Contextual Exemplar Classifier-Based Image Representation for Classification

Chunjie Zhang, Qingming Huang, and Qi Tian, *Fellow, IEEE*

***Abstract*— The use of local features for image representation has become popular in recent years. Local features are often used in the bag-of-visual-words scheme. Although proven effective, this method still has two drawbacks. First, local regions from which local features are extracted are not discriminative enough for visual tasks. Hence, the combination of local features is necessary. Second, the semantic gap between visual features and human perception also hinders the performance. To address these two problems, in this paper, we propose a novel contextual exemplar classifier-based method for image representation and apply it for classification tasks. Each exemplar classifier is trained to separate one training image from the other images of different classes. We partition each image into a number of regions and use the responses of these exemplar classifiers as the image region's representation. The contextual relationship is then modeled using mixture Dirichlet distributions. A bilayer model is used to predict image classes with L_2 constraints. Experimental results on the Natural Scene, Caltech-101/256, Flower-17/102, and SUN-397 data sets show that the proposed method is able to outperform the state-of-the-art local feature-based methods for image classification.**

***Index Terms*— Computer vision, image processing, pattern classification.**

I. Introduction

LOCAL features are widely used for various visual applications, such as image classification [1], retrieval [2], and segmentation [3]. There are many local features (e.g.,Scale-invariant feature transform (SIFT) [4], histogram of gradients (HoG) [5], speeded up robust features SURF [6], multiple support regions of gradient histogram [7], KAZE [8], color [9], and texture [10]) designed for various visual tasks. Usually, local features are encoded to get the histogram representations of images. This method is very efficient as it makes use of the discriminative power of local features and can also be computed very efficiently.

However, there are mainly two problems with the local feature-based image representation scheme. First, a local region should be predefined in order to extract the local feature. Usually, the local region is determined by dense sampling [11] or by detection [12]. However, local regions are often too small to be discriminative enough for reliable recognition. To solve this problem, local features are often combined together using the predefined rules [13], [14] or by region detection [15]. However, predefining the combining strategy may not be able to cope with the complex situations. Besides, the objectiveness of the region detection method is inconsistent with classification tasks. Another way to solve this problem is by image region selections [16], [17] or by inferring the locations of objects [18].

Second, the semantic gap between visual features and human perception also hinders the performances. To alleviate this drawback, researchers try to use the semantic-based representation methods [19]–[30]. Some try to use the training images directly [19]–[25], while others make use of the information from other sources [26]–[30]. However, training images are often used per class without considering the intra-class and inter-class variations. Besides, the performances decrease with the increment of classes. The use of images from other sources helps to alleviate this problem. However, images from other sources may be biased and contaminated with irrelevant images. Attribute is also used for bridging the semantic discrepancy between the visual features and the human perception [31]–[34]. An attribute is defined as the specification of a property of an object. This strategy is efficient for images with separatable properties. However, attributes are often predefined by human experts. This requires domain knowledge and is also labor-intensive. Besides, there are many images that cannot be efficiently represented by predefined attributes. Exemplar image is also used for semantic representation [21], [22], [35]. This method copes with the intra- and inter-class variations by separating each training image from other images.

To solve the above-mentioned problems, in this paper, we propose a novel contextual exemplar classifier (CEC)-based representation method for image classification tasks. We densely select multiscale image regions with overlap. Each exemplar classifier is trained to separate one training image from the other images of different classes. For each image region, the response of the exemplar classifier

Manuscript received March 29, 2015; revised July 3, 2015, August 13, 2015, September 15, 2015, September 25, 2015, November 30, 2015, and December 18, 2015; accepted January 9, 2016. Date of publication February 8, 2016; date of current version August 2, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61303154 and Grant 61332016, in part by the Open Project of Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, and in part by the National Basic Research Program of China (973 Program) under Grant 2012CB316400 and Grant 2015CB351802. This paper was recommended by Associate Editor R. Hamzaoui. *(Corresponding author: Chunjie Zhang.)*

C. Zhang is with Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China; and also with the Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhangcj@ucas.ac.cn).

Q. Huang is with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China, also with the Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China.

Q. Tian is with the Department of Computer Sciences, University of Texas at San Antonio, San Antonio, TX 78249 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2527380

can indicate the semantic relationship with the particular training image, and hence, bears some semantic meanings. We then model the contextual relationships of exemplar classifiers using mixture Dirichlet distributions. We combine each image region's representation in a predefined order to form a matrix-based image representation. A bilayer model is then learned for image-class predictions. L_2 norm is also used to avoid overfitting. We evaluate the proposed method on several public data sets. The experimental results show the usefulness and efficiency of the proposed method.

