Artif Intell Rev (2016) 45:333–368
DOI 10.1007/s10462-015-9448-4

Visual descriptors for scene categorization: experimental evaluation

Xue Wei[1] · **Son Lam Phung**[1] · **Abdesselam Bouzerdoum**[1]

Published online: 17 November 2015
© The Author(s) 2015. This article is published with open access at Springerlink.com

Abstract Humans are endowed with the ability to grasp the overall meaning or the gist of a complex visual scene at a glance. We need only a fraction of a second to decide if a scene is indoors, outdoors, on a busy street, or on a clear beach. In recent years, computational gist recognition or scene categorization has been actively pursued, given its numerous applications in image and video search, surveillance, and assistive navigation. Many visual descriptors have been developed to address the challenges in scene categorization, including the large number of semantic categories and the tremendous variations caused by imaging conditions. This paper provides a critical review of visual descriptors used for scene categorization, from both methodological and experimental perspectives. We present an empirical study conducted on four benchmark data sets assessing the classification accuracy and class separability of state-of-the-art visual descriptors.

Keywords Visual descriptors · Scene categorization · Survey and evaluation · Gist recognition

1 Introduction

Humans can grasp rapidly the overall meaning of a complex visual scene. With a single glance, they can determine whether they are looking at a room, a beach, or a forest (Pavlopoulou and Yu 2010). Viewers need only a fraction of a second to associate a picture with an abstract concept such as girl clapping or busy street (Goldstein 2010). This ability to understand the

✉ Son Lam Phung
phung@uow.edu.au

Xue Wei
xw158@uowmail.edu.au

Abdesselam Bouzerdoum
a.bouzerdoum@uow.edu.au

[1] School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, Northfields Av, Wollongong, NSW 2522, Australia

conceptual meaning of a scene at a glance, regardless of its visual complexity and without attention to details, is known as *gist recognition*. The gist of a scene plays many roles in visual perception. It guides the viewer's attention, aids object recognition, and affects the viewer's recollection of the scene (Snowden et al. 2004).

Understanding gist recognition in humans and replicating this ability in computers has been an ongoing research goal (Fei-Fei et al. 2007b; Loschky et al. 2007; Wei et al. 2010; Walther et al. 2011; Lee and Choo 2013; Linsley and MacEvoy 2014; Malcolm et al. 2014). In the early psychovisual studies, Potter found that a scene is processed rapidly at an abstract level before intentional selection occurs (Potter 1975). Gist recognition is shown to happen in the first 30–300 ms of viewing a scene (Goldstein 2010). This perception occurs much earlier than the time required for recognizing individual objects in the scene. Consequently, gist recognition is considered to rely significantly more on holistic and low-level properties than on the detection of individual objects (Loschky et al. 2007). These low-level properties include color (Castelhano and Henderson 2008; Oliva 2000), edges (Pavlopoulou and Yu 2010), and texture (Renninger and Malik 2003). It was even found that object shape and identity are not necessary for the rapid perception of scenes (Oliva and Torralba 2001).

In computer vision, gist recognition is also known as *scene categorization,* which aims to classify a scene into semantic categories (Lazebnik et al. 2006; Xiao et al. 2014a; Krapac et al. 2011; Yao et al. 2012; Zhao and Xing 2014). The scene could be a static image or dynamic video, and the semantic categories could be indoor versus outdoor, gas station versus restaurant, or slow traffic versus flowing traffic. Scene categorization can be used to provide cues about objects and actions, detect abnormal events in public places, sense dangerous situations, and search for images and video; therefore, it is highly useful for applications in surveillance (Zhang et al. 2008; Gowsikhaa et al. 2014), navigation (Chang et al. 2010; Jia et al. 2014; Khosla et al. 2014; Fuentes-Pacheco et al. 2015), and multimedia (Chella et al. 2001; Maggio and Cavallaro 2009; Huang et al. 2008; Zhou et al. 2010; Dai et al. 2015).

In computational scene categorization, visual descriptors play a central role in recognition performance. A good visual descriptor must possess discriminative power to characterize different semantic categories, while remaining stable in the presence of inter- and intra-class variations caused by photometric and geometric image distortions. Note that previously Douze et al. (2009) compared different visual descriptors for image search, but the compared descriptors were restricted to only GIST descriptor (Oliva and Torralba 2001) and its variants. Mikolajczyk and Schmid (2005) evaluated the local descriptors, such as SIFT-based features, steerable filters, complex filters, and moment invariants. Van de Sande et al. (2010) analyzed mainly SIFT-based color descriptors on the PASCAL VOC Challenge 2007. Xiao et al. (2014a) compared fourteen descriptors on the SUN397 data set. Their comparison also focused on the local features, such as *scale-invariant feature transform* (SIFT) (Lowe 2004), *histogram of oriented gradients* (HOG) (Dalal and Triggs 2005), and *local binary pattern* (LBP) (Ojala et al. 1996).

This paper aims to assess the state-of-the-art visual descriptors for scene categorization of static images, from both methodological and experimental perspectives. The compared descriptors range from biologically-inspired features, local features to global features. The paper is structured as follows. Section 2 reviews visual descriptors used for scene categorization. Section 3 describes the public data sets and evaluation protocols for scene categorization. Section 4 presents results of experimental evaluations on four benchmark data sets. Section 5 gives concluding remarks.

Table 1 List of acronyms used in this paper

Acronym	Definition
BIF	Biologically-inspired features (Song and Tao 2010)
BOP	Bag-of-parts (Juneja et al. 2013)
BoW	Bag-of-words (Fei-Fei et al. 2007a)
CENTRIST	Census transform histogram (Wu and Rehg 2011)
CNN	Convolutional neural network (LeCun et al. 1998; Krizhevsky et al. 2012)
DeCAF	Deep convolutional activation feature (Donahue et al. 2014)
GIST	An abstract representation of the scene (Oliva and Torralba 2001)
HIK	Histogram intersection kernel (Maji et al. 2008)
HMAX	Hierarchical model and X (Riesenhuber and Poggio 1999)
HOG	Histogram of oriented gradients (Dalal and Triggs 2005)
HOG-SPM	HOG with spatial pyramid matching (Lazebnik et al. 2006)
HSOG	Histogram of the second-order gradients (Huang et al. 2014a)
LBP	Local binary pattern (Ojala et al. 1996)
LBP-HF	LBP with Fourier histogram (Ahonen et al. 2009)
LCS	Local color statistic (Clinchant et al. 2007)
OB	Object bank (Li et al. 2010)
PLBP	LBP with pyramid representation (Qian et al. 2011)
RBF	Radial basis function kernel (Scholkopf et al. 1997)
SEV	Sequential edge vectors (Morikawa and Shibata 2012)
SIFT	Scale-invariant feature transform (Lowe 2004)
SIFT-FV	SIFT with Fisher vector (Sanchez et al. 2013)
SIFT-LLC	SIFT with locality-constrained linear coding (Wang et al. 2010)
SIFT-ScSPM	SIFT with sparse coding based spatial pyramid matching (Yang et al. 2009)
SIFT-SPM	SIFT with spatial pyramid matching (Lazebnik et al. 2006)
SURF	Speeded up robust features (Bay et al. 2008)
SVM	Support vector machine (Burges 1998)

2 Visual descriptors

Many algorithms have been developed for scene categorization from static images. For convenience, the list of the most widely used acronyms in this paper is given in Table 1. In most algorithms, visual features are first extracted from the input image, and then classified into semantic categories using a trained classifier, e.g., support vector machines. Hence, feature extraction plays a vital role in scene categorization. The existing approaches for visual feature extraction in scene categorization can be divided into three broad categories: biologically-inspired methods, local features, and global features. Table 2 summarizes the three categories and gives the representative visual descriptors of each category.

2.1 Biologically-inspired feature extraction models

To mimic the gist recognition capability of the human vision, researchers have developed a variety of computational algorithms for scene categorization from static images. For example,

Table 2 Classification of scene categorization descriptors

Approaches	Representative works
Biologically-inspired feature extraction	
Visual cell model	HMAX (Serre et al. 2005)
Layout properties	GIST (Oliva and Torralba 2001)
Deep learning features	Convolutional neural networks (Zhou et al. 2014)
	OverFeat (Sermanet et al. 2014)
	DeCAF Donahue et al. (2014)
Local feature extraction	
Patch-based features	SIFT (Lowe 2004)
	HOG (Dalal and Triggs 2005)
	LBP (Ojala et al. 1996)
	CENTRIST (Wu and Rehg 2011)
	HSOG (Huang et al. 2014a)
Object-based model	Object bank (Li et al. 2010)
Region-based model	Edge vectors (Morikawa and Shibata 2012)
	Bag of parts (Juneja et al. 2013)
Keypoint-based features	SURF (Bay et al. 2008)
	FREAK (Alahi et al. 2012)
	BRISK (Leutenegger et al. 2011)
Global feature formation	
Principal component analysis	SIFT features with PCA (Ke and Sukthankar 2004)
Histogram	Multi-resolution histogram (Hadjidemetriou et al. 2004)
Feature encoding	SPM (Lazebnik et al. 2006)
	ScSPM (Yang et al. 2009)
	LLC (Wang et al. 2010)
	FV (Sanchez et al. 2013; Dixit et al. 2015)

Lee and Mumford (2003) developed a hierarchical Bayesian inference for scene reconstruction based on the *early visual neurons*. Song and Tao (2010) suggested a gist recognition model, where intensity, color, and $C1$ visual features are extracted. Grossberg and Huang (2009) proposed the ARTSCENE system based on gist and texture features for natural scene classification. In the following subsections, we discuss three major biologically-inspired feature extraction approaches: the HMAX model, the GIST model, and the deep learning model.

2.1.1 HMAX model

Riesenhuber and Poggio (1999) proposed a feed-forward architecture, called HMAX, which is inspired by the hierarchical nature of the primate visual cortex. The HMAX model has four layers: two simple layers ($S1$ and $S2$) and two complex layers ($C1$ and $C2$). In layer $S1$, the input image is densely filtered with Gaussian filters at several orientations and scales. The feature maps generated from layer $S1$ are then arranged into filter bands that contain neighboring scale maps with different orientations. After layer $S1$, the spatial maximum values in each filter band are computed at the same orientation to form layer $C1$. Let $P_i, i = 1, \dots, K$, be a dictionary that is learned from samples of layer $C1$. P_i is used as a prototype to represent intermediate-level feature $S2$. The maximum values of $S2$ are computed over all positions and scales, and are used as the output features $C2$, which are shift- and scale-invariant.

The HMAX architecture has the advantages of both template-based features (Viola and Jones 2004) and histogram-based features (Lowe 2004; Dalal and Triggs 2005). HMAX features preserve object geometry similarly to template-based features, and they are robust to small distortions in objects, like histogram-based features. Serre et al. (2007) later used HMAX features for object recognition and scene understanding. In object recognition tasks, the HMAX model outperforms the HOG (Dalal and Triggs 2005), the part-based model (Leibe et al. 2004), and the local patch correlation (Torralba et al. 2004). However, HMAX has a longer processing time than other algorithms, such as the HOG and SIFT.

Several extensions to the HMAX have been proposed. Serre and Riesenhuber (2004) introduced a new HMAX that uses Gabor filters to model simple cell receptive field instead of Gaussian filters. Mutch and Lowe (2006) extended the HMAX by introducing lateral inhibition and feature localization. The extended HMAX achieved an improvement of 14 % in the classification rate compared to the original HMAX, in object categorization on the Caltech-101 data set. Brumby et al. (2009) developed a large-scale functional model based on the HMAX and applied it to detect vehicles in remote-sensing images. Inspired by the HMAX, to describe the scene, Jiang et al. (2010) proposed a new approach that combines features from simple cells and complex cells with the *sparse coding based spatial pyramid matching* (ScSPM) (Yang et al. 2009).

2.1.2 GIST model

Oliva and Torralba (2001) proposed a computational model, known as GIST, for scene categorization. They suggested that images in a scene category possess a similar spatial structure that can be extracted without segmenting the image. The GIST features are the statistical summary of the scene spatial layout. That is, they capture the dominant perceptual properties, such as naturalness, openness, roughness, expansion, and ruggedness, of a scene.

The GIST model can be described as follows. An input image I is first padded, whitened, and normalized to reduce the blocking artifact. Next, the image is processed with a set of multi-scale oriented Gabor filters. The impulse response of a Gabor filter is a Gaussian modulated by a harmonic function:

$$g(x, y) = \cos\left(2\pi \frac{x'}{\lambda} + \Phi\right) \exp\left(\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right), \qquad (1)$$

where $x' = x \cos\theta + y \sin\theta$, $y' = -x \sin\theta + y \cos\theta$, θ is the rotation angle, Φ is the phase offset, λ is the wavelength of the harmonic function, σ is the standard deviation of the Gaussian function, and γ is the spatial aspect ratio. In the original GIST model, 32 filters at four scales and eight orientations are used. Each output filtered image is partitioned into 16 blocks, and the average value of each block is used as a feature. Overall, a GIST feature vector has 512 elements. Figure 1b shows the GIST features extracted from an input image shown in Fig. 1a. Here, the input image is divided into 4 × 4 regions. From each region 32 gist features (4 scales and 8 orientations) are extracted and visualized in polar coordinates.

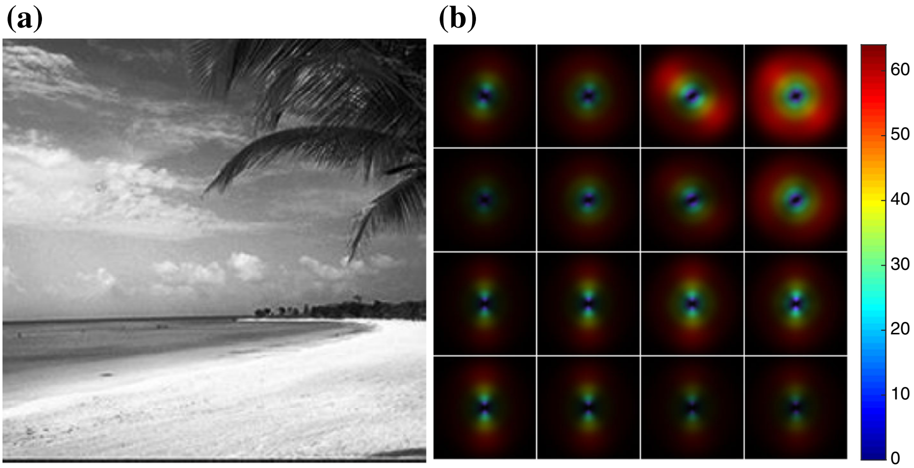


Fig. 1 Visual illustration of GIST feature extraction. **a** input image, **b** GIST features in polar form

The GIST features were found to be more effective for recognizing outdoor scenes than the indoor scenes (Wu and Rehg 2011). They have been combined with other features for scene categorization. Torralba et al. (2006) proposed a new GIST model that combines the local features, global features, bottom-up saliency, and top-down mechanisms to predict which image regions are likely to be fixated by human observers. Han and Liu (2010) developed a hierarchical GIST model for scene classification with two layers: (1) a perceptual layer based on the GIST model proposed by Oliva and Torralba (2001); and (2) a conceptual layer based on the kernel PCA (Scholkopf et al. 1998).

