coding (LDC) method (Wang *et al.*, 2013), which transforms local features of an image into more discriminative distance vectors, yields an accuracy of 82.6%. Wu and Rehg (2011) and Liu and Shah (2007) both reported an accuracy of 83.3%, which is slightly below the accuracy of 83.7% achieved by the proposed method. The pairwise rotation-invariant co-occurrence local binary pattern (PRICoLBP) feature (Qi *et al.*, 2014) leads to an accuracy of 84.3%, which outperforms our method by 0.6%. This can be explained by the fact that the PRICoLBP feature has larger supporting regions than a single visual word used in the bag-of-words model and hence can depict more subtle and complex structures in an image.

Table 5 Average recognition accuracy for OT, LP, and LSP

Dataset	Method	Accuracy (%)
OT	Ours	88.8
	Oliva and Torralba (2001)	83.7
LP	Ours	86.7
	Li and Perona (2005)	65.2
	Quelhas *et al.* (2007)	66.5
	Lu *et al.* (2009)	75.0
LSP	Ours	83.7
	Lazebnik *et al.* (2006)	81.1
	Wu and Rehg (2011)	83.3
	Liu and Shah (2007)	83.3
	Kwitt *et al.* (2012)	82.3
	Wang *et al.* (2013)	82.6
	Qi *et al.* (2014)	84.3

6 Conclusions

We studied feature generation and multi-class classification for natural scene category recognition. With regard to feature generation, we proposed pyramid topic histograms to represent an image. These heterogeneous features aim to improve pLSA in two ways. Given an interest point detector and a local region descriptor, the pyramid topic histogram is used to compute the latent topic histograms of individual image blocks by using pLSA in a spatial grid layout and get them together to create a global scene representation used for the final scene categorization. We found that the pyramid topic histogram consistently outperforms standard pLSA for an arbitrary detector/descriptor pair. For example, on the LSP dataset we observe about 10% improved performance. On the other hand, significant improvement can be made by combining various interest point detectors and local region descriptors involved in the bag-of-words representation.

With regard to multi-class classification, we proposed a two-stage framework. From the individual pyramid topic histograms, in the first stage AdaBoost is used to select the most discriminative latent topics and further compute a posteriori probabilities of an unknown image belonging to each of a fixed number of scene categories. Based on the preliminary results of the first stage, in the second stage the prod-max rule is used to fuse information cues coming from heterogeneous features. An assumption taken by prod-max that the individual features are statistically independent is not satisfied in most cases. In practice, however, this seems not to be a problem and works well in the case of a large number of scene categories.

References

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**(1):119-139. [doi:10.1006/jcss.1997.1504]

Harris, C., Stephens, M., 1988. A combined corner and edge detector. Alvey Vision Conf., p.147-151. [doi:10.5244/C.2.23]

Hofmann, T., 1999. Probabilistic latent semantic indexing. Proc. 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.50-57. [doi:10.1145/312624.312649]

Hu, Z.H., Cai, Y.Z., Li, Y.G., *et al.*, 2005. Data fusion for fault diagnosis using multi-class support vector machines. *J. Zhejiang Univ.-Sci.*, **6A**(10):1030-1039. [doi:10.1631/jzus.2005.A1030]

Kadir, T., Brady, M., 2001. Saliency, scale and image description. *Int. J. Comput. Vis.*, **45**(2):83-105. [doi:10.1023/A:1012460413855]

Kwitt, R., Vasconcelos, N., Rasiwasia, N., 2012. Scene recognition on the semantic manifold. European Conf. on Computer Vision, p.359-372. [doi:10.1007/978-3-642-33765-9_26]

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, p.2169-2178. [doi:10.1109/CVPR.2006.68]

Li, F.F., Perona, P., 2005. A Bayesian hierarchical model for learning natural scene categories. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, p.524-531. [doi:10.1109/CVPR.2005.16]

Liu, J.G., Shah, M., 2007. Scene modeling using co-clustering. IEEE Int. Conf. on Computer Vision, p.1-7. [doi:10.1109/ICCV.2007.4408866]

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, **60**(2):91-110. [doi:10.1023/B:VISI.0000029664.99615.94]

Lu, F.X., Yang, X.K., Zhang, R., *et al.*, 2009. Image classification based on pyramid histogram of topics. IEEE Int. Conf. on Multimedia and Expo, p.398-401. [doi:10.1109/ICME.2009.5202518]

Lu, F.X., Yang, X.K., Lin, W.Y., *et al.*, 2011. Image classification with multiple feature channels. *Opt. Eng.*, **50**(5):057210.1-057210.9. [doi:10.1117/1.3582852]

Matas, J., Chum, O., Urban, M., *et al.*, 2004. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis. Comput.*, **22**(10):761-767. [doi:10.1016/j.imavis.2004.02.006]

Mikolajczyk, K., Schmid, C., 2004. Scale & affine invariant interest point detectors. *Int. J. Comput. Vis.*, **60**(1):63-86. [doi:10.1023/B:VISI.0000027790.02288.f2]

Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.*, **42**(3):145-175. [doi:10.1023/A:1011139631724]

Qi, X.B., Xiao, R., Li, C.G., *et al.*, 2014. Pairwise rotation invariant co-occurrence local binary pattern. *IEEE Trans. Patt. Anal. Mach. Intell.*, **36**(11):2199-2213. [doi:10.1109/TPAMI.2014.2316826]

Quelhas, P., Monay, F., Odobez, J., *et al.*, 2007. A thousand words in a scene. *IEEE Trans. Patt. Anal. Mach. Intell.*, **29**(9):1575-1589. [doi:10.1109/ TPAMI.2007.1155]

Shechtman, E., Irani, M., 2007. Matching local self-similarities across images and videos. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-8. [doi:10.1109/CVPR.2007.383198]

Wang, Z.L., Feng, J.S., Yan, S.C., *et al.*, 2013. Linear distance coding for image classification. *IEEE Trans. Image Process.*, **22**(2):537-548. [doi:10.1109/TIP.2012.2218826]

Wu, J.X., 2012. Efficient HIK SVM learning for image classification. *IEEE Trans. Image Process.*, **21**(10):4442-4453. [doi:10.1109/TIP.2012.2207392]

Wu, J.X., Rehg, J.M., 2011. CENTRIST: a visual descriptor for scene categorization. *IEEE Trans. Patt. Anal. Mach. Intell.*, **33**(8):1489-1501. [doi:10.1109/TPAMI.2010.224]

Zhang, J.G., Marszałek, M., Lazebnik, S., *et al.*, 2006. Local features and kernels for classification of texture and object categories: a comprehensive study. *Int. J. Comput. Vis.*, **73**(2):213-238. [doi:10.1007/s11263-006-9794-4]