The main contributions of this paper lie in three aspects. First, we combine the region-based image representations with exemplar classifiers to obtain the semantic representations of images. This makes the image representations more consistent with human perception compared with directly using visual features. Second, we model the contextual relationships of the exemplar classifier-based representations with mixture Dirichlet distributions. This ensures us to combine the discriminative power of training images for more efficient representations. Third, a bilayer model is proposed to use the relationships of the semantic image representations, which finally improves the classification performances.

The rest of this paper is organized as follows. Related work is given in Section II. The details of the proposed CEC-based method are given in Section III. Section IV gives the experimental results on several public data sets. Finally, the conclusion is drawn in Section V.

II. Related Work

Local features were efficiently used for various visual tasks [1]–[3]. Many local features [4]–[10] had been proposed for different applications. The SIFT feature was proposed in [4] and widely used for classification, retrieval, and matching. To speed up the computation, the HoG feature was proposed in [5]. Bay *et al.* [6] proposed the SURF descriptor which was scale and rotation invariant and was also robust to noise. Fan *et al.* [7] tried to aggregate the gradient information to intensity orders for object matching. Alcantarilla *et al.* [8] proposed the KAZE feature which used an additive operator-splitting technique for nonlinear scale space. Rao *et al.* [9] explored the color information for person detection to cope with the variations of different scenarios. Nguyen *et al.* [10] used support local pattern for visual applications.

Usually, the local regions from which local features were extracted were often determined by dense sampling [11] or by detection [12]. Local features were often combined to increase the discriminative power [13]–[15]. Ni *et al.* [13] modeled the local feature's spatial context by random forest, while Zhang *et al.* [14] used the Haar-like transformation. Wu *et al.* [15] tried to bundle local features for partial-duplicate image search with heavy computational cost. Other researchers leveraged less computational cost methods for spatial information usages [16]–[18], [36], [37]. Lazebnik *et al.* [16] proposed the spatial-pyramid matching technique that was widely used with good performance. Zhang *et al.* [17] used the component for image representation and proposed boosted bilinear model for object recognition. In order to make use of the detection information, Russakovsky *et al.* [18] tried to detect objects and then separate the object and background representations. Marin *et al.* [36] used random forests for pedestrian detection, while Wang *et al.* [37] detected human action as the spatiotemporal tube.

To get semantically meaningful representations, many works had been done [19]–[30]. Some researchers made use of the training samples for semantic modeling. Rasiwasia and Vasconcelos [19] proposed to use low-dimensional semantic spaces for scene classification, while Vogel and Schiele [20] used the semantic modeling for image retrieval. A weak semantic representation was proposed in [21] and then extended as subsemantic space [22]. A holistic context model was used in [23] for image classification, while Oliva and Torralba [24] used the holistic representation of scenes for recognition. Inoue and Shinoda [25] proposed a fast video semantic-indexing method. Other researchers tried to construct semantic spaces using images collected from the Internet. Yang *et al.* [26] used Web images for semantic video indexing by minimizing the sample-specific loss. Hauptmann *et al.* [27] evaluated the usage of semantic concepts for video retrieval using the broadcast news. Russell *et al.* [28] introduced the Labelme data set. An interactive image-tagging scheme was proposed in [29] by reducing the human-labeling effort. The object bank was proposed in [30] to semantically represent images. Vázquez *et al.* [38] explored the domain adaptation problem for human detection with active learning, while Xu *et al.* [39] proposed the domain adaptation of deformable part-based models and improved the detection accuracy.

Attribute-based methods [31]–[34] were also widely explored. Torresani *et al.* [31] proposed to use classesmes for object categorization. Farhadi *et al.* [32] described objects by attributes, while Parikh and Grauman [33] used the relative attributes to model the relativeness of different attributes for ranking. In order to detect unseen objects, Lampert *et al.* [34] proposed an attribute transfer method among image classes. The exemplar classifier-based method was also widely explored [21], [22], [35], [40]. Malisiewicz *et al.* [35] used an exemplar support vector machines for object detection in an ensemble way and then extended to subsemantic spaces [22]. Xu *et al.* [40] adapted the pedestrian detector using the boosted-latent dirichlet allocation exemplar classifiers.

III. Classification With Contextual Semantic Space-Based Representation

In this section, we give the details of the proposed contextual semantic space-based representation method for image classification. We densely extract image regions with overlap and train exemplar classifiers. The responses of exemplar classifiers are combined with mixture Dirichlet distributions to get the semantic representations of image regions. A bilayer model is learned to predict image classes with L_2 constraints. Fig. 1 shows the flowchart of the proposed method.

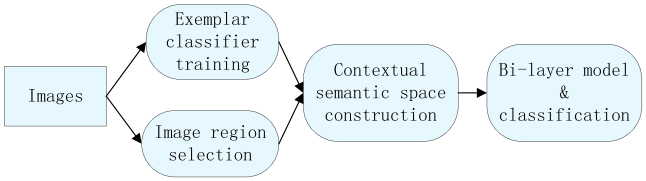


Fig. 1. Flowchart of the proposed contextual semantic space-based representation method for image classification.