2.1.3 Deep learning

In recent years, deep learning architectures have gained fervent research interest for image recognition. One of the major deep learning architectures is the *convolutional neural networks* (CNN) developed by LeCun et al. (1998). CNNs are inspired by the discoveries of Hubel and Wiesel (1968) about the receptive fields in mammal visual cortex. The CNNs are based on three key architectural ideas: (1) *local receptive fields* for extracting local features; (2) *weight sharing* for reducing network complexity; (3) *sub-sampling* for handling local distortions and reducing feature dimensionality. An advantage of CNNs is that they can be trained to map raw pixels to image categories, thereby alleviating the need for hand-designed features.

A CNN is a feed-forward architecture with three main types of layers: (1) 2-D convolution layers; (2) 2-D sub-sampling layers; and (3) 1-D output layers (see Fig. 2 for an example). A *convolution layer* consists of several adjustable 2-D filters. The output of each filter is called a feature map, because it indicates the presence of a feature at a given pixel location. A *sub-sampling layer* follows each convolution layer, and reduces the size of each input feature map, via mean pooling or max pooling. The 1-D layers map the extracted 2-D features to the final network output.

Designing and training a CNN or deep learning architecture is a computation-intensive task that requires significant engineering efforts. Hinton et al. (2006) proposed a new approach for training deep networks. Their main idea is to pre-train each layer of the network using an unsupervised learning algorithm, e.g. the *restricted Boltzmann machine*, the *denoising auto-*

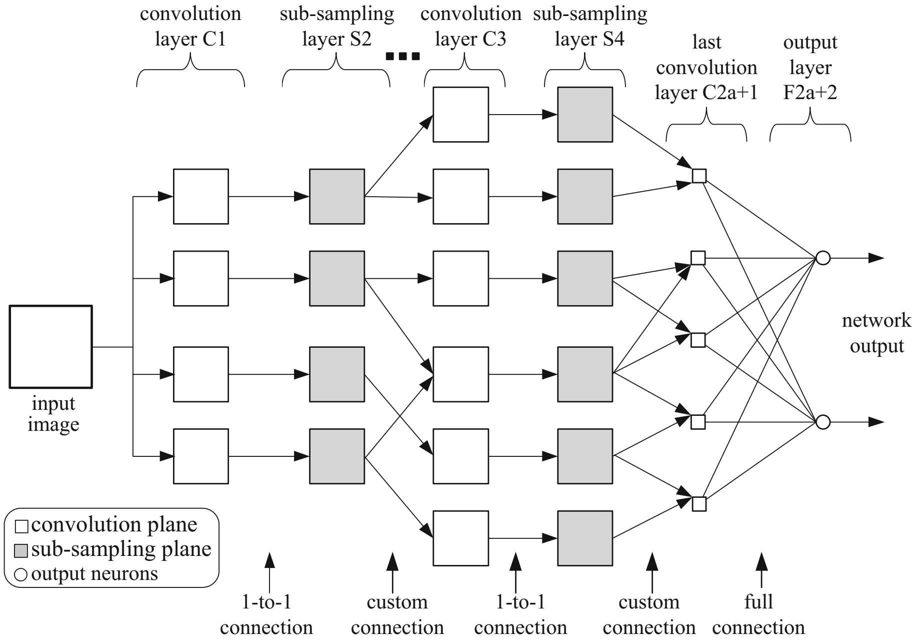


Fig. 2 An example of layers in a convolutional neural network

encoder, and the *kernel principal component analysis*. Once the pre-training is completed, a supervised learning algorithm, e.g. the error backpropagation, is employed to adjust the connection weights of the hidden layers and the output layer (Theodoridis 2015).

Krizhevsky et al. (2012) developed a CNN for image classification, which has 5 convolution layers, 650,000 neurons and 60 million parameters, and produces 4096 features. On the ImageNet benchmark, which comprises 1.2 million images with 1000 object categories, the CNN achieved a top-1 classification rate of 62.5 %, which was higher than the previous results obtained by other methods. Sermanet et al. (2014) later developed a CNN-based integrated framework, called OverFeat, to perform both localization and detection tasks; their system achieved very competitive results (1st in localization and 4th in classification) on the ImageNet benchmark. CNNs have also been applied for scene categorization by Zhou et al. (2014) and Donahue et al. (2014). CNNs have been shown to be less effective for moderate-size data sets (Goh et al. 2014), however, they perform well when trained on large-scale data sets (Krizhevsky et al. 2012; Donahue et al. 2014).

2.2 Local feature extraction

Local descriptors capture low-level properties of the scene, whereas global descriptors represent the overall spatial information. Vogel et al. (2007) studied the use of local features and global features in the categorization of natural scenes. Their results suggested that humans rely as much on local region information as on global configural information. Existing algorithms for local descriptors can be divided into three main categories: patch-based, object-based, and region-based.

2.2.1 Patch-based local features

Patch-based algorithms extract features from small patches of input images. For the SIFT (Lowe 2004) and *speeded up robust features* (SURF) (Bay et al. 2008), patches are generated from local windows around interest points. For the LBP (Ojala et al. 1996), patches are formed from rectangular regions of each pixels. For the HOG (Dalal and Triggs 2005), patches are non-overlapping blocks where orientation voting and normalization are applied. For SIFT-ScSPM (Yang et al. 2009), patches are overlapping blocks formed from a regular grid at the same scale. For the SIFT-LCS-FV (Sanchez et al. 2013), patches are overlapping blocks formed from a regular grid at five scales.

(A) Scale-invariant feature transform

Lowe (2004) developed the SIFT algorithm to extract image features that are invariant to image scale, rotation, and changing illumination. Extracting SIFT local features involves four main steps. First, the difference-of-Gaussian filters are applied to identify the location and scale of interest points. Second, the interest points with high stability are selected as the key points. Third, dominant orientation is assigned to each key point based on local image gradient. Fourth, the SIFT features that are partially invariant to affine distortions and illumination changes are extracted from key-point regions. The SIFT features are computed from image gradient magnitude and orientation in a region centered at key point. An example of SIFT key points is shown in Fig. 3a. The centers of circles are the key points, the radiuses of circles are the average scales of the key points, and the arrows inside the circles are the average orientations of the key points.

To apply the SIFT for scene categorization, Fei-Fei and Perona (2005) proposed to extract local features from dense patches. A dictionary is formed from random local patches using k-means algorithm. Then, for each input image, a feature vector is generated using the trained dictionary. An example of SIFT feature map that is extracted from dense patches is shown in Fig. 3b. In this figure, the SIFT features are averaged at each pixel location and shown.

The original SIFT algorithm has been extended by several researchers. Brown and Susstrunk (2011) proposed the *multi-spectral SIFT* (MSIFT) on color and near-infrared images for scene categorization. They showed that compared with the SIFT, HMAX, and GIST on the 8-outdoor-scene data set (Oliva and Torralba 2001), the MSIFT reduced feature dimensionality and improved recognition accuracy. Liu et al. (2011) proposed the *SIFT flow* to align images across scenes, and applied it for image alignment, video retrieval, face recognition, and motion prediction. Bo et al. (2010) improved the low-level SIFT features using a kernel approach. The kernel descriptors provide a principled tool to convert pixel attributes to patch-level features.

(B) Speeded-up robust features

Bay et al. (2008) proposed scale- and rotation-invariant visual descriptor, called SURF. The SURF detects interest points using determinant of Hessian matrix. The computation time of interest point detection is reduced by using integral images (Viola and Jones 2001). The key points are then selected from interest points using non-maximum suppression in multi-scale space. The orientation of key point is assigned using sliding orientation windows on Haar wavelet response maps. The longest vector over all windows defines the orientation of the key point. The Haar wavelet response in the horizontal direction (d_x) and the vertical direction (d_y) are computed from a 4×4 sub-region over the key point. The feature vector for each sub-region is $v = \left(\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|\right)$. SURF key points of an example image is shown in Fig. 3c. The centers of circles are the key points, the radiuses of circles are the average scales of the key points, and the arrows inside the circles are the average orientations of the key points.

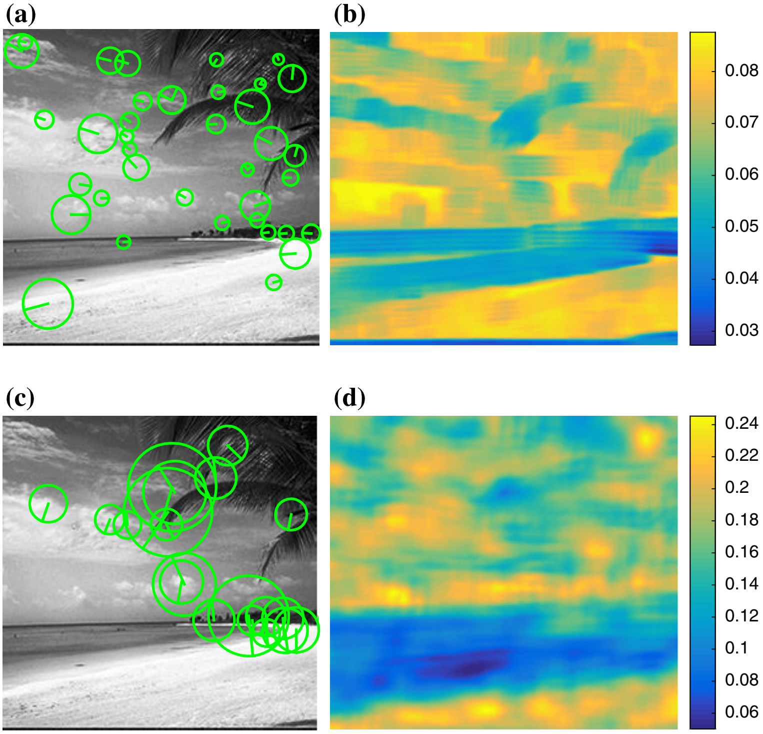


Fig. 3 Visual illustration of SIFT, SURF, and HOG feature extraction of the input image in Fig. 1a. **a** SIFT key points, **b** SIFT dense feature map, **c** SURF key points, **d** HOG dense feature map

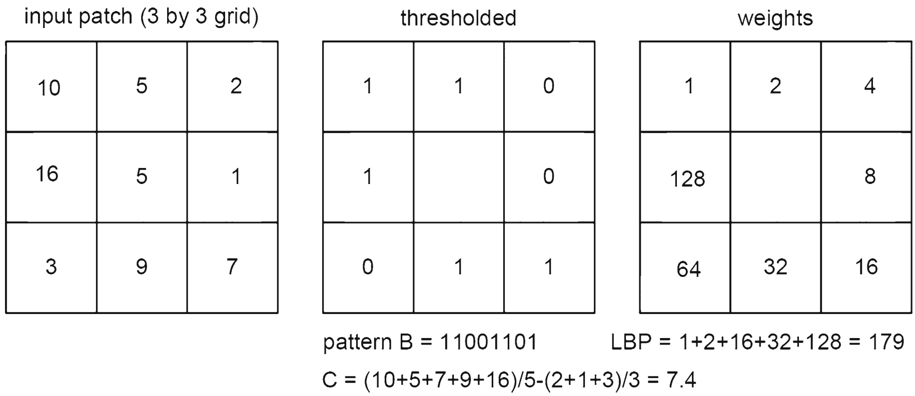
(C) Histogram of oriented gradients

Dalal and Triggs (2005) originally developed the HOG descriptor for pedestrian detection in gray-scale images. The HOG features have since been applied for recognition of other image categories, such as cars (Rybski et al. 2010), bicycles (Felzenszwalb et al. 2010), and facial expressions (Bai et al. 2009). The HOG feature extraction involves four main steps. First, an input image is normalized by the power-law, and the image gradients are computed along the horizontal and vertical directions. Next, the image is divided into cells; a cell can be a rectangular or circular region. In the third step, the histograms for multiple orientations are computed for each cell, where each pixel in the cell contributes a weighted score to a histogram. Finally, the cell histograms are normalized and grouped in blocks to form the HOG features. An example of HOG features is shown in Fig. 3d, which illustrates the strength of averaged HOG features at each pixel location.

For scene categorization, the HOG features are useful for capturing the distribution of image gradients and edge directions in a regular grid. Xiao et al. (2014a) compared the HOG with other descriptors, such as GIST and SIFT, on the SUN397 data set. The HOG achieved a higher classification rate (CR) compared with other hand-designed descriptors.

(D) Local binary pattern

Ojala et al. (1996) first developed the LBP algorithm for texture classification. Since then, the LBP has been applied to many computer vision tasks, including face recognition (Ahonen et al. 2006), pedestrian detection (Zheng et al. 2011), and scene categorization (Xiao et al.



input patch (3 by 3 grid)

10	5	2
16	5	1
3	9	7

thresholded

1	1	0
1		0
0	1	1

weights

1	2	4
128		8
64	32	16

pattern B = 11001101 LBP = 1+2+16+32+128 = 179

C = (10+5+7+9+16)/5-(2+1+3)/3 = 7.4

Fig. 4 Illustration of the basic LBP algorithm

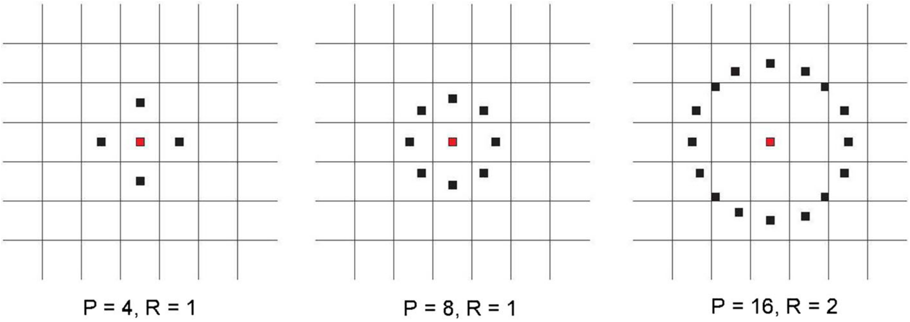


Fig. 5 The circular regions in a generic form of the LBP. Here, P is the number of neighboring pixels, and R is the circle radius. When $P = 8$ and $R = 1$, the basic LBP operator is obtained

2014a). The LBP algorithm analyzes the textures of a local patch by comparing each center pixel with the neighboring pixels.

The basic LBP operates on 3-by-3 blocks. Each pixel in the block is compared to the center pixel, and a binary value of 1 or 0 is returned (see Fig. 4). A pattern B is formed by concatenating the binary values from the neighboring pixels. The decimal LBP code is obtained by summing the thresholded differences weighted by powers of two. Furthermore, a contrast measure C is obtained by subtracting the average of pixel values smaller than the center pixel value p from the average of pixel values larger than or equal to p. An example of LBP code map is shown in Fig. 6. In the example, the pattern B is 11001101; the LBP code is 179; the contrast C is 7.4. The histogram of local contrast (LBP/C) is used as a feature vector. The computational simplicity of the LBP algorithm makes it suitable for real-time image analysis.

Several variants of the LBP have been developed. Ojala et al. (2002a) proposed a generic form of LBP that supports arbitrary neighborhood sizes. By contrast to the basic LBP that uses 8 neighboring pixels in a 3-by-3 block, the generic operator $LBP_{P,R}$ is circularly symmetric, see Fig. 5.

Ojala et al. (2002a) also developed the uniform patterns, denoted as $LBP^{u}_{P,R}$. Here, u is the number of transitions (0 to 1, or 1 to 0) in an LBP pattern. A local binary pattern is called uniform if $u \leq 2$. Different output labels are assigned to uniform LBP codes, and a single

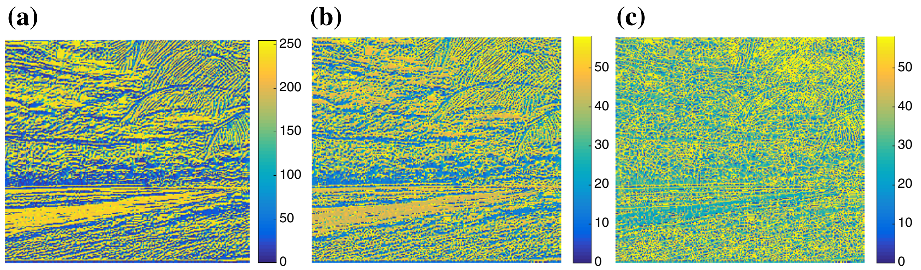


Fig. 6 Visual illustration of LBP-based feature extraction of the input image in Fig. 1a. **a** LBP, **b** uniform LBP, **c** LBP-HF

label is assigned to non-uniform patterns. Experiments by Ojala et al. (2002a) indicated that uniform patterns can be considered as fundamental textures because they represent the vast majority of local texture patterns.

Ahonen et al. (2009) proposed an algorithm, called LBP-HF, that combines the uniform LBP and Fourier coefficients. They showed that the LBP-HF has better rotation invariance than the uniform LBP. In the LBP-HF algorithm, the uniform pattern $LBP^{u}_{P,R}$ at pixel location(x, y) is replaced by $LBP^{u}_{P,(R+\theta) \bmod P}$, where θ is a rotation angle: $\theta = 0, \frac{2\pi}{P}, \frac{2\times 2\pi}{P}, \ldots, \frac{(P-1)\times 2\pi}{P}$. Then, the histograms h_{θ} of $LBP^{u}_{P,(R+\theta) \bmod P}$ are computed. Finally, the Discrete Fourier Transform is applied on h_{θ} to form LBP-HF features (Fig. 6).

Guo et al. (2010) proposed a *completed local binary pattern* (CLBP) to extract image features for texture classification. The original LBP only encodes the signs of differences between center pixel and its neighbors (see Fig. 4). The CLBP encodes both the signs (CLBP-S) and magnitudes (CLBP-M) of the differences. Furthermore, the intensity of center pixel (CLBP-C) is encoded as the third part of CLBP. Guo et al. (2010) showed that the texture classification accuracy of CLBP was better than the original LBP algorithm. Li et al. (2012) proposed a scale- and rotation-invariant LBP descriptor. The scale-invariance of this method is achieved by searching for the maximum response over scale spaces. The rotation invariance is achieved by locating the dominant orientations of the uniform-LBP patterns. The scale- and rotation-invariant LBP outperforms the classical uniform LBP in texture classification. In another approach, Qian et al. (2011) proposed an LBP descriptor with pyramid representation (PLBP). By cascading the LBP features obtained from hierarchical spatial pyramids, the PLBP descriptor extracts texture resolution information. The pyramid representation for LBP is more efficient than the multi-resolution representation for LBP proposed by Ojala et al. (2002b).

(E) Census transform histogram

Wu and Rehg (2011) proposed a visual descriptor called CENTRIST, which is a holistic representation of structural and geometrical properties of images. In the CENTRIST, feature maps are calculated by the census transform (CT), which is equivalent to the local binary pattern $LBP_{8,1}$. Wu and Regh presented an experiment showing that CT values encode shape information. The authors identified 6 CT values with highest counts (31, 248 240, 232, 15, and 23) in the 15-scene data set (Fei-Fei and Perona 2005; Lazebnik et al. 2006). The 6 CT values correspond to local 3-by-3 neighborhoods that have horizontal or close-to-diagonal edge structures. To encode the global structure of an image, the CT values are processed by the spatial pyramid algorithm, described in (Lazebnik et al. 2006). The features extracted from the pyramid feature maps are then reduced using the spatial PCA or BoW methods.

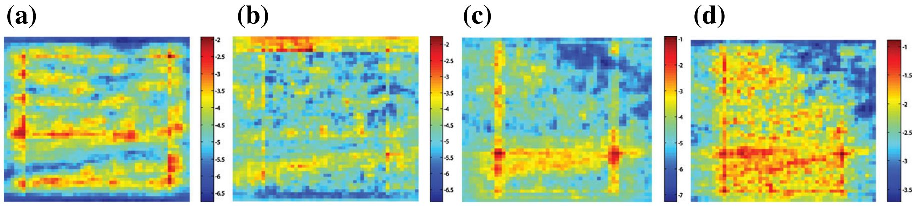


Fig. 7 Object bank feature maps of input image in Fig. 1a. **a** 'tree', **b** 'sky', **c** 'beach', **d** 'coast'

Wu and Rehg (2011) showed experimentally that the CENTRIST with the spatial PCA outperforms the state-of-the-art algorithms, such as the SIFT and GIST, on several scene categorization data sets. However, the CENTRIST has a number of limitations. First, the CENTRIST is not invariant to rotations. Second, it is not designed for extracting color information. Third, the CENTRIST still has difficulty in learning semantic concepts from images with varied viewing angles and scales. Based on the CENTRIST, a multi-channel feature generation mechanism was proposed by Xiao et al. (2014b). The CENTRIST features with multi-channel information (RGB channels and infrared channel) improves grayscale CENTRIST's performance on scene categorization.

2.2.2 Object-based local features

Object-based algorithms rely on landmark objects to classify scenes. They have been applied for scene perception in robotic navigation systems (Abe et al. 1999; Yu et al. 2009b; Schauerte et al. 2011). Bao et al. (2011) proposed a scene layout reconstruction algorithm by determining the 3-D locations and the support planes of objects. In these methods, the scene is classified based on landmark objects and their configuration. A challenge of object-based methods is to detect small objects, especially in outdoor conditions. Another challenge is to select a small set of landmark objects to represent a scene (Siagian and Itti 2007).

An example of object-based algorithms for visual categorization are *object bank* proposed by Li et al. (2010). The OB algorithm was designed to decrease the gap between low-level visual features of objects and high-level semantic information of the scene.

In the OB approach, an image is represented by scale-invariant response maps produced by pre-trained object detectors. Objects are classified by two types of detectors: (1) the SVM detector, proposed by Felzenszwalb et al. (2010), for objects such as humans, cars, and tables; (2) the texture detector, proposed by Hoiem et al. (2006), for objects such as sky, road, and sea. Li et al. (2010) analyzed common object types in four data sets ESP (Von Ahn 2006), LabelMe (Russell et al. 2008), ImageNet (Jia et al. 2009), and Flickr (Yahoo 2004), and selected 200 object detectors that were trained with 12 detection scales and 3 spatial pyramid levels (Peters and Itti 2007). The example feature maps produced by four object detectors are shown in Fig. 7a–d. Li et al. (2010) compared the OB algorithm with SIFT, GIST, and *spatial pyramid matching* (SPM) (Lazebnik et al. 2006) on several scene data sets. The results showed that OB outperforms the other algorithms on the UIUC data set (Li and Fei-Fei 2007), the 67-indoor-scene data set (Quattoni and Torralba 2009), and the 15-scene data set (Lazebnik et al. 2006).

2.2.3 Region-based local features

Region-based algorithms segment images and extract features from different regions. Boutell et al. (2006) proposed an algorithm to classify scenes using the identities of regions and

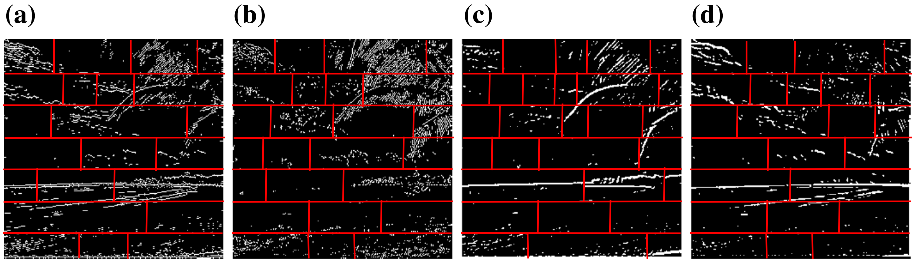


Fig. 8 Illustration of SEV regions of input image in Fig. 1a and the computed edge maps along the **a** horizontal direction, **b** vertical direction, **c** +45 degree direction, and **d** −45 degree direction

the spatial relations between regions. Gokalp and Aksoy (2007) proposed a bag-of-regions algorithm for scene classification. In their algorithm, an image is partitioned into regions, and the structure of the image is represented by a bag of individual regions and a bag of region pairs. Juneja et al. (2013) used distinctive parts for scene categorization. First, the distinctive parts in each category are detected and learned based on HOG features. Then, the features of parts are extracted and encoded using bag of words or Fisher vector.

An example of region-based algorithms is the *sequential edge vectors* (SEV) proposed by Morikawa and Shibata (2012). The SEV reduces the ambiguity of features caused by the inter-class variations among scene categories. Unlike other region-based algorithms that segment the entire image, the SEV segments only local regions. The SEV method considers an image as a *document* consisting of *sentences*, in which the separated regions play the role of *words*, and oriented edges play the role of *letters*. The SEV identifies each word from the letters, forms the document vector from words, and finally determines the topic vector.

The main steps of the SEV method can be described as follows. First, oriented edges of input image are detected by horizontal, vertical, and diagonal edge filters. Second, the horizontal or vertical edge map is used to generate fixed sentences (also called local threads). All local threads are scanned from one end to the other, producing edge distribution vectors in four directions. Next, the sum of absolute differences (SAD) between neighboring local windows is calculated. The boundaries of meaningful sequences are determined by locating peaks in the SAD histograms. Figure 8 shows the boundaries in four edge maps. Then, the meaningful sequences form words that describe images. Finally, probabilistic latent semantic analysis (PLSA) (Quelhas et al. 2005) is applied on meaningful sequences (words) to generate topic vectors.

Using the 8-outdoor-scene data set, Morikawa and Shibata evaluated the SEV algorithm on seven scene categories: coast, open country, forest, highway, mountain, street, and tall building. The SEV method with a 16-by-16 scan window achieved a higher F-measure (65 %) than a model that uses SIFT and PLSA (55 %).

2.3 Global feature formation

For scene categorization, global features are often extracted without image segmentation or object detection to summarize the statistics of the image. For example, Renninger and Malik (2003) computed histograms of features generated by a bank of Gaussian derivative filters to represent scenes. Serrano et al. (2004) used quantized color histograms and wavelet texture as global features for scene categorization. Furthermore, Mikolajczyk and Schmid (2005) showed that the performance of visual descriptors depends not only on local regions but also global information. Therefore, several scene categorization algorithms focus on first extracting suitable local features and then forming global features from regular grids. In the following subsections, four representative algorithms for global feature formation in scene categorization are described: the PCA, histograms, BoW, and Fisher vector.

2.3.1 Principal component analysis

Principal component analysis (PCA) represents data by a small number of principal components. It has been used for dimensionality reduction in a wide range of computer vision tasks, such as image categorization (Han and Chen 2009), face recognition (Guangda et al. 2002; Xie and Lam 2006), and feature selection (Malhi and Gao 2004; Yang and Li 2010; Ebied 2012).

Ke and Sukthankar (2004) proposed a visual descriptor, known as PCA-SIFT, that combines the SIFT and PCA. In their approach, a projection matrix P is calculated from a large number of image patches. For each training image patch, a feature vector is generated from the horizontal and vertical gradient maps. The covariance matrix C of all feature vectors is calculated. Eigenanlysis is applied to C, and the n most-significant eigenvectors of C are selected to form the projection matrix P. In Ke and Sukthankar's experiments, n was selected to be 20, which is significantly smaller than the number of features (128) in the standard SIFT algorithm. Compared to the SIFT features, the PCA-SIFT features lead to an improvement in image matching, when evaluated on the INRIA Graffiti data set (2004).

2.3.2 Histogram

Histogram is a method to represent the statistical distribution of data. It is efficient and robust to noise compared to other feature formation methods (Hadjidemetriou et al. 2004). Therefore, it has been used in many image processing tasks, including image and video retrieval (Lee et al. 2003; Song et al. 2004; Jeong et al. 2003; Liu and Yang 2013), image structure analysis (Koenderink and Van Doorn 1999), image filtering (Kass and Solomon 2010; Igarashi et al. 2013), and color indexing (Swain and Ballard 1991; Stricker and Orengo 1995; Funt and Finlayson 1995; Kikuchi et al. 2013). In several scene categorization algorithms, such as the LBP and its variants, global features are formed by calculating the histogram of local features (Ojala et al. 1996; Gupta et al. 2009; Wu and Rehg 2011).

A weakness of the histogram approach is that it does not capture the spatial information. To encode spatial information, Hadjidemetriou et al. (2004) proposed multi-resolution histograms that extract shape and texture features at several resolution levels. Given an image $I(x, y)$ with n gray-levels, the spatial resolution of the image is decreased by convolving it with a Gaussian kernel $G(x, y; \sigma)$. The multi-resolution histograms are calculated as $\mathbf{h}[I * G(\sigma)]$. Then, the cumulative histograms corresponding to each image resolution is computed. Next, the differences between the cumulative histograms of consecutive levels are calculated. The difference histograms are sub-sampled and normalized to make them independent of the sub-sampling factor. Finally, the normalized difference histograms are concatenated to form a feature vector.

2.3.3 Bag-of-words

The *bag-of-word* (BoW) algorithms were first used in document classification to simplify the representation of natural language. Recently, BoW has been used to classify images based on the appearance of image patches. In the BoW, features are extracted from regular grids (*feature extraction*) and quantized into discrete visual words (*encoding*). The coding strategy aims to generate similar codes for similar features. A compact representation of visual words is built for each image based on a trained dictionary (*codebook*).

Qin and Yung (2010) proposed an algorithm based on contextual visual words. To train the visual words, SIFT features are calculated from both the *region of interest* (ROI) and the regions surrounding the ROI. Fei-Fei and Perona (2005) proposed a BoW algorithm based on *latent Dirichlet allocation* (Blei et al. 2003). They compared different local patch detectors, such as the regular grids, random sampling, saliency detector (Kadir and Brady 2001), and the difference-of-Gaussian detector (Lowe 2004). Their experiment results showed that the regular grids perform better than the random sampling, saliency detector, and difference-of-Gaussian detector for scene categorization.

Next, we describe three representative BoW algorithms for training visual words on regular grids: SPM (Lazebnik et al. 2006), ScSPM (Yang et al. 2009), and LLC (Wang et al. 2010).

Spatial pyramid matching (SPM) for recognizing natural scenes was proposed by Lazebnik et al. (2006). Unlike the traditional BoW algorithms that extract order-less features, the SPM algorithm retains the global geometric correspondence of images. It divides the input image into regular grids and computes local features, such as SIFT and HOG in each grid. The visual vocabulary is formed by k-means clustering, and then all features are formed using vector quantization (VQ). Based on the trained dictionary, local features are represented. Finally, the spatial histograms (average pooling) of coded features are used as feature vectors. To recognize multiple scene categories, a support vector machine with the one-versus-all strategy is used.