A. Exemplar Classifier Training and Image Region Selection

We use the bag-of-visual-words model (BoW) with spatial pyramid ($L = 0, 1, 2$) as the initial image representation [16]. Exemplar classifiers are then trained to separate each training image from the other images of different classes. Let $\boldsymbol{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}] \in \mathbb{R}^{D\times N}$ be the N training images with the corresponding labels $\boldsymbol{Y} = [y_1, \ldots, y_N]$ of C classes. Linear classifier is used as the exemplar classifier

$$\hat{y}_{n,i} = \boldsymbol{w_n}^T \boldsymbol{x_i} + b_n, \quad n = 1, \ldots, N; \quad i = 1, \ldots, N \tag{1}$$

where $(\boldsymbol{w_n}, b_n)$, $n = 1, \ldots, N$ are the parameters for the n-th exemplar classifier. $\hat{y}_{n,i}$ is the predicted label of the n-th exemplar classifier for the i-th image. To learn the parameters, we try to minimize the summed loss of training images as

$$(\boldsymbol{w_n}, b_n) = \underset{(\boldsymbol{w_n}, b_n)}{\operatorname{argmin}} \|\boldsymbol{w_n}\|^2 + \lambda_1 \ell(\hat{y}_{n,n}, y_n) + \sum_{i=1, y_i \neq y_n}^{N} \ell(\hat{y}_{n,i}, y_n) \tag{2}$$

where λ_1 is the parameter, which controls the relative influences of the exemplar image, and $\ell(*, *)$ is the loss function. For classification tasks, the hinge loss is often used. However, the hinge loss does not differentiable. To alleviate this problem, we adopt the quadratic hinge loss as

$$\ell(\hat{y}_{n,i}, y_n) = (\max(\hat{y}_{n,i} \times y_n - 1, 0))^2 \tag{3}$$

and learn the exemplar classifier for each training image. Besides, to use the spatial and context information of local features, we use image region as the basic element for image representation.

Suppose, we select M regions for the n-th image with the BoW representation as $\overline{\boldsymbol{x}}_n^m \in \mathbb{R}^{D\times 1}, m = 1, \ldots, M$. We use the responses of exemplar classifiers as the initial semantic representations of image regions: $\widetilde{\boldsymbol{x}}_n^m = [\widetilde{x}_{n,1}^m; \ldots; \widetilde{x}_{n,N}^m] \in \mathbb{R}^{N\times 1}$. Each dimension can be calculated as

$$\widetilde{x}_{n,i}^m = \boldsymbol{w_i}^T \overline{\boldsymbol{x}}_n^m + b_i, \quad i = 1, \ldots, N. \tag{4}$$

We combine $\widetilde{\boldsymbol{x}}_{n,i}^m$ in a predefined order (left to right and top to bottom) as $\widetilde{\boldsymbol{X}}_n = [\widetilde{\boldsymbol{x}}_n^1, \ldots, \widetilde{\boldsymbol{x}}_n^M]$. Instead of representing each image as a vector, we go one step beyond by representing the image as a matrix. Note that this strategy is different from simply organizing the local features into the matrix form. By imposing image regions, we can cope with the variations of images. Besides, it is also more semantically meaningful than directly using local features.

B. CEC-Based Image Representation

We can use $\widetilde{\boldsymbol{X}}_n$ for image representation and apply it for classification tasks directly. However, since each exemplar classifier is learned with only one positive sample, it may be unable to distinguish the complex semantics well. Besides, the dimension of exemplar-based representation increases with the number of training images N.

We use the mixture Dirichlet distributions to model the contextual relationships of exemplar classifier-based representation. Let $\widetilde{\boldsymbol{X}} = [\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_N]$, the mixture Dirichlet distribution can be written as

$$P_{\widetilde{\boldsymbol{X}}|Y}(\widetilde{\boldsymbol{x}}|y, \Lambda^y) = \sum_{k=1}^{K} \gamma_c^y \mathrm{Dir}(\widetilde{\boldsymbol{x}}; \boldsymbol{\xi}_k^y) \tag{5}$$

where $\Lambda^y = \{\gamma_c^y, \boldsymbol{\xi}_k^y\}$ are the parameters with $\sum_c \gamma_c^y = 1$. Y is a random variable defined on the classes of images $y \in \{1, \ldots, C\}$. K is the mixture number. The Dirichlet distribution $\mathrm{Dir}(\widetilde{\boldsymbol{x}}; \boldsymbol{\xi})$ is calculated as