Sparse coding based spatial pyramid matching (ScSPM) was suggested by Yang et al. (2009) to improve the efficiency of the SPM. The original SPM uses k-means vector quantization, whereas the ScSPM uses sparse coding to quantize the local features. Furthermore, for spatial pooling, the original SPM uses histograms, whereas the ScSPM applies the *max* operator, which is more robust to local spatial translations. For the ScSPM, the linear-SVM classifier is used to reduce the computation cost. The experiments by Yang *et al.* show that the sparse coding of SIFT descriptors with the linear-SVM outperforms several methods, including kernel codebooks (Gemert et al. 2008), SVM-K-Nearest-Neighbor (Zhang et al. 2006), and naive Bayes nearest-neighbor (NBNN) (Boiman et al. 2008).

Locality-constrained linear coding (LLC) was proposed by Wang et al. (2010) to reduce the computational cost of the SPM. Because of the importance of locality, as demonstrated by Yu et al. (2009a), the LLC replaces the sparsity constraint used in ScSPM with the locality constraint to select similar bases for local features from the trained visual words. In Wang et al.'s approach, a linear weighted combination of these bases is learned to represent local features. Their experiments on the Catech-101 data set (Fei-Fei et al. 2007a) show that the LLC achieved higher classification rates than the ScSPM, NBNN (Boiman et al. 2008), and kernel codebooks (Gemert et al. 2008). Recently, Goh et al. (2014) improved the ScSPM and LLC by using a deep architecture. The deep architecture merges the strengths of the BoW framework and the deep learning method to encode the SIFT features.

2.3.4 Fisher vector

Fisher vector (FV) is a feature encoding algorithm proposed by Sanchez et al. (2013). The local features extracted from dense multi-scale grids are represented by their deviation from a *Gaussian mixture model* (GMM). The local features are mapped to a higher-dimensional space which is more amenable to linear classification.

In the FV algorithm, a GMM is first computed from a training set of local features using the Expectation–Maximization algorithm. The parameters of a GMM are denoted by $\lambda = \{(w_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k), k = 1, \ldots, K\}$, where w_k is the mixture weight, $\boldsymbol{\mu}_k$ is the mean vector, and $\boldsymbol{\sigma}_k$ is the covariance matrix of the kth Gaussian component.

Let $X = \{\mathbf{x}_t, t = 1, \ldots, T\}$ be the set of local descriptors extracted from an input image. For local descriptor $\mathbf{x}_t$, let $p_\lambda(\mathbf{x}_t)$ be the probability density function of local descriptor, as computed by the GMM model. Let L_λ be the square-root of the inverse of the Fisher information matrix. The normalized gradient statistics are computed as

$$\varphi(\mathbf{x}_t) = L_\lambda \nabla_\lambda \log p_\lambda(\mathbf{x}_t). \tag{2}$$

A Fisher vector is the sum of normalized gradient statistics:

$$\mathcal{G}_\lambda^X = \sum_{t=1}^{T} \varphi(\mathbf{x}_t) \tag{3}$$

The final Fisher vector is the concatenation of the gradients $\mathcal{G}_{w_k}^X$, $\mathcal{G}_{\boldsymbol{\mu}_k}^X$, and $\mathcal{G}_{\boldsymbol{\sigma}_k}^X$. To improve the classification accuracy, two normalization steps, l_2-normalization and power normalization (Perronnin et al. 2010), can be applied on the Fisher vector.

Compared with other bag-of-words algorithms, the Fisher vector has many advantages. First, it provides a generalized method to define a kernel from a generative process of the data. Second, Fisher vector can be computed from small vocabularies with a lower computational cost. However, the Fisher vector is dense, which leads to storage issues for large-scale applications.

2.3.5 Composite global features

Global features are also formed by combining different global features. For example, the CENTRIST features are generated by combining histograms and PCA (Wu and Rehg 2011). The HOG-SPM and SIFT-SPM features are first extracted by BoW algorithms, and then accumulated by spatial histograms (Lazebnik et al. 2006). The SIFT-ScSPM (Yang et al. 2009) and SIFT-LLC (Wang et al. 2010) represent local features by the BoW algorithm and the max pooling. The image categorization method proposed by Krapac et al. (2011) combines the SPM and the Fisher kernel to encode spatial layout of images.

3 Data sets and performance measures

Progress in image recognition is due in part to the existence of comprehensive data sets on which new or existing algorithms can be rigorously evaluated (Salzberg 1997; Phillips et al. 2000). In this section, we review the publicly available data sets and the performance measures for scene categorization algorithms, and discuss their characteristics.

3.1 Data sets for scene categorization

Because of the difficulty of finding one representative data set, it is not sufficient to evaluate a scene categorization algorithm on only one data set. For example, on the Caltech-101 data set, the LLC algorithm is found to have a higher classification rate than the ScSPM algorithm (Wang et al. 2010). However, on the 15-scene data set, the ScSPM algorithm has a higher classification rate than the LLC algorithm (see Sect. 4.4). One possible reason is that built-in

Table 3 Data sets for scene categorization

Data set	Data set size (images)	Image categories
8-outdoor-scene (Oliva and Torralba 2001)	2600	8 outdoor scenes
13-natural-scene (Fei-Fei and Perona 2005)	3759	13 natural scenes
15-scene (Lazebnik et al. 2006)	4485	15 indoor/outdoor scenes
67-indoor-scene (Quattoni and Torralba 2009)	15,620	67 indoor scenes
SUN397 (Xiao et al. 2014a)	108,754	397 general scenes
Places205 (Zhou et al. 2014)	2,448,873	205 general scenes

biases are present when collecting image data for a recognition task, e.g. the viewing angle, the type of background scene, and the composition of objects. These intrinsic biases cause every data set to represent the physical world differently. Therefore, scene categorization algorithms should be evaluated on multiple large-scale data sets that exhibit more diversity and less bias.

Table 3 summarizes the major data sets used for scene categorization. For each data set, the source, the number of images, and the number of image categories are given. In the following, we describe the major data sets. Three benchmark data sets, 8-outdoor-scene (Oliva and Torralba 2001), 13-natural-scene (Fei-Fei and Perona 2005), and 15-scene (Lazebnik et al. 2006), have been used by many researchers (Grossberg and Huang 2009; Qin and Yung 2010; Jiang et al. 2010; Kazuhiro 2011; Cakir et al. 2011; Morikawa and Shibata 2012; Meng et al. 2012; Goh et al. 2014; Zhang et al. 2015). The 8-outdoor-scene data set has eight categories of only-outdoor scenes (coast, forest, highway, inside city, mountain, open country, street, and tall buildings), whereas the 13-natural-scene data set contains the same eight categories of outdoor scenes and five additional categories of indoor scenes (bedroom, kitchen, living-room, office, and store). The 15-scene data set includes all images from the 13-natural-scene, and two additional outdoor scenes of man-made structures (suburb and industry). The images in each category are from different sources, such as COREL data set, Google image search, and personal photographs.

Large data sets have been developed to increase the number of semantic categories and the diversity of images. The 67-indoor-scene data set (Quattoni and Torralba 2009) divides five indoor scenes (working places, home, leisure, store, and public spaces) into 67 sub-categories. Common objects appear in multiple categories; therefore, it is harder to distinguish between the image categories. The SUN397 data set (Xiao et al. 2014a) has 397 scene categories, from abbey, bedroom, and castle to highway, theater, and yard. There are at least 100 images in each category. The Places205 data set (Zhou et al. 2014) has 205 scene categories; each category has at least 5000 images. Among the benchmark data sets listed in Table 3, the SUN397 has the most number of categories and the Places205 has the most number of images.

3.2 Performance measures

To evaluate a scene categorization system, a data set is typically partitioned into three separate subsets: training, validation, and test. The training subset is used to determine the system's adjustable parameters, the validation set is used to prevent over-training, and the test set is used to estimate the system's generalization capability. The generalization capability of a system is commonly measured using the classification rate (*CR*), which is the percentage of test images that are correctly classified. For example, the *CR* has been used for scene categorization in (Elfiky et al. 2012; Wu and Rehg 2011; Kazuhiro 2011; Meng et al. 2012).

To prevent bias in partitioning the data set and to estimate more reliably the generalization capability, an alternative technique known as n-fold *cross-validation* is usually adopted. The image data set is divided into n subsets of equal size. For each validation fold, one subset is used for testing, and the remaining $(n-1)$ subsets are used for training and validation. This is repeated n times, each time a different subset fold is used for testing. Finally, the n classification rates are averaged to give the overall *CR*.

Scene categorization is a multi-class recognition problem. Many scene categorization algorithms are also evaluated using the *confusion matrix* (Wu and Rehg 2011; Perina et al. 2010; Meng et al. 2012; Morikawa and Shibata 2012; Cakir et al. 2011; Qin and Yung 2010). For a problem involving K categories, the confusion matrix has K rows and K columns. Each row represents an actual category, and each column represents a predicted category. The entry at row r, column c is the number of category r samples that are classified as category c. Clearly, the correct classification for individual categories are the diagonal entries, whereas the miss-classification are the non-diagonal entries.

Evaluation measures for two-class problems are also applied for scene categorization. Consider a scene category r. The positive class consists of all samples belonging to category r, whereas the negative class consists of all samples belonging to the remaining categories. Four measures can be computed:

– *True positives* (tp) is the number of test samples in the positive class that are correctly classified.
– *False positives* (fp) is the number of test samples in the negative class that are incorrectly classified.
– *False negatives* (fn) is the number of test samples in the positive class that are incorrectly classified.
– *True negatives* (tn) is the number of test samples in the negative class that are correctly classified.

The precision rate P and the recall rate R are then defined as

$$P = \frac{tp}{tp + fp}, \text{ and } R = \frac{tp}{tp + fn}. \tag{4}$$

A good scene categorization system should have a high precision rate and a high recall rate. These two requirements can be reflected in a single measure called the *F-measure*, which is the harmonic mean of the precision rate and recall rate:

$$F = \frac{2PR}{P + R}. \tag{5}$$

The plot of the true positive rate versus the false positive rate is called the *receiver operating characteristic* (*ROC*) curve (Fawcett 2006). It is a useful tool for visualizing the scene categorization performance, when a system parameter is varied. Another performance measure is the *area-under-the-ROC-curve* or *AUC*. The *ROC* curve and the *AUC* have been used for scene categorization in (Jia et al. 2009) and (Xiao et al. 2014a).

The measures described above are suitable for evaluating the performance of a complete scene categorization system that includes both a feature extractor and a classifier. To evaluate the performance of the feature extraction independently of the classifier, we can use the class separability of the extracted features. The *Fisher's discriminant analysis* (FDA) is a tool for analyzing the separability of features. For example, Li et al. (2003) and Chin et al. (2011) used

FDA to analyze multi-class image classification. Consider a scene categorization problem that involves K classes. The within-class covariance matrix C_w is calculated as

$$C_w = \sum_{k=1}^{K} \sum_{\mathbf{x} \in \omega_k} (\mathbf{x} - \bar{\mathbf{x}}_k)(\mathbf{x} - \bar{\mathbf{x}}_k)^T, \tag{6}$$

where $\bar{\mathbf{x}}_k$ is the mean vector of class ω_k. The between-class covariance matrix C_b is given by

$$C_b = \sum_{k=1}^{K} N_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^T, \tag{7}$$

where $\bar{\mathbf{x}}$ is the mean vector of all classes and N_k is the number of samples in class ω_k. The *S score* of the feature vector is given by

$$S = \frac{|\mathrm{trace}(C_b)|}{|\mathrm{trace}(C_w)|}. \tag{8}$$

A high S score means there is a high separability between the K classes using the given feature vector.

4 Experimental evaluation and results

In this section, we present an extensive experimental evaluation of the performance of state-of-the-art descriptors, which include biologically-inspired, local, and global feature extraction methods. The selected descriptors were evaluated on four benchmark data sets with respect to scene categorization accuracy and class separability of feature vectors. For scene categorization, five measures are used to evaluate the classification accuracy, namely classification rate, precision, recall, F-measure, and *AUC*. By using the same classification protocol described in (Krizhevsky et al. 2012), we are able to compare descriptors on the SUN397 data set with several recent methods: ImageNet-CNN (Krizhevsky et al. 2012), BOP-FV (Juneja et al. 2013), discriminative patches (Doersch et al. 2013), OverFeat (Sermanet et al. 2014), Places-CNN (Zhou et al. 2014), and DeCAF (Donahue et al. 2014). For class separability, the visual descriptors are compared using Fisher discriminant analysis.

Section 4.1 describes the four image data sets used in the experiments. Section 4.2 and 4.3 describe the implementation of the descriptors and classifiers. Section 4.4 presents results of a comparative study using different classifiers with four data sets. Finally, Sect. 4.5 evaluates the class separability and stability of feature vectors.

4.1 Image data sets

The experiments were conducted on four data sets: the 8-outdoor-scene data set, the 15-scene data set, the 67-indoor-scene data set, and the SUN397 data set. The 8-outdoor-scene data set and the 15-scene data set have been used as benchmark for scene categorization by many researchers (Fei-Fei and Perona 2005; Yang et al. 2009; Quattoni and Torralba 2009; Li et al. 2010; Jiang et al. 2010; Wu and Rehg 2011; Kazuhiro 2011; Karayev et al. 2014). The 67-indoor-scene data set contains 67 indoor scene categories. There are at least 100 images per category and all images have a size of at least 200 × 200 pixels. The 67-indoor-scene data set has been used to evaluate scene categorization in (Quattoni and Torralba 2009; Wang et al. 2010; Wu and Rehg 2011; Doersch et al. 2013; Juneja et al. 2013; Bergamo and Torresani 2014; Cheng et al. 2015; Dixit et al. 2015). The SUN397 data set contains 397 scene categories and 108,754 images. It has been used as a benchmark for scene categorization by Huang et al. (2014b), Sanchez et al. (2013), Sun et al. (2013), Bergamo and Torresani (2014), Zhou et al. (2014), Donahue et al. (2014), and Dixit et al. (2015).

4.2 Implementation of visual descriptors

This section describes the implementation of the visual descriptors compared in our experiments. There are two biologically-inspired descriptors (GIST and HMAX), four SIFT-based descriptors (SIFT-SPM, SIFT-ScSPM, SIFT-LLC, and SIFT-FV), two other BOW-based descriptors (HOG-SPM and SURF-ScSPM), five LBP-based descriptors (LBP, Uniform LBP, LBP-HF, PLBP, CENTRIST), and one object-based descriptor (OB). Most of the 14 descriptors combine local and global features.

The biologically-inspired descriptors implemented in this paper are GIST and HMAX. The *GIST descriptor* is a low dimensional representation of an image. In our experiment, the normalized input image was convolved with Gabor filters at 4 scales and 8 orientations. Each filtered output was down-sampled to a 4 by 4 patch and reshaped to a 16 element vector. The GIST descriptor assembled all outputs from the 32 filters to form a feature vector with 512 elements. The *HMAX descriptor* contains two simple layer $S1$ and $S2$, and two complex layer $C1$ and $C2$. In our experiment, the $S1$ layer was formed from the outputs of Gaussian filters with 4 orientations and 12 scales. In layer $C1$, the $S1$ feature maps were grouped into 4 filter bands of a certain scale range. The max pooling operation was applied to each filter bank. Only $S1$ units with the same preferred orientation fed into a given C1 unit. The $S2$ features were formed from $C1$ features with 256 visual words (learned from $C1$). The $C2$ features were generated from $S2$ features using max pooling. The final HMAX descriptor had 4069 features.

The SIFT-based descriptors were compared to evaluate the encoding capability of SPM, ScSPM, LLC, and FV. For the four SIFT-based descriptors (SIFT-SPM, SIFT-ScSPM, SIFT-LLC, and SIFT-FV), their local features were calculated from overlapping patches (16×16) on a dense grid every 8 pixels. The local patch was first filtered with Gaussian filters to generate 8 orientation maps. The histograms were then generated from the orientation maps and further weighted by a Gaussian function. The SIFT descriptor was formed by concatenating the orientation histograms. The *SIFT-SPM descriptor* extracted global features from dense SIFT features by the SPM algorithm. In the experiments, k-means clustering and PCA were used to train and extract 200 visual words from random samples of local features. The local features were quantized by the trained visual words. Finally, histograms of quantized features were formed with 1000 bins. The *SIFT-ScSPM descriptor* formed global features from dense-SIFT features using the ScSPM algorithm. A feature dictionary including 1024 visual words was obtained by applying k-means clustering to the local features. Finally, the ScSPM algorithm was employed to convert the local SIFT features to global features. The *SIFT-LLC descriptor* formed global features from dense-SIFT features using the LLC algorithm. The SIFT-LLC descriptor converted the local feature maps to global features based on the n nearest neighbors of the feature dictionary. In the experiment, the number of neighbors was set to 5. The *SIFT-FV descriptor* extracted global features from dense-SIFT features using the Fisher vector encoding. The dictionary was trained using a Gaussian mixture model with 256 Gaussians.

Other BOW-based descriptors that are similar to the SIFT-based descriptors include HOG-SMP and SURF-ScSPM. In our experiment, the *HOG-SPM descriptor* extracted local features from overlapping patches (16×16) on a dense grid every 8 pixels. The input image was first normalized globally by the power-law. Then, image gradients were computed along the horizontal and vertical directions. For each patch, histograms for multiple orientations were computed to form HOG features. The global features were then generated by the spatial pyramid matching (SPM). The *SURF-ScSPM descriptor* used the SURF for local feature extraction. The ScSPM algorithm was applied for global feature formation. Different from the SIFT-based descriptors and HOG-SMP that extracted local features from dense patches, the SURF-ScSPM descriptor extracted local features from interest-points (sparse patches). In our experiment, at least 100 interest points were found in each image.

The LBP-based descriptors compared in the experiment include LBP, uniform LBP, LBP-HF, PLBP, and CENTRIST. The *LBP descriptor* is a histogram of the LBP feature map. In our experiment, the feature map was generated from 3-by-3 blocks of the entire image. The number of histogram bins was selected as 256. When the number of histogram bins was reduced, our preliminary experiments indicated that the classification rate dropped by about 10 %. The *uniform LBP descriptor* is similar to the LBP descriptor, but it only encodes uniform patterns of the LBP. For the uniform LBP, we represented each input image with 59 uniform patterns. The histograms of the uniform LBP were calculated with 59 bins. The *LBP-HF descriptor* is an extension of the uniform LBP. In our experiment, the histograms of the uniform LBP were first calculated on the input image and its 90-degree rotated version. Then, the Fourier transform was applied on the histograms. The magnitudes of Fourier coefficients were calculated as the LBP-HF features. The final LBP-HF vector had 76 elements: half of the elements were generated from the original image, and the other half from the rotated image. The *PLBP descriptor* is an LBP descriptor with spatial pyramid representation. To calculate PLBP features, each input image was decomposed into 5 Gaussian pyramid images with dyadic scales. The histograms of the LBP in each pyramid image were combined to form the PLBP features. The *CENTRIST descriptor* first converted input image into the CENTRIST feature map (similar to LBP feature map). Then, spatial histogram with 3 spatial levels and PCA with 40 eigenvectors were employed to form the CENTRIST feature map. The dimension of the CENTRIST was 1240.

The object-based descriptor OB extracts features from a large number of pre-trained object detectors. It has the longest feature-extraction stage among the compared descriptors. In our experiment, 176 object detectors with 12 detection scales and 3 spatial pyramid levels were used. An OB descriptor with 44,604 features was formed by max pooling.

4.3 Classification procedure

Support Vector Machines (SVM) with linear, RBF and HIK kernels were used to classify the different descriptors. The linear kernel is given by

$$K(\mathbf{f}_i, \mathbf{f}_j) = \mathbf{f}_i \cdot \mathbf{f}_j, \tag{9}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are two feature vectors. The RBF kernel is given by

$$K(\mathbf{f}_i, \mathbf{f}_j) = \exp\left\{-\gamma ||\mathbf{f}_i - \mathbf{f}_j||^2\right\}, \tag{10}$$

where γ is a positive scalar. The HIK kernel is computed as

$$K(\mathbf{f}_i, \mathbf{f}_j) = \sum_{n=0}^{N} \min\left[\mathbf{h}_i(n), \mathbf{h}_j(n)\right], \tag{11}$$

where $\mathbf{h}_i(n)$ and $\mathbf{h}_j(n)$ are, respectively, the N-bin histograms of $\mathbf{f}_i$ and $\mathbf{f}_j$.

Five-fold cross validation was used to evaluate the performance of the SVM classifiers with different visual descriptors. In each fold, four subsets were used for training and validation, and the remaining subset was used for testing. The parameters of the SVM classifiers were determined using a validation set in each fold. The average values of *CR, P, R, F*, and *AUC* were calculated over the five folds. The standard deviations of *CR, P, R, F*, and *AUC* over the five folds are used as a measure of variation in the classification performance.

4.4 Classification results

The following four subsections present and discuss the results of image classification on the 15-scene, 8-scene, 67-indoor-scene, and SUN397 data sets.

4.4.1 Classification results on the 15-scene data set

In the first experiment, the 14 selected descriptors were evaluated on the 15-scene data set. Tables 4, 5 and 6 present the classification performance measures of the different visual descriptors using the linear-SVM, RBF-SVM, and HIK-SVM. In these tables (and also Tables 7, 8, and 9), the best performance measure is indicated in bold font. The results in these tables show that each descriptor achieves its best classification rate using a different SVM classifier. For example, the SIFT-ScSPM, SIFT-LLC, SIFT-FV, SURF-ScSPM, CENTRIST, and OB achieve higher classification rates with the linear-SVM than with the RBF- or HIK-SVM. The LBP-HF and GIST have their highest classification rates when using the RBF-SVM. The HOG-SPM, SIFT-SPM, LBP, uniform LBP, PLBP, and HMAX achieve their highest classification rates with the HIK-SVM. The highest classification rates on the 15-scene data set for individual descriptors are (in a descending order) the SIFT-ScSPM (84.5 %), SIFT-LLC (83.0 %), SIFT-FV (80.2 %), HOG-SPM (80.0 %), OB (79.9 %), SIFT-SPM (77.9 %), SURF-ScSPM (73.3 %), PLBP (73.0 %), GIST (72.8 %), CENTRIST (72.7 %), LBP (71.9 %), uniform LBP (70.6 %), LBP-HF (67.3 %), and HMAX (63.9 %).

The biologically-inspired algorithms HMAX algorithm has the lowest *CRs* among the compared algorithms. As shown in Tables 4, 5 and 6, the *CRs* of the HMAX using the linear-SVM, RBF-SVM, and HIK-SVM are 61.1, 62.4, and 63.9 %, respectively. The other biologically-inspired descriptor, namely the GIST, has *CRs* of 71.5, 72.8, and 72.1 %, respectively. Furthermore, the GIST algorithm performed better than the LBP-based and HMAX algorithms with all three SVM kernels.

The SIFT-ScSPM outperforms all other 13 descriptors on the 15-scene data set; its *CRs* is 84.5 % for the linear-SVM, 83.8 % for the RBF-SVM, and 83.6 % for the HIK-SVM. The SIFT-ScSPM algorithm also has higher values of precision, recall, F-measure, and *AUC* than other algorithms.

The SIFT-LLC and SIFT-FV perform better than the HOG-SPM and the SIFT-SPM. The SIFT-LLC uses the locality-constrained linear coding and the SIFT-FV uses Fisher kernel coding, whereas the HOG-SPM and the SIFT-SPM uses vector quantization for global feature formation. The result indicates that a better encoding algorithm like ScSPM and Fisher vector improves the classification performance. Note that among the top-seven algorithms, SIFT-ScSPM, SIFT-LLC, SIFT-FV, HOG-SPM, SIFT-SPM, and SURF-ScSPM all encode local features using BoW.

It is interesting to note that the classification rates of HOG-SPM and SIFT-SPM improve by 8.3 and 12.0 %, respectively, when using the RBF-SVM compared to using the linear-SVM. In addition, using the HIK-SVM increases the classification rates of HOG-SPM and SIFT-SPM by 9.4 and 17.0 %, respectively. However, for the SIFT-LLC, SIFT-ScSPM, and SIFT-FV, the classification performance is not improved by using the RBF-SVM and HIK-SVM. Note that previous tests (Yang et al. 2009; Wang et al. 2010) on several benchmark data sets also

Table 4 Scene categorization performance on the 15-scene data set using the linear-SVM

ID	Algorithms	CR (%)	Precision (%)	Recall (%)	F-measure (%)	AUC (%)
1	SIFT-ScSPM	**84.5 ± 1.5**	**84.8 ± 1.7**	**84.0 ± 1.4**	**84.1 ± 1.4**	**98.9 ± 0.1**
2	SIFT-LLC	83.0 ± 1.3	83.0 ± 1.5	82.2 ± 1.5	82.3 ± 1.4	98.8 ± 0.2
3	SIFT-FV	80.2 ± 1.8	79.6 ± 1.9	79.2 ± 1.9	79.0 ± 2.0	98.4 ± 0.1
4	HOG-SPM	70.6 ± 2.7	71.7 ± 2.7	68.7 ± 2.6	70.0 ± 2.4	96.3 ± 0.3
5	OB	79.9 ± 1.6	80.0 ± 1.6	79.1 ± 2.0	79.0 ± 1.9	97.7 ± 0.3
6	SIFT-SPM	60.9 ± 4.0	62.3 ± 5.6	57.8 ± 4.3	57.0 ± 4.7	94.4 ± 0.5
7	SURF-ScSPM	73.3 ± 2.1	72.9 ± 2.1	72.3 ± 2.3	72.3 ± 2.4	97.3 ± 0.4
8	GIST	71.5 ± 1.2	71.1 ± 1.4	70.6 ± 1.5	71.2 ± 1.0	95.7 ± 0.4
9	CENTRIST	72.7 ± 1.4	72.8 ± 1.1	72.2 ± 1.9	71.9 ± 1.7	95.9 ± 0.3
10	LBP	71.1 ± 2.9	69.8 ± 4.0	70.3 ± 3.4	69.3 ± 3.9	94.9 ± 1.2
11	Uniform LBP	56.2 ± 3.0	55.6 ± 5.8	54.7 ± 3.2	52.6 ± 3.6	92.8 ± 1.0
12	LBP-HF	64.9 ± 3.4	63.8 ± 4.6	63.9 ± 3.8	62.5 ± 4.3	92.8 ± 1.0
13	PLBP	53.8 ± 3.6	52.7 ± 4.7	52.4 ± 3.6	51.5 ± 4.1	92.1 ± 1.5
14	HMAX	61.1 ± 4.1	60.8 ± 4.0	59.8 ± 4.5	59.9 ± 4.3	79.7 ± 6.9

Table 5 Scene categorization performance on the 15-scene data set using the RBF-SVM

ID	Algorithms	CR (%)	Precision (%)	Recall (%)	F-measure (%)	AUC (%)
1	SIFT-ScSPM	**83.8 ± 1.7**	**84.2 ± 1.6**	**83.2 ± 1.6**	**83.5 ± 1.3**	**98.3 ± 0.0**
2	SIFT-LLC	82.2 ± 1.1	82.5 ± 1.2	81.5 ± 1.1	81.5 ± 1.2	98.2 ± 0.1
3	SIFT-FV	79.4 ± 1.7	78.9 ± 2.1	78.1 ± 2.1	77.6 ± 2.3	97.5 ± 0.2
4	HOG-SPM	78.9 ± 1.2	76.8 ± 3.2	76.3 ± 2.9	76.1 ± 3.1	96.5 ± 0.5
5	OB	73.2 ± 2.0	73.5 ± 1.9	72.0 ± 2.4	72.1 ± 2.3	95.5 ± 0.4
6	SIFT-SPM	72.9 ± 1.3	72.0 ± 1.6	71.6 ± 1.4	71.3 ± 1.4	95.2 ± 0.4
7	SURF-ScSPM	72.4 ± 1.6	71.6 ± 1.8	71.1 ± 1.7	70.6 ± 1.8	95.8 ± 0.4
8	GIST	72.8 ± 1.2	72.2 ± 1.0	71.8 ± 1.4	71.5 ± 1.2	95.2 ± 0.2
9	CENTRIST	72.6 ± 1.8	72.7 ± 1.1	72.0 ± 2.1	71.8 ± 1.8	91.2 ± 1.0
10	LBP	70.9 ± 4.2	69.6 ± 5.2	70.1 ± 4.8	69.2 ± 5.3	96.2 ± 0.7
11	Uniform LBP	67.8 ± 3.9	66.8 ± 5.0	66.8 ± 4.4	65.6 ± 4.9	94.0 ± 0.8
12	LBP-HF	67.3 ± 4.4	66.0 ± 5.4	66.5 ± 4.9	66.3 ± 4.4	94.0 ± 0.8
13	PLBP	69.4 ± 3.5	69.2 ± 4.9	68.2 ± 4.1	68.1 ± 4.4	95.4 ± 1.0
14	HMAX	62.4 ± 3.7	61.6 ± 3.4	61.0 ± 4.0	60.8 ± 3.7	79.5 ± 6.0

indicate that the choice of the SVM kernel does not affect the performance of the LLC and ScSPM significantly.

The SURF-ScSPM is also a BOW-based descriptor. However, it extracts local features from the interest point patches. The *CRs* of SURF-ScSPM is lower than the *CRs* of SIFT-ScSPM, SIFT-LLC, and HOG-SPM, which extract local features from the dense patches. The result indicates that sparse local features can not cover enough information for scene categorization compared to dense local features.