$$\mathrm{Dir}(\widetilde{\boldsymbol{x}}; \boldsymbol{\xi}) = \Gamma\left(\sum_{n=1}^{N} \xi_n\right) \prod_{n=1}^{N} (\widetilde{\boldsymbol{x}}_n)^{\xi_n - 1} \Big/ \prod_{n=1}^{N} \Gamma(\xi_n) \tag{6}$$

where $\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_N\}$. $\Gamma(.)$ is the Gamma function. We learn the parameters using the generalized expectation-maximization algorithm [23]. An image region's posterior probability is predicted as

$$P_{Y|\widetilde{\boldsymbol{X}}}(y|\widetilde{\boldsymbol{x}}) = P_{\widetilde{\boldsymbol{X}}|Y}(\widetilde{\boldsymbol{x}}|y) P_Y(y) / P_{\widetilde{\boldsymbol{X}}}(\widetilde{\boldsymbol{x}}) \tag{7}$$

where $P_Y(y)$ and $P_{\widetilde{\boldsymbol{X}}}(\widetilde{\boldsymbol{x}})$ are set to $1/C$ and $1/N$, respectively. The posterior probabilities are used for image region's representation $\boldsymbol{h}$ as $\boldsymbol{h} = (P_{Y|\widetilde{\boldsymbol{X}}}(1|\widetilde{\boldsymbol{x}}), \ldots, P_{Y|\widetilde{\boldsymbol{X}}}(K|\widetilde{\boldsymbol{x}}))^T$. In this way, we can ensure the new semantic representation does not increase with the number of training images as long as the number of mixture distributions is fixed.

C. BiLayer Model for Classification

Each image region can be classified as belonging to a particular class using $\boldsymbol{h}$. We use a nonlinear function $f(\boldsymbol{h})$ to model this relationship as

$$z = f(\boldsymbol{h}) = \boldsymbol{\alpha}^T \boldsymbol{q} \tag{8}$$

with $\boldsymbol{q}$ defined as the linear and quadratic correlations between each element of $\boldsymbol{h}$ as

$$\boldsymbol{q} = [h_1; \ldots; h_K; h_1 h_1; h_1 h_2; \ldots; h_i h_j; \ldots; h_K h_K]. \tag{9}$$

We use the linear term to model the influences of each semantic index and use the quadratic term to model the correlations. Higher order relationships can also be incorporated to improve the performance with the dimension of $\boldsymbol{q}$ increases rapidly.

Let $\boldsymbol{z} = [z_1; \ldots; z_M]$, we use the linear combination of each z_m, $m = 1, \ldots, M$ for the image-level prediction as

$$y = \boldsymbol{z}^T \boldsymbol{\beta}. \tag{10}$$

Equations (8) and (10) can be rewritten in a matrix form as

$$y = \boldsymbol{\alpha}^T \boldsymbol{Q} \boldsymbol{\beta} \tag{11}$$

with $Q = [q^1, \ldots, q^M]$.

Algorithm 1 Procedures for Solving (14)

Input:

The initial parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$; the regularization parameters λ_2 and λ_3; training images ($\boldsymbol{Q_n}$, y_n), n=1,…, N; maximum iteration number $maxiter$; stopping threshold θ

Output:

The learned parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$;

1: **for** $iter = 1, 2, \ldots, maxiter$

2: Solve for the optimal $\boldsymbol{\alpha}$ while keeping $\boldsymbol{\beta}$ fixed by solving Eq. 15;

3: Solve for the optimal $\boldsymbol{\beta}$ while keeping $\boldsymbol{\alpha}$ fixed by solving Eq. 16;

4: Calculate the change of objective value of Eq. 14 falls below θ.

If not satisfied,

Go to Step 1;

else

Break, go to step 5.

5: **return** The learned encoding parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$.

To learn the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we try to minimize the prediction errors on the training images as

$$[\boldsymbol{\alpha}, \boldsymbol{\beta}] = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{n=1}^{N} \operatorname{loss}\left(\boldsymbol{\alpha}^T \boldsymbol{Q_n} \boldsymbol{\beta}, y_n\right) \tag{12}$$

where $\boldsymbol{Q_n}$ is the new image representation with its corresponding label as y_n, $n = 1, \ldots, N$. We use the exponential loss as

$$\operatorname{loss}(\boldsymbol{\alpha}^T \boldsymbol{Q} \boldsymbol{\beta}, y) = e^{-\boldsymbol{\alpha}^T \boldsymbol{Q} \boldsymbol{\beta} \times y}. \tag{13}$$

To avoid overfitting, we add L_2 norm to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The final objective function can be written as

$$[\boldsymbol{\alpha}, \boldsymbol{\beta}] = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{n=1}^{N} e^{-\boldsymbol{\alpha}^T \boldsymbol{Q} \boldsymbol{\beta} \times y} + \lambda_2 \|\boldsymbol{\alpha}\|_2^2 + \lambda_3 \|\boldsymbol{\beta}\|_2^2 \tag{14}$$

where λ_2 and λ_3 are the two parameters, which control the L_2 norm influences of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. However, (14) cannot be optimized jointly over $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Hence, we resort to the alternative optimization strategy by optimizing over $\boldsymbol{\alpha}/\boldsymbol{\beta}$ while keeping $\boldsymbol{\beta}/\boldsymbol{\alpha}$ fixed. When $\boldsymbol{\beta}$ is fixed, the optimization over $\boldsymbol{\alpha}$ can be solved as

$$[\boldsymbol{\alpha}] = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{n=1}^{N} e^{-\boldsymbol{\alpha}^T \boldsymbol{Q} \boldsymbol{\beta} \times y} + \lambda_2 \|\boldsymbol{\alpha}\|_2^2. \tag{15}$$

When $\boldsymbol{\alpha}$ is fixed, the optimization over $\boldsymbol{\beta}$ can be solved as

$$[\boldsymbol{\beta}] = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{n=1}^{N} e^{-\boldsymbol{\alpha}^T \boldsymbol{Q} \boldsymbol{\beta} \times y} + \lambda_3 \|\boldsymbol{\beta}\|_2^2. \tag{16}$$

This alternative optimization over $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be iterated for a predefined times or the decrease of objective value of (14) falls below a threshold. Algorithm 1 gives the procedures for solving (14).

IV. Experiments

To evaluate the effectiveness of the proposed CEC-based method, we conduct experiments on several public image data sets: the Natural Scene data set [16], the Caltech-101 data set [41], the Caltech-256 data set [42], the Flower-17 data set [43], and the Flower-102 data set [44]. Fig. 2 shows some example images of these data sets.

A. Experimental Setup

To extract local features, we densely choose local regions with multiscales. Images are first resized to the same size for each data set. The smallest local region is set to 16×16 pixels with an overlap of 6 pixels. We use the sparse-coding technique [45] and set the codebook size to 1000. Max pooling with spatial pyramid matching ($L = 0, 1, 2$) is used to get the initial BoW representation of images. The minimum size of image region is set to 64×64 pixels with 16 pixels overlap. Hence, we can get about 150 image regions for a 300×200 image. The initial image region's visual representation is obtained by max pooling the encoded parameters within this region. The learned exemplar classifiers are then used to predict image region's classes. We set the maximum number of iterations in Algorithm 1 to 50. For a fair comparison, we compare the performances reported by other methods directly. We also give the performances of using CEC-based representation on the image level (CE).

B. Natural Scene Data Set

There are 15 classes of images in this data set. Each class has 200–400 images with the average size of 300×250 pixels. We resize all the images to 300×250 pixels and randomly select 100 images per class for training. The other images are used for the performance evaluation. This process is repeated for ten times to get reliable results.

We give the performance comparisons of CE and CEC with other methods [11], [16], [19], [21]–[23], [30], [45]–[47] in Table I. Since the visual representations of images do not have semantic correspondences with human perception, directly using the text-processing technique cannot achieve comparable performances [19], [46] as the semantic-based methods. The visual only based methods [11], [16], [45], [47] improve the recognition by designing discriminative models and making use of the spatial and correlation relationships, and hence, can boost the performances compared with [19] and [46]. The semantic-based methods [19], [21]–[23], [30] can alleviate the semantic and visual discrepancy problems and can improve the recognition accuracy. However, the histogram-based image representation cannot fully explore the spatial layout and the structure relationships of visual features. By combining semantic correspondences with image regions for joint representation, the proposed CEC can eventually outperform visual only and semantic-only-based methods. Besides, compared with [23], which also explores the contextual relationships of semantics, CEC can make use of the discriminative property of exemplar classifiers for representation. The use of region-based strategy also helps to make use of more information. Moreover, compared