The LBP, uniform LBP, LBP-HF, PLBP, and CENTRIST have higher classification rates than the HMAX on the 15-scene data set. The PLBP has the highest classification rate

Table 6 Scene categorization performance on the 15-scene data set using the HIK-SVM

ID	Algorithms	CR (%)	Precision (%)	Recall (%)	F-measure (%)	AUC (%)
1	SIFT-ScSPM	**83.6 ± 1.6**	**84.2 ± 1.7**	**83.2 ± 1.4**	**83.2 ± 1.4**	**98.2 ± 0.1**
2	SIFT-LLC	82.6 ± 1.5	82.8 ± 1.4	82.1 ± 1.3	82.1 ± 1.3	98.2 ± 0.0
3	SIFT-FV	80.0 ± 1.7	80.1 ± 2.0	79.7 ± 1.4	79.3 ± 1.7	97.7 ± 0.2
4	HOG-SPM	80.0 ± 2.6	79.5 ± 2.5	79.7 ± 2.5	79.6 ± 2.4	97.8 ± 0.2
5	OB	76.9 ± 2.0	76.9 ± 1.8	76.0 ± 2.3	76.0 ± 2.2	97.4 ± 0.2
6	SIFT-SPM	77.9 ± 1.2	77.5 ± 1.3	76.8 ± 1.0	76.8 ± 1.0	97.0 ± 0.3
7	SURF-ScSPM	72.3 ± 2.4	71.4 ± 2.5	71.2 ± 2.5	70.6 ± 2.6	95.6 ± 0.4
8	GIST	72.1 ± 0.7	71.8 ± 0.7	71.4 ± 0.9	72.1 ± 0.9	95.2 ± 0.2
9	CENTRIST	70.9 ± 2.4	71.3 ± 1.8	70.1 ± 2.6	70.0 ± 2.3	95.9 ± 0.4
10	LBP	71.9 ± 2.5	71.0 ± 3.2	71.1 ± 3.2	70.5 ± 3.4	95.7 ± 0.8
11	Uniform LBP	70.6 ± 3.5	70.4 ± 4.4	69.8 ± 4.0	69.3 ± 4.5	95.7 ± 1.0
12	LBP-HF	66.4 ± 3.3	65.6 ± 4.4	65.5 ± 3.7	64.8 ± 4.2	95.7 ± 1.0
13	PLBP	73.0 ± 3.4	72.3 ± 5.1	72.9 ± 4.2	72.6 ± 4.7	96.6 ± 1.0
14	HMAX	63.9 ± 3.4	63.5 ± 3.6	62.5 ± 3.7	62.6 ± 3.6	82.9 ± 8.3

(73.0 %) among the LBP-based descriptors. For the original LBP, uniform LBP, and PLBP, scene categorization performance is better using the HIK-SVM than the linear-SVM and the RBF-SVM. The highest classification rates of LBP, uniform LBP, PLBP are 71.9, 70.6, and 73.0 %, respectively. For the LBP-HF algorithm, a higher classification rate (67.3 %) is achieved using the RBF-SVM, compared to the linear-SVM (64.9 %) and HIK-SVM (66.4 %). The highest *CR* of the CENTRIST is 72.7 %, achieved with the linear-SVM. In fact, the CENTRIST is also a LBP-based algorithm because it extracts local features using the LBP feature map. The difference is that the CENTRIST forms the global feature vector using the spatial PCA, whereas the LBP, uniform LBP, and LBP-HF form the global feature vector using histograms. The PCA with 3 spatial levels accounts for the higher performance of the CENTRIST over the original LBP algorithm.

The OB has a higher classification rate (79.9 %) than the SIFT-SPM, the LBP-based algorithms, and the biologically-inspired algorithms on the 15-scene data set. The OB algorithm achieves its highest classification rate of 79.9 % when using the linear-SVM. A similar observation was reported in (Li et al. 2010); the OB performs better than the SIFT-SPM on the UIUC-sport-event data set, the 15-scene data set, and the 67-indoor scene data set. However, in our experiment, the OB has a lower *CR* than the SIFT-ScSPM and SIFT-LLC on the 15-scene data set. This result indicates that the global feature formation used in the OB is not as good as in the SIFT-ScSPM and SIFT-LLC.

Based on this experiment, we determined a suitable SVM kernel for each of the descriptors. The selected SVM classifiers were used in the subsequent experiments, where we evaluated the visual descriptors on three other data sets: the 8-outdoor-scene, the 67-indoor-scene, and the SUN397 data sets. The aim of the subsequent experiments is to identify the algorithms that had consistent performance on multiple data sets.

4.4.2 Classification results on the 8-outdoor-scene data set

Table 7 shows scene categorization results of the 14 selected descriptors on the 8-outdoor -scene data set. Among the 14 descriptors, the SIFT-ScSPM has the highest *CR* (89.8 %). The

Table 7 Scene categorization performance on the 8-nature-outdoor-scene data set

ID	Algorithms	CR (%)	Precision (%)	Recall (%)	F-measure (%)	AUC (%)
1	SIFT-ScSPM (linear)	**89.8 ± 1.8**	**90.4 ± 1.8**	**90.0 ± 2.2**	**90.0 ± 2.1**	**98.4 ± 0.4**
2	SIFT-LLC (linear)	88.1 ± 2.1	88.6 ± 2.0	88.3 ± 2.6	88.3 ± 2.4	98.2 ± 0.6
3	SIFT-FV (linear)	88.1 ± 3.3	88.2 ± 3.4	88.0 ± 3.9	87.9 ± 3.8	98.7 ± 0.4
4	HOG-SPM(HIK)	88.1 ± 2.5	89.0 ± 2.1	88.2 ± 3.2	88.3 ± 2.9	98.3 ± 0.6
5	OB (linear)	87.5 ± 2.5	88.0 ± 2.3	87.6 ± 3.1	87.5 ± 2.8	98.4 ± 0.5
6	SIFT-SPM(HIK)	87.4 ± 2.3	87.9 ± 2.2	87.8 ± 2.5	87.6 ± 2.4	98.1 ± 0.4
7	SURF-ScSPM (linear)	85.9 ± 2.6	85.4 ± 2.8	85.0 ± 3.4	85.9 ± 3.2	98.0 ± 0.5
8	GIST (RBF)	85.3 ± 2.5	85.9 ± 2.4	85.4 ± 3.1	85.4 ± 2.9	98.0 ± 0.8
9	CENTRIST (linear)	83.5 ± 4.9	84.3 ± 4.8	83.6 ± 5.8	83.5 ± 5.5	97.2 ± 1.1
10	LBP (HIK)	76.6 ± 4.3	77.0 ± 4.6	76.8 ± 5.1	76.5 ± 4.9	95.8 ± 1.2
11	Uniform LBP (HIK)	75.2 ± 4.6	76.0 ± 4.6	75.2 ± 5.8	74.9 ± 5.5	95.0 ± 1.6
12	LBP-HF (RBF)	71.9 ± 3.7	72.1 ± 4.5	71.9 ± 4.7	71.4 ± 4.6	92.5 ± 2.6
13	PLBP (HIK)	81.4 ± 2.9	81.6 ± 3.6	81.3 ± 3.7	81.2 ± 3.8	96.7 ± 1.1
14	HMAX (HIK)	79.8 ± 0.8	80.1 ± 0.7	80.2 ± 0.7	80.0 ± 0.7	96.5 ± 0.3

SIFT-ScSPM, SIFT-LLC, SIFT-FV, HOG-SPM, OB, SIFT-SPM, and SURF-ScSPM are the top-7 algorithms, listed in a descending order of *CR*. For the last 7 descriptors, the classification rates of GIST, CENTRIST and PLBP are higher than 80.0 %. Note that the HMAX achieves a higher *CR* (79.8 %) on the outdoor scene categorization than three LBP-based algorithms (76.6 % for LBP, 75.2 % for uniform LBP, and 71.9 % for LBP-HF). This result indicates that the biologically-inspired features are useful for outdoor scene categorization.

4.4.3 Classification results on the 67-indoor-scene data set

Table 8 shows scene categorization results of the 14 selected descriptors on the 67-indoor-scene data set. The SIFT-ScSPM still has the highest *CR* (45.6 %) compared with the other 13 descriptors. The top-7 descriptors are the SIFT-ScSPM, OB, SIFT-LLC, SIFT-FV, SURF-ScSPM, HOG-SPM, and SIFT-SPM, listed in a descending order of *CR*. These seven algorithms (except for the OB) use BoW methods for global feature formation. These results indicate that using BoW methods is more robust than using histograms, PCA and down-sampling, especially when the complexity of images is increased. Note that the OB descriptor outperformed most of the BoW methods on the 67-indoor-scene data set. This result indicates that the object information is useful for indoor scene categorization.

From Tables 4, 5, 6, 7 and 8, we can see that the ranking based on different measures were consistent for the top-seven algorithms. A higher *CR* was also accompanied by a higher precision, recall, F-measure, and *AUC* values.

4.4.4 Classification results on the SUN397 data set

Using the SUN397 data set, we compared the 14 visual descriptors and 4 recent methods based on deep learning: OverFeat (Sermanet et al. 2014), DeCAF (Donahue et al. 2014), ImageNet-CNN (Krizhevsky et al. 2012), and Places-CNN (Zhou et al. 2014). Note that training a CNN for scene categorization on a large data set requires significant engineering efforts for parameter tuning. To achieve a fair comparison, we evaluated the 14 visual descriptors using

Table 8 Scene categorization performance on the 67-indoor data set

ID	Algorithms	CR (%)	Precision (%)	Recall (%)	F-measure (%)	AUC (%)
1	SIFT-ScSPM (linear)	**45.6 ± 1.0**	**45.4 ± 2.5**	**35.8 ± 1.5**	**36.9 ± 1.7**	**91.7 ± 0.4**
2	SIFT-LLC (linear)	43.9 ± 0.4	44.7 ± 1.2	35.2 ± 0.8	36.3 ± 0.9	91.3 ± 0.5
3	SIFT-FV (linear)	41.5 ± 1.2	46.0 ± 1.3	31.4 ± 1.0	32.5 ± 1.0	90.4 ± 0.6
4	HOG-SPM (HIK)	31.0 ± 0.5	27.5 ± 0.8	25.5 ± 0.6	25.8 ± 0.6	85.1 ± 0.6
5	OB (linear)	45.3 ± 0.5	43.7 ± 1.0	39.4 ± 0.7	39.2 ± 0.7	91.7 ± 0.4
6	SIFT-SPM (HIK)	31.4 ± 0.8	28.4 ± 1.1	25.9 ± 0.7	26.4 ± 0.8	85.3 ± 0.5
7	SURF-ScSPM (linear)	34.5 ± 0.7	35.5 ± 0.9	25.4 ± 0.4	26.2 ± 0.2	87.6 ± 0.3
8	GIST (RBF)	30.9 ± 0.9	28.0 ± 0.1	25.2 ± 0.3	25.7 ± 0.4	84.4 ± 0.4
9	CENTRIST (linear)	12.2 ± 11.0	15.4 ± 8.5	10.8 ± 9.6	10.1 ± 9.6	78.4 ± 3.8
10	LBP (HIK)	22.9 ± 0.6	21.0 ± 1.0	17.7 ± 0.9	18.4 ± 0.9	81.7 ± 0.5
11	Uniform LBP (HIK)	22.0 ± 1.1	18.5 ± 0.4	16.2 ± 1.0	16.6 ± 0.9	80.0 ± 0.6
12	LBP-HF (RBF)	15.4 ± 1.2	12.7 ± 0.8	11.9 ± 0.7	11.8 ± 0.7	76.3 ± 1.0
13	PLBP (HIK)	27.2 ± 1.0	24.1 ± 1.3	21.6 ± 1.0	22.2 ± 1.0	84.5 ± 0.3
14	HMAX (HIK)	11.6 ± 2.3	10.5 ± 2.7	9.4 ± 1.9	9.5 ± 2.1	72.9 ± 3.0

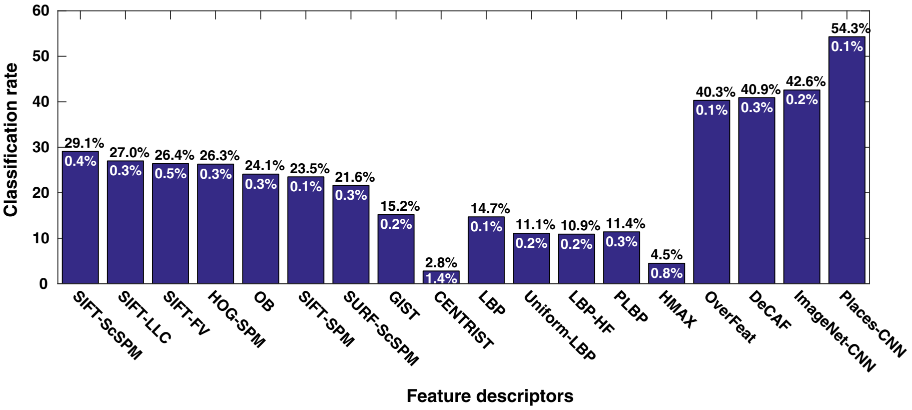


Fig. 9 Comparison of scene categorization methods on the SUN397 data set. For each method, the *top number* in *black* is the classification rate, and the *bottom number* in *white* is the standard deviation

the same evaluation protocol described in (Xiao et al. 2014a) for the SUN397 data set. This data set is divided into fixed partitions. In each partition, 50 training images and 50 test images per class are used for evaluation. The classification rate, averaged over the fixed partitions, is used for comparison. The four deep learning methods had been evaluated using the same protocol, and their results have been reported in (Donahue et al. 2014; Zhou et al. 2014; Xiao et al. 2014a).

Figure 9 presents the *CRs* and their standard deviation of the 14 visual descriptors and the four deep-learning methods (OverFeat, DeCAF, ImageNet-CNN, and Places-CNN) on the SUN397 data sets. All the four deep features outperform the other visual descriptors. However, even the best algorithm (Places-CNN) had a *CR* of only 54.3 %, which is still significantly lower than human performance of 68.0 % (Xiao et al. 2014a).

To rank the data sets in terms of their degree of difficulty, we compared the average classification rates of the top 7 descriptors on each of the four data set. The highest average *CR* of 87.8 % is obtained with the 8-outdoor-scene data set, compared to 79.8 % for the 15-scene data set, 39.0 % for the 67-indoor-scene data set, and 25.4 % for the SUN397 data set. These results indicate that among the four data sets, the 8-outdoor-scene data set is the easiest and the SUN397 data set is the hardest to classify. Similar ranking is obtained if we use the median *CR* on each data set. Apart from the difference in the number of images and scene categories, the 8-outdoor-scene data set consists of only outdoor images, whereas the SUN397 data set contains not only outdoor scenes but also indoor and man-made scenes.

4.5 Class separability and stability of feature vectors

We evaluated the class separability of the feature vectors using the Fisher score S (see Sect. 3.2). Note that this evaluation is independent of the classifier used.

Table 9 presents the class separability scores (S) for the compared features on the four data sets. Among the biologically-inspired features, the HMAX has a low S score on the four data sets. The HMAX features are formed by using max pooling in layer C1, BoW in layer S2, and max pooling in layer C2. The GIST has a higher S score than the HMAX, CENTRIST, OB, and the LBP-based features. As shown in Tables 4, 5, 6, 7 and 8, the GIST algorithm also has higher classification rates than the OB, HMAX, LBP, uniform LBP, and LBP-HF. However, the GIST algorithm has a lower class separability score than BoW algorithms (SIFT-ScSPM, SURF-ScSPM, SIFT-LLC, SIFT-SPM, and HOG-SPM). Note that the GIST forms global features using only down-sampling and averaging. This result indicates that biologically-inspired features can benefit from better schemes for forming global features.

Table 9 The S score for class separability of feature vectors

ID	Algorithms	8-outdoor-scene	15-scene	67-indoor-scene	SUN397
1	SIFT-ScSPM	1.8827	1.3005	1.0364	1.0064
2	SIFT-LLC	2.4339	1.5626	1.1441	1.0061
3	SIFT-FV	**2.5470**	**2.3111**	**1.1484**	**1.0132**
4	HOG-SPM	1.6340	1.6010	1.0959	1.0042
5	OB	1.0003	1.0001	1.0001	1.0000
6	SIFT-SPM	1.4521	1.3573	1.0516	1.0052
7	SURF-ScSPM	1.8323	1.7330	1.0525	1.0060
8	CENTRIST	1.0045	1.0039	1.0000	1.0000
9	GIST	1.0604	1.0283	1.0047	1.0008
10	LBP	1.0434	1.0136	1.0021	1.0004
11	Uniform LBP	1.0187	1.0070	1.0015	1.0003
12	LBP-HF	1.0146	1.0091	1.0012	1.0002
13	PLBP	1.0630	1.0255	1.0034	1.0002
14	HMAX	1.0006	1.0003	1.0004	1.0001

A high value of S means the extracted scene categories are highly separable using the given feature vector

Among the 14 descriptors, the SIFT-FV (which calculates SIFT features on dense-grids, and forms global features using the Fisher kernel coding) has the highest *S* score on the 8-outdoor-scene data set (2.5470), the 15-outdoor-scene data set (2.3111), the 67-indoor-scene data set (1.1484), and the SUN397 data set (1.0132). The result indicates the Fisher vector extracts discriminative global features. Note that in our experiment, the SIFT-FV has lower *CRs* than the SIFT-ScSPM on the four data sets. However, in the paper of Sanchez et al. (2013), the *CR* of SIFT-FV on the SUN397 data set is 43.3 %, which is higher than other hand-designed features compared in our experiments. The reason is the SIFT-FV in (Sanchez et al. 2013) extracts local SIFT features from overlapping patches (24 × 24) on a regular grid every 4 pixels at 5 scales. The SIFT-FV in our experiment extracts SIFT features from overlapping patches (16 × 16) on a regular grid every 8 pixels at one scale. The difference between the *CRs* of SIFT-FV in (Sanchez et al. 2013) and in our paper indicates that informative local features and discriminative global feature formation methods can improve the classification performance.