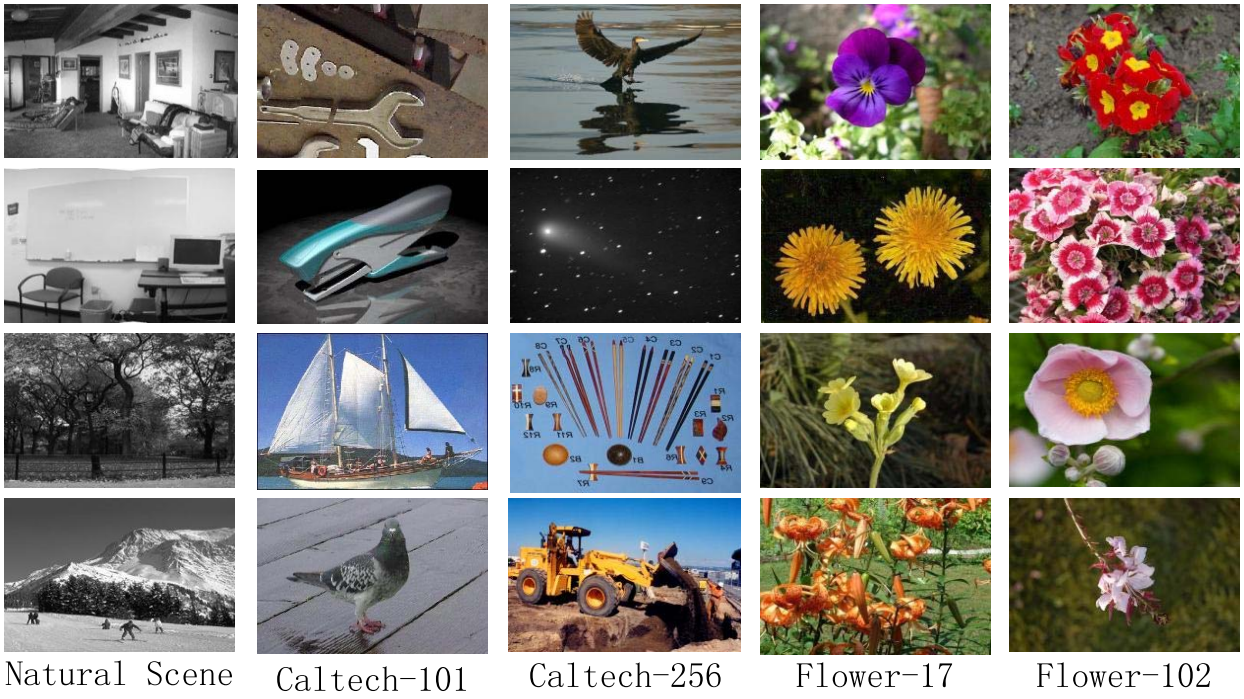


Fig. 2. Example images of the Natural Scene data set, the Caltech-101 data set, the Caltech-256 data set, the Flower-17 data set, and the Flower-102 data set.

TABLE I

PERFORMANCE COMPARISON OF THE PROPOSED CONTEXTUAL EXEMPLAR SPACE-BASED IMAGE REPRESENTATION AND CLASSIFICATION METHOD WITH OTHER METHODS ON THE NATURAL SCENE DATA SET. SS: SEMANTIC SPACE, CM: CONTEXTUAL MODELS, AND KC: KERNEL CODEBOOK. NUMERICAL VALUES STAND FOR MEAN AND STANDARD DERIVATION

Algorithms	Classification Rate
L^2BLM[11]	83.40 ± 1.30
SBLM[11]	85.60 ± 1.50
SPM[16]	81.40 ± 0.50
LDA[19]	59.0
SS[19]	73.95 ± 0.74
WSR-EC[21]	81.54 ± 0.59
S^3R[22]	83.72 ± 0.78
CM[23]	77.20 ± 0.39
ObjectBank [30]	80.9
ScSPM[45]	80.28 ± 0.93
pLSA[46]	72.7
KC[47]	76.67 ± 0.39
CE	83.20 ± 0.53
CEC	87.59 ± 0.48

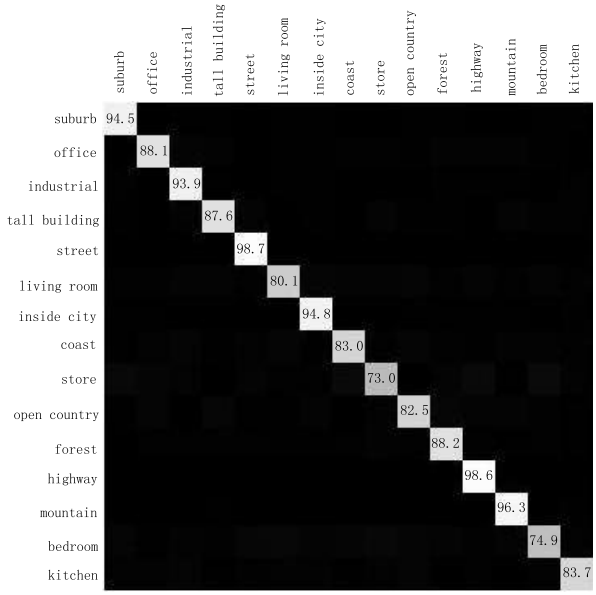


Fig. 3. Confusion matrix on the Natural Scene data set.

with [11], which uses the bilinear model with visual clues only, CEC jointly uses the visual and semantic information. We give the confusion matrix in Fig. 3.

C. Caltech-101/256 Data Sets

There are more than 9000 images of 101 classes in the Caltech-101 data set. The number of images per class varies from 31 to 80. The 15 and 30 training images per class are randomly selected for training. We repeat the random selection process for ten times.

The performance comparisons of CEC with others [11], [38], [42], [47]–[51] are given in Table II. Compared with the visual-based methods, the proposed CEC method can improve the recognition accuracy dramatically. By learning discriminative models instead of using visual distances [48] directly, CEC can outperform [48] by 9.5%/8.4% with 15/30 training images, respectively. Besides, using the matrix-based technique can preserve more information for accurate recognition compared with the histogram-based methods, and hence, the proposed CEC and sparsity-constrained bilinear model (SBLM) [17] are able to

TABLE II
PERFORMANCE COMPARISONS ON THE CALTECH-101 DATA SET

Methods	15 training	30 training
SBLM [17]	71.68 ± 1.12	77.50 ± 0.77
KSPM [41]	56.40	64.40 ± 0.80
ScSPM [45]	67.00 ± 0.45	73.20 ± 0.54
KC [47]	–	64.14 ± 1.18
NBNN [48]	65.00 ± 1.14	70.40
LLC [49]	65.43	73.44
KMTJSRC [50]	65.00 ± 0.70	–
SVM-KNN [51]	59.10 ± 0.60	66.20 ± 0.50
CNN [54]	–	84.77 ± 0.70
CE	69.28 ± 0.75	75.10 ± 0.58
CEC	74.52 ± 0.68	78.89 ± 0.63
CEC-CNN	–	86.14 ± 0.69

outperform [41], [45], [46], [49], [51]. Moreover, by only using the SIFT features, the proposed CEC can outperform [50], which combines the different types of features. This proves the effectiveness of the proposed method.

The Caltech-256 data set is an extension of the Caltech-101 data set. There are 29 780 images of 256 classes. The intra-class variability is also larger compared with the Caltech-101 data set. There are at least 80 images per class. 15/30/45 training images per class are randomly selected for ten times for the performance evaluation.

We give the classification performances of CEC and others [11], [41], [45], [47]–[51] in Table III. We can have similar conclusions as on the Caltech-101 data set. The proposed CEC is able to outperform many visual- and semantic-based methods. This is not only because we explore the discriminative power of exemplar classifiers and contextual relationships, but also because the use of region-based representation can incorporate more spatial and structure information. Besides, CEC is able to outperform ObjectBank [30] which leverages images beyond training samples. Moreover, CEC also outperforms methods which only use exemplar semantic spaces [21], [22].

The proposed method can also be combined with the other visual-based representation methods, such as the Fisher Vector (FV) [53] and the convolutional neural network (CNN) [54]. This is achieved by first obtaining the visual representation of images using the FV/CNN and then using the proposed CEC-based image representation method for classification. In this way, we are able to achieve 49.7% classification rate on the Caltech-256 data set when 30 training images are used. This outperforms FV [53] by 2.3%. Besides, on the Caltech-101 data set, we are able to achieve a classification rate of 86.1% which is 1.4% better than the CNN-based strategy when 30 training images per class are used.

D. Flower-17/102 Data Sets

There are 17 classes (*buttercup, colts' foot, daffodil, daisy, dandelion, fritillary, iris, pansy, sunflower, windflower, snowdrop, lilyvalley, bluebell, crocus, tigerlily, tulip*, and *cowslip*) of flower images in the Flower-17 data set. Each class has 80 images with a total number of 1360 images. We follow the same experimental setting as in [26] using the three splits of images (40/20/20 for train/validate/test, respectively). As color

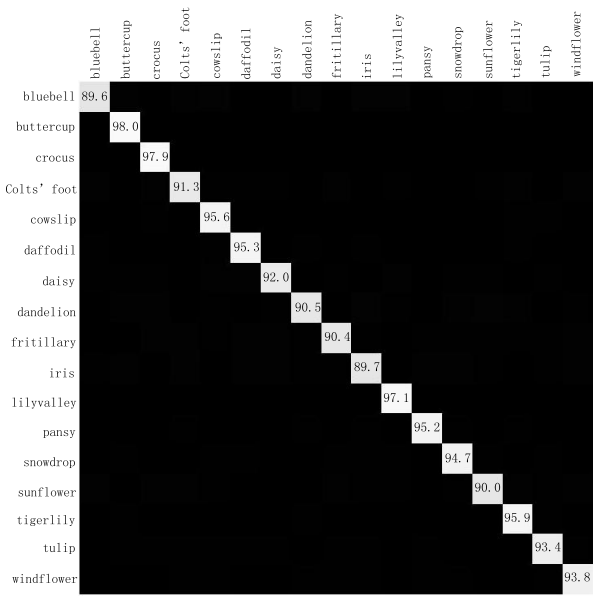


Fig. 4. Confusion matrix on the Flower-17 data set.

plays an important role for flower recognition, we also extract color SIFT features [55].