The other BoW algorithms (SIFT-LLC, SIFT-ScSPM, SIFT-SPM, HOG-SPM, and SURF-ScSPM) also has higher *S* scores than the LBP-based algorithms (LBP, uniform LBP, LBP-HF, PLBP, and CENTRIST) on the four data sets. Note that to form global features, the SIFT-ScSPM and SIFT-LLC combine the BoW algorithms with spatial histograms and max pooling, whereas the LBP-based methods only use histograms.

The SURF-ScSPM has lower *S* score than most of the BOW-based descriptors, such as SIFT-ScSPM, SIFT-LLC, and SIFT-FV, but it has higher *S* score than OB, SIFT-SPM, LBP-based, and biologically-inspired descriptors. This shows that compared to the SIFT-ScSPM, the discriminative power of the SURF-ScSPM is reduced by extracting the SURF features from key points. The *S* score of SURF-ScSPM is 1.8323 on the 8-outdoor-scene data set, 1.7330 on the 15-outdoor-scene data set, 1.0525 on the 67-indoor-scene data set, and 1.0060 on the SUN397 data set.

The LBP-based features (LBP, uniform LBP, PLBP, LBP-HF, and CENTRIST) have lower *S* scores than the BoW features. However, the LBP-based features achieve higher *S* scores than the OB and HMAX. Note that the LBP-based features are more efficient to compute than the BoW features. Among the LBP-based algorithms, the PLBP achieved the highest class separability score. This is also reflected in the higher *CR* of the PLBP, compared to the LBP, uniform LPB, LBP-HF, and CENTRIST (see Sect. 4.4). The uniform LBP and the LBP-HF reduce the dimension and also the class separability of features. This result indicates that the class separability and classification accuracy of the LBP-based algorithms can be improved by using a better global feature formation, instead of the histograms.

The OB has the lowest *S* score among the compared features. Its *S* score is 1.0003 on the 8-outdoor-scene data set, 1.0001 on the 15-outdoor-scene data set, 1.0001 on the 67-indoor-scene data set, and 1.0000 on the SUN397 data set. The reason may be that many similar objects appear in different scene categories. Note that the OB uses only max pooling to form global features, and it extracts a large number of features (44,604 per image).

Next, we evaluated the stability of the scene categorization algorithms in the presence of noise. In this experiment, Gaussian noise with varied standard deviation was added to the original images. Then, the class separability scores were computed for the noisy images. Figure 10 shows the results for the top-7 algorithms: SIFT-ScSPM, SIFT-LLC, SIFT-FV, HOG-SPM, OB, SIFT-SPM, and SURF-ScSPM. These algorithms (based on the bag-of-words) are identified as having high classification accuracy in Sect. 4.4. When noise is added, the class separability (*S* score) of all features reduces. At all noise levels, the SIFT-FV descriptors has a higher *S* score than all other descriptors.

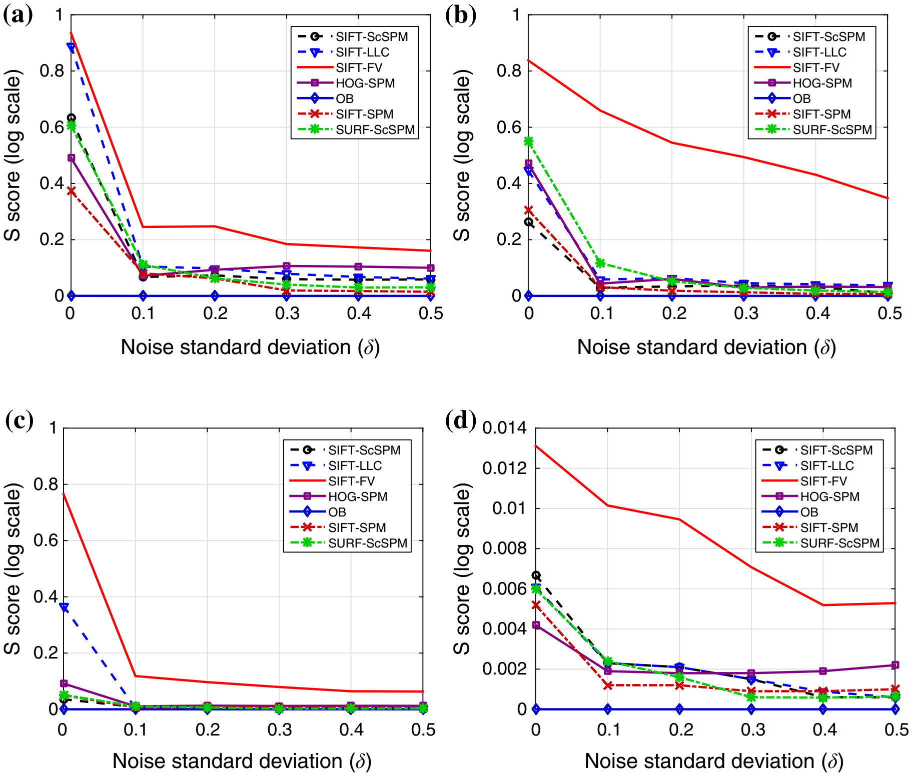


Fig. 10 Stability of features under the presence of Gaussian noise of varying standard deviation, on the four data sets. **a** 8-scene data set, **b** 15-scene data set, **c** 67-indoor-scene data set, **d** SUN397 data set

The results presented in this section indicate that the method for forming global features affects the class separability significantly. Using BoW algorithms before applying histograms, PCA or max pooling (as in SIFT-FV, SIFT-ScSPM, and HOG-SPM) produces feature vectors with more discriminative power. Using histograms as the first step of the global feature formation (as in CENTRIST) decreases the class separability of features. Using only one method for global feature formation (as in the LBP-based algorithms) does not yield high separability scores and nor high classification rates.

5 Conclusion

This paper presented an experimental evaluation of existing visual descriptors for scene categorization. Fourteen descriptors were grouped into three categories: biologically-inspired features, local features, and global feature formation. The existing benchmark data sets and performance measures for scene categorization were also discussed.

The experimental results indicate that the SIFT-ScSPM outperforms all other tested descriptors on the 15-scene data set. The SIFT-ScSPM uses SIFT as its local descriptor and ScSPM as its global feature formation. Local descriptors, SIFT, HOG, and SURF achieve higher classification rates than LBP, CENTRIST, and HMAX. The global feature formation methods affect the class separability of feature vectors significantly. Using ScSPM, LLC, and FV for global feature formation leads to higher class separability than using histograms and PCA. Using BoW before histograms, PCA, and max pooling makes feature vectors more distinguishable. The results on 67-indoor-scene data set show that the mid-level features like objects, bag-of-parts, and the efficient patch encoding algorithm like Fisher vector improve the classification rates for indoor scenes. The results on the SUN397 data set indicate that the SIFT-ScSPM outperforms all other hand-designed descriptors. The learned features produced by deep learning establish the new state-of-the art performance in scene categorization. However, there is still a large performance gap between the best computational algorithm and humans.

Based on this survey and evaluation, several promising research directions can be highlighted. First, local feature descriptors can be built that combine the properties of SIFT, HOG, SURF, GIST, or the early stages of deep learning architecture. A good local descriptor leads to a high classification rate. Second, global feature formation algorithms can be developed based on ScSPM and FV. Third, most existing studies on gist recognition have been concerned with static scenes, which is the focus of this study. In recent years, gist recognition of dynamic scenes has attracted the attention of researchers (Shroff et al. 2010; Derpanis et al. 2012; De Geest and Tuytelaars 2004; Ng et al. 2015), and therefore, the extension of this study to dynamic scenes would be invaluable.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Abe Y, Shikano M, Fukuda T, Arai F, Tanaka Y (1999) Vision based navigation system for autonomous mobile robot with global matching. In: Proceedings of IEEE international conference on robotics and automation, pp 1299–1304

Ahonen T, Hadid A, Pietikainen M (2006) Face description with local binary patterns: application to face recognition. IEEE Trans Pattern Anal Mach Intell 28(12):2037–2041

Ahonen T, Matas J, He C, Pietikainen M (2009) Rotation invariant image description with local binary pattern histogram Fourier features. In: Proceedings of Scandinavian conference on image analysis, pp 61–70

Alahi A, Ortiz R, Vandergheynst P (2012) FREAK: fast retina keypoint. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 510–517

Bai Y, Guo L, Jin L, Huang Q (2009) A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition. In: Proceedings of IEEE international conference on image processing, pp 3305–3308

Bao SY, Sun M, Savarese S (2011) Toward coherent object detection and scene layout understanding. Image Vis Comput 29(9):569–579

Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (SURF). Comput Vis Image Underst 110(3):346–359

Bergamo A, Torresani L (2014) Classemes and other classifier-based features for efficient object categorization. IEEE Trans Pattern Anal Mach Intell 36(10):1988–2001

Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Bo L, Ren X, Fox D (2010) Kernel descriptors for visual recognition. In: Lafferty J, Williams C, Shawe-Taylor J, Zemel RS, Culotta A (eds) Advances in neural information processing systems 23, Curran Associates, Inc., pp 244–252

Boiman O, Shechtman E, Irani M (2008) In defense of nearest-neighbor based image classification. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1–8

Boutell MR, Jiebo L, Brown CM (2006) Factor graphs for region-based whole-scene classification. In: Proceedings of IEEE conference on computer vision and pattern recognition workshop, pp 104–104

Brown M, Susstrunk S (2011) Multi-spectral SIFT for scene category recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 177–184

Brumby SP, Kenyon G, Landecker W, Rasmussen C, Swaminarayan S, Bettencourt LMA (2009) Large-scale functional models of visual cortex for remote sensing. In: Proceedings of IEEE applied imagery pattern recognition workshop, pp 1–6

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2):121–167

Cakir F, Gudukbay U, Ulusoy O (2011) Nearest-neighbor based metric functions for indoor scene recognition. Comput Vis Image Underst 115(11):1483–1492

Castelhano MS, Henderson JM (2008) The influence of color on the perception of scene gist. J Exp Psychol Hum Percept Perform 34(3):660–675

Chang CK, Siagian C, Itti L (2010) Mobile robot vision navigation and localization using gist and saliency. In: Proceedings of IEEE-RSJ international conference on intelligent robots and systems, pp 4147–4154

Chella A, Frixione M, Gaglio S (2001) Conceptual spaces for computer vision representations. Artif Intell Rev 16(2):137–152

Cheng G, Han J, Guo L, Liu T (2015) Learning coarse-to-fine sparselets for efficient object detection and scene classification. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1173–1181

Chin T, Suter D, Wang H (2011) Boosting histograms of descriptor distances for scalable multiclass specific scene recognition. Image Vis Comput 29(4):241–250

Clinchant S, Csurka G, Perronnin F, Renders J (2007) XRCEs participation to ImagEval. In: Proceedings of workshop at content visualization and intermedia representations, pp 1–8

Dai Q, Zhao RW, Wu Z, Wang X, Gu Z, Wu W, Jiang YG (2015) Detecting violent scenes and affective impact in movies with deep learning. In: Proceedings of MediaEval workshop, pp 1–3

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 886–893

De Geest R, Tuytelaars T (2014) Dense interest features for video processing. In: Proceedings of IEEE international conference on image processing, pp 5771–5775

Derpanis KG, Lecce M, Daniilidis K, Wildes RP (2012) Dynamic scene understanding: the role of orientation features in space and time in scene classification. In: Proceedings of IEEE international conference on computer vision and pattern recognition, pp 1306–1313

Dixit M, Chen S, Gao D, Rasiwasia N, Vasconcelos N (2015) Scene classification with semantic fisher vectors. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 2974–2983

Doersch C, Gupta A, Efros AA (2013) Mid-level visual element discovery as discriminative mode seeking. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds) Advances in neural information processing systems 26, Curran Associates, Inc., pp 494–502

Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) DeCAF: a deep convolutional activation feature for generic visual recognition. In: Proceedings of international conference on machine learning, pp 647–655

Douze M, Jegou H, Sandhawalia H, Amsaleg L, Schmid C (2009) Evaluation of gist descriptors for web-scale image search. In: Proceedings of international conference on image and video retrieval, pp 140–147

Ebied HM (2012) Feature extraction using PCA and kernel-PCA for face recognition. In: Proceedings of international conference on informatics and systems, pp 72–77

Elfiky NM, Khan SF, Weijer J, Gonzalez J (2012) Discriminative compact pyramids for object and scene recognition. Pattern Recognit 45(4):1627–1636

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861–874

Fei-Fei L, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 524–531

Fei-Fei L, Fergus R, Perona P (2007a) Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. J Comput Vis Image Underst 106(1): 59–70

Fei-Fei L, Lyer A, Koch C, Perona P (2007b) What do we perceive in a glance of a real-world scene? J Vis 7(1):1–29

Felzenszwalb P, Girshick R, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645

Fuentes-Pacheco J, Ruiz-Ascencio J, Rendon-Mancha JM (2015) Visual simultaneous localization and mapping: a survey. Artif Intell Rev 43(1):55–81

Funt BV, Finlayson GD (1995) Color constant color indexing. IEEE Trans Pattern Anal Mach Intell 17(5):522–529

Gemert JCV, Geusebroek JM, Veenman CJ, Smeulders AWM (2008) Kernel codebooks for scene categorization. In: Proceedings of European conference on computer vision, pp 696–709

Goh H, Thome N, Cord M, Lim JH (2014) Learning deep hierarchical visual feature coding. IEEE Trans Neural Netw Learn Syst 25(12):2212–2225

Gokalp D, Aksoy S (2007) Scene classification using bag-of-regions representations. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1–8

Goldstein EB (2010) Encyclopedia of perception. SAGE, Thousand Oaks

Gowsikhaa D, Abirami S, Baskaran R (2014) Automated human behavior analysis from surveillance videos: a survey. Artif Intell Rev 42(4):747–765

Grossberg S, Huang T (2009) ARTSCENE: a neural system for natural scene classification. J Vis 9(4):1–19

Guangda S, Cuiping Z, Rong D, Cheng D (2002) MMP-PCA face recognition method. Electron Lett 38(25):1654–1656

Guo Z, Zhang D, Zhang D (2010) A completed modeling of local binary pattern operator for texture classification. IEEE Trans Image Process 19(6):1657–1663

Gupta P, Arrabolu SS, Brown M, Savarese S (2009) Video scene categorization by 3D hierarchical histogram matching. In: Proceedings of IEEE international conference on computer vision, pp 1655–1662

Hadjidemetriou E, Grossberg MD, Nayar SK (2004) Multiresolution histograms and their use for recognition. IEEE Trans Pattern Anal Mach Intell 26(7):831–847

Han X, Chen Y (2009) Image categorization by learned PCA subspace of combined visual-words and low-level features. In: Proceedings of international conference on intelligent information hiding and multimedia signal processing, pp 1282–1285

Han Y, Liu G (2010) A hierarchical GIST model embedding multiple biological feasibilities for scene classification. In: Proceedings of international conference on pattern recognition, pp 3109–3112

Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554

Hoiem D, Efros AA, Hebert M (2006) Putting objects in perspective. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 2137–2144

Huang D, Zhu C, Wang Y, Chen L (2014a) HSOG: a novel local image descriptor based on histograms of the second-order gradients. IEEE Trans Image Process 23(11):4680–4695

Huang Y, Wu Z, Wang L, Tan T (2014b) Feature coding in image classification: a comprehensive study. IEEE Trans Pattern Anal Mach Intell 36(3):493–506

Huang Z, Huang H, Zhang W, Hou L (2008) Face recognition using the global image features based on scene gist. J Inf Comput Sci 5(2):919–928

Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol 1(195):215–243

Igarashi M, Mizuno A, Ikebe M (2013) Accuracy improvement of histogram-based image filtering. In: IEEE international conference on image processing, pp 1217–1221

INRIA Graffiti data set (2004) Viewpoint change sequences. http://kahlan.eps.surrey.ac.uk/featurespace/web/data.htm

Jeong S, Won CS, Gray RM (2003) Histogram-based image retrieval using Gauss mixture vector quantization. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing, pp 677–680

Jia D, Wei D, Socher R, Li-Jia L, Kai L, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 248–255

Jia X, Schwing AG, Urtasun R (2014) Tell me what you see and i will show you where it is. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 3190–3197

Jiang A, Wang C, Xiao B, Dai R (2010) A new biologically inspired feature for scene image classification. In: Proceedings of international conference on pattern recognition, pp 758–761

Juneja M, Vedaldi A, Jawahar C, Zisserman A (2013) Blocks that shout: distinctive parts for scene classification. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 923–930

Kadir T, Brady M (2001) Scale, saliency and image description. Int J Comput Vis 45(2):83–105

Karayev S, Fritz M, Darrell T (2014) Anytime recognition of objects and scenes. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 572–579

Kass M, Solomon J (2010) Smoothed local histogram filters. ACM Trans Graph 29(4):100:1–100:10

Kazuhiro H (2011) Local autocorrelation of similarities with subspaces for shift invariant scene classification. Pattern Recogn 44(4):794–799

Ke Y, Sukthankar R (2004) PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 506–513

Khosla A, Byoungkwon A, Lim JJ, Torralba A (2014) Looking beyond the visible scene. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 3710–3717

Kikuchi H, Kataoka S, Muramatsu S, Huttunen H (2013) Color-tone similarity of digital images. In: Proceedings of IEEE international conference on image processing, pp 393–397

Koenderink JJ, Van Doorn AJ (1999) The structure of locally orderless images. Int J Comput Vis 31(2):159–168

Krapac J, Verbeek J, Jurie F (2011) Modeling spatial layout with fisher vectors for image categorization. In: Proceedings of IEEE international conference on computer vision, pp 1487–1494

Krizhevsky A, Sutskever I, Geoffrey EH (2012) ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges C, Bottou L,Weinberger K (eds) Advances in neural information processing systems 25, Curran Associates, Inc., pp 1097–1105

Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 2169–2178

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324

Lee HY, Lee HK, Ha YH (2003) Spatial color descriptor for image retrieval and video segmentation. IEEE Trans Multimedia 5(3):358–367

Lee KW, Choo H (2013) A critical review of selective attention: an interdisciplinary perspective. Artif Intell Rev 40(1):27–50

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am 20(7):1434–1448

Leibe B, Leonardis A, Schiele B (2004) Combined object categorization and segmentation with an implicit shape model. In: Proceedings of European conference on computer vision workshop, pp 17–32

Leutenegger S, Chli M, Siegwart RY (2011) BRISK: binary robust invariant scalable keypoints. In: Proceedings of IEEE international conference on computer vision, pp 2548–2555

Li L, Fei-Fei L (2007) What, where and who? Classifying events by scene and object recognition. In: Proceedings of IEEE international conference on computer vision, pp 1–8

Li L, Su H, Lim Y, Fei-Fei L (2010) Object bank: a high-level image representation for scene classification and semantic feature sparsification. In: Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A (eds) Advances in neural information processing systems 23, Curran Associates, Inc., pp 1378–1386

Li T, Zhu S, Ogihara M (2003) Using discriminant analysis for multi-class classification. In: Proceedings of IEEE international conference on data mining, pp 589–592

Li Z, Liu G, Yang Y, You J (2012) Scale- and rotation-invariant local binary pattern using scale-adaptive texton and subuniform-based circular shift. IEEE Trans Image Process 21(4):2130–2140

Linsley D, MacEvoy SP (2014) Evidence for participation by object-selective visual cortex in scene category judgments. J Vis 14(9):1–17

Liu C, Yuen J, Torralba A (2011) Sift flow: dense correspondence across scenes and its applications. IEEE Trans Pattern Anal Mach Intell 33(5):978–994

Liu GH, Yang JY (2013) Content-based image retrieval using color difference histogram. Pattern Recogn 46(1):188–198

Loschky LC, Sethi A, Simons DJ, Pydimarri TN, Ochs D, Corbeille JL (2007) The importance of information localization in scene gist recognition. J Exp Psychol Hum Percept Perform 33(6):1431–1450

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110

Maggio E, Cavallaro A (2009) Learning scene context for multiple object tracking. IEEE Trans Image Process 18(8):1873–1884

Maji S, Berg AC, Malik J (2008) Classification using intersection kernel support vector machines is efficient. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1–8

Malcolm GL, Nuthmann A, Schyns PG (2014) Beyond gist: strategic and incremental information accumulation for scene categorization. Psychol Sci 25(5):1087–1097

Malhi A, Gao RX (2004) PCA-based feature selection scheme for machine defect classification. IEEE Trans Instrum Meas 53(6):1517–1525

Meng X, Wang Z, Wu L (2012) Building global image features for scene recognition. Pattern Recogn 45(1):373–380

Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630

Morikawa S, Shibata T (2012) Scene image recognition based on the sequence of local image vectors represented by oriented edges. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing, pp 1313–1316

Mutch J, Lowe DG (2006) Multiclass object recognition with sparse, localized features. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 11–18

Ng JYH, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 4694–4702

Ojala T, Pietikainen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distributions. Pattern Recognit 29(1):51–59

Ojala T, Pietikainen M, Maenpaa T (2002a) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987

Ojala T, Pietikainen M, Maenpaa T (2002b) Multiresolution grayscale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987

Oliva A (2000) Diagnostic colors mediate scene recognition. Cogn Psychol 41:176–210

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

Pavlopoulou C, Yu SX (2010) Indoor–outdoor classification with human accuracies: Image or edge gist? In: Proceedings of IEEE conference on computer vision and pattern recognition workshops, pp 41–47

Perina A, Cristani M, Murino V (2010) Learning natural scene categories by selective multi-scale feature extraction. Image Vis Comput 28(6):927–939

Perronnin F, Liu Y, Sanchez J, Poirier H (2010) Large-scale image retrieval with compressed fisher vectors. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 3384–3391

Peters RJ, Itti L (2007) Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1–8

Phillips PJ, Hyeonjoon M, Rizvi SA, Rauss PJ (2000) The FERET evaluation methodology for face-recognition algorithms. IEEE Trans Pattern Anal Mach Intell 22(10):1090–1104

Potter MC (1975) Meaning in visual search. Science 187(4180):965–966

Qian X, Hua XS, Chen P, Ke L (2011) PLBP: an effective local binary patterns texture descriptor with pyramid representation. Pattern Recognit 44(1011):2502–2515

Qin J, Yung NHC (2010) Scene categorization via contextual visual words. Pattern Recognit 43(5):1874–1888

Quattoni A, Torralba A (2009) Recognizing indoor scenes. In: IEEE conference on computer vision and pattern recognition, pp 413–420

Quelhas P, Monay F, Odobez JM, Gatica-Perez D, Tuytelaars T, Van Gool L (2005) Modeling scenes with local descriptors and latent aspects. In: Proceedings of IEEE international conference on computer vision, pp 883–890

Renninger LW, Malik J (2003) When is scene recognition just texture recognition? Vis Res 44:2301–2311

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025

Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe: a database and web-based tool for image annotation. Int J Comput Vis 77(1):157–173

Rybski P, Huber D, Morris D, Hoffman R (2010) Visual classification of coarse vehicle orientation using histogram of oriented gradients features. In: Proceedings of IEEE intelligent vehicles symposium, pp 921–928

Salzberg SL (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Disc 1(3):317–328

Sanchez J, Perronnin F, Mensink T, Verbeek J (2013) Image classification with the fisher vector: theory and practice. Int J Comput Vis 105(3):222–245

Schauerte B, Kuhn B, Kroschel K, Stiefelhagen R (2011) Multimodal saliency-based attention for object-based scene analysis. In: Proceedings of IEEE/RSJ international conference on intelligent robots and systems, pp 1173–1179

Scholkopf B, Sung KK, Burges CJC, Girosi F, Niyogi P, Poggio T, Vapnik V (1997) Comparing support vector machines with Gaussian kernels to radial basis function classifiers. Comput Vis Image Underst 45(11):2758–2765

Scholkopf B, Smola A, Muller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5):1299–1319

Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2014) OverFeat: integrated recognition, localization and detection using convolutional networks. In: Proceedings of international conference on learning representations, pp 1–15

Serrano N, Savakis A, Luo J (2004) Improved scene classification using efficient low-level features and semantic cues. Pattern Recognit 37(9):1773–1784

Serre T, Riesenhuber M (2004) Realistic modeling of simple and complex cell tuning in the hmax model, and implications for invariant object recognition in cortex. Tech. rep., Massachusetts Institute of Technology

Serre T, Wolf L, Poggio T (2005) Object recognition with features inspired by visual cortex. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 994–1000

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007) Robust object recognition with cortex-like mechanisms. IEEE Trans Pattern Anal Mach Intell 29(3):411–426

Shroff N, Turaga P, Chellappa R (2010) Moving vistas: exploiting motion for describing scenes. In: Proceedings of IEEE international conference on computer vision and pattern recognition, pp 1911–1918

Siagian C, Itti L (2007) Biologically-inspired robotics vision Monte-Carlo localization in the outdoor environment. In: Proceedings of IEEE-RSJ international conference on intelligent robots and systems, pp 1723–1730

Snowden R, Thompson P, Troscianko T (2004) Basic vision: an introduction to visual perception. Oxford University Press, New York

Song D, Tao D (2010) Biologically inspired feature manifold for scene classification. IEEE Trans Image Process 19(1):174–184

Song YJ, Park WB, Kim DW, Ahn JH (2004) Content-based image retrieval using new color histogram. In: Proceedings of international symposium on intelligent signal processing and communication systems, pp 609–611

Stricker M, Orengo M (1995) Similarity of color images. In: Proceedings of storage and retrieval for image and video databases, pp 381–392

Sun M, Huang W, Savarese S (2013) Find the best path: an efficient and accurate classifier for image hierarchies. In: Proceedings of IEEE international conference on computer vision, pp 265–272

Swain MJ, Ballard DH (1991) Color indexing. Int J Comput Vis 7(1):11–32

Theodoridis S (2015) Machine learning: a Bayesian and optimization perspective. Academic Press, London

Torralba A, Murphy KP, Freeman WT (2004) Sharing features: efficient boosting procedures for multiclass object detection. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 762–769

Torralba A, Oliva A, Castelhano M, Henderson J (2006) Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev 113(4):766–786

Van de Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. IEEE Trans Pattern Anal Mach Intell 32(9):1582–1596

Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of IEEE Computer Society conference on computer vision and pattern recognition, pp 511–518

Viola P, Jones M (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154

Vogel J, Schwaninger A, Wallraven C, Bülthoff HH (2007) Categorization of natural scenes: local versus global information. ACM Trans Appl Percept 4(3):19

Von Ahn L (2006) Games with a purpose. Computer 39(6):96–98

Walther DB, Chai B, Caddigan E, Beck DM, Fei-Fei L (2011) Simple line drawings suffice for functional MRI decoding of natural scene categories. Proc Nat Acad Sci 108(23):9661–9666

Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 3360–3367

Wei L, Sang N, Wang Y (2010) A biologically inspired object-based visual attention model. Artif Intell Rev 34(2):109–119

Wu J, Rehg JM (2011) CENTRIST: a visual descriptor for scene categorization. IEEE Trans Pattern Anal Mach Intell 33(8):1489–1501

Xiao J, Ehinger KA, Hays J, Torralba A, Oliva A (2014a) SUN database: exploring a large collection of scene categories. Int J Comput Vis 1–20

Xiao Y, Wu J, Yuan J (2014b) mCENTRIST: a multi-channel feature generation mechanism for scene categorization. IEEE Trans Image Process 23(2):823–836

Xie X, Lam KM (2006) Gabor-based kernel PCA with doubly nonlinear mapping for face recognition with a single face image. IEEE Trans Image Process 15(9):2481–2492

Yahoo (2004) Flickr. https://www.flickr.com/

Yang J, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1794–1801

Yang JL, Li HX (2010) PCA based sequential feature space learning for gene selection. In: Proceedings of international conference on machine learning and cybernetics, pp 3079–3084

Yao J, Fidler S, Urtasun R (2012) Describing the scene as a whole: joint object detection, scene classification and semantic segmentation. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 702–709

Yu K, Zhang T, Gong Y (2009a) Nonlinear learning using local coordinate coding. In: Bengio Y, Schuurmans D, Lafferty J, Williams C, Culotta A (eds) Advances in neural information processing systems 22, Curran Associates, Inc., pp 2223–2231

Yu Y, Mann GKI, Gosine RG (2009b) A novel robotic visual perception method using object-based attention. In: Proceedings of IEEE international conference on robotics and biomimetics, pp 1467–1473

Zhang H, Berg AC, Maire M, Malik J (2006) SVM-KNN: discriminative nearest neighbor classification for visual category recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 2126–2136

Zhang L, Ji R, Xia Y, Zhang Y, Li X (2015) Learning a probabilistic topology discovering model for scene categorization. IEEE Trans Neural Netw Learn Syst 26(8):1622–1634

Zhang Z, Li M, Huang K, Tan T (2008) Robust automated ground plane rectification based on moving vehicles for traffic scene surveillance. In: Proceedings of IEEE international conference on image processing, pp 1364–1367

Zhao B, Xing E (2014) Hierarchical feature hashing for fast dimensionality reduction. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 2051–2058

Zheng Y, Shen C, Hartley R, Huang X (2011) Pyramid center-symmetric local binary/trinary patterns for effective pedestrian detection. In: Proceedings of Asian conference on computer vision, pp 281–292

Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using Places database. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K (eds) Advances in neural information processing systems 27, Curran Associates, Inc., pp 487–495

Zhou H, Hermans T, Karandikar AV, Rehg JM (2010) Movie genre classification via scene categorization. In: Proceedings of the international conference on multimedia, pp 747–750