There are 8189 images of 102 classes with 40–250 images per class in the Flower-102 data set. This data set is more difficult to distinguish than the Flower-17 data set, as there are more images and classes. We follow the same experimental setup as [44] and [50] did and use ten images per class for training, ten images per class for validation, and the rest of images for testing.

Tables IV and V give the performance comparisons on the Flower-17 data set and the Flower-102 data set, respectively. For the Flower-17 data set, we also give the confusion matrix in Fig. 4. Compared with the simple local feature transformation [14], CEC makes use of the local features more sufficiently, and hence, it outperforms [14] by 3.5%/2.5% on the Flower-17/102 data sets, respectively. Besides, the use of discriminatively trained classifiers is more efficient than simple nearest-neighbor-based strategy [47]. Moreover, region-based representation is more efficient than the histogram-based methods [43], [56]–[58]. These results again demonstrate the effectiveness of the proposed CEC method.

E. Parameter Influences

Larger λ_1 places more importance on the exemplar training image, while smaller λ_1 encourages the learned classifier to predict images as not belonging to the particular class. We give the performance changes with λ_1 in Fig. 5. We can observe from Fig. 5 that the proposed method is able to achieve a good performance with a wide range of λ_1. This is because the proposed CEC does not rely on the output of one particular exemplar classifier, but combines them for joint representation. The performances are relatively stable as long as the learned exemplar classifiers can make consistent predictions. This increases the robustness of the proposed CEC method.

λ_2 and λ_3 are the two parameters, which control the regularization terms of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Larger λ_2/λ_3 places more importance on the regularization term $\boldsymbol{\alpha}/\boldsymbol{\beta}$. We set λ_2 and λ_3

TABLE III
PERFORMANCE COMPARISONS ON THE CALTECH-256 DATA SET

Methods	15 training	30 training	45 training
SBLM [17]	35.60 ± 0.87	42.93 ± 060	–
WSR-EC[21]	35.28 ± 0.65	42.01 ± 0.47	45.82 ± 0.54
S^3R[22]	37.85 ± 0.48	43.52 ± 0.44	46.86 ± 0.63
ObjectBank [30]	–	39.00	–
Classemes [31]	–	36.00	–
KSPM [42]	–	34.10	–
KSPM [45]	23.34 ± 0.42	29.51 ± 0.52	–
ScSPM [45]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55
KC [47]	–	27.17 ± 0.46	–
NBNN [48]	30.45	38.18	–
LLC [49]	34.36	41.19	45.31
LScSPM[52]	30.00 ± 0.14	35.74 ± 0.10	38.54 ± 0.36
FV[53]	38.5	47.4	52.1
CE	35.43 ± 0.41	42.68 ± 0.39	45.96 ± 0.38
CEC	39.20 ± 0.37	44.75 ± 0.33	47.93 ± 0.31
CEC-FV	40.50 ± 0.42	49.71 ± 0.36	48.81 ± 0.34

TABLE IV
PERFORMANCE COMPARISON ON THE FLOWER-17 DATA SET

Methods	Classification rate
Harr-SIFT [14]	90.15 ± 1.39
Nilsback and Zisserman [43]	71.76 ± 1.76
KMTJSRC-CG [50]	88.90 ± 2.30
Varma and Ray [56]	82.55 ± 0.34
χ^2 [57]	87.45 ± 1.13
LP-B [58]	85.40 ± 2.40
CE	89.51 ± 1.33
CEC	93.70 ± 1.06

TABLE V
PERFORMANCE COMPARISON ON THE FLOWER-102 DATA SET

Methods	Classification rate
Harr-SIFT [14]	75.6
Nilsback and Zisserman [44]	72.8
KMTJSRC-CG [50]	74.1
Xie et al. [59]	86.8
CE	75.3
CEC	78.1

to be equal in this paper. We give the influences of λ_2/λ_3 in Fig. 6. We can observe from Fig. 6 that if we set λ_2/λ_3 too large or too small, the performances decrease. We set λ_2/λ_3 to 0.2 for the Natural Scene data set and Flower-17 data set. 0.8 is used for the rest three data sets.

The mixture Dirichlet number K is another important parameter which influences the performances. Fig. 7 shows the performance changes with K on the five data sets. We can observe that a small K is unable to separate images well. With the increment of K, we can gradually improve the performance. However, if K is too large, the computational cost increases as the image representation is of $O(K^2)$.

F. Convergence

During each iteration, the optimization over (15) and (16) can reduce the objective value of (14). Besides, the objective value of (14) is larger than zero. This means, we can gradually reduce the objective value of (14). For intuitive illustration, we give the changes of objective values of (14) with the number of iterations in Fig. 8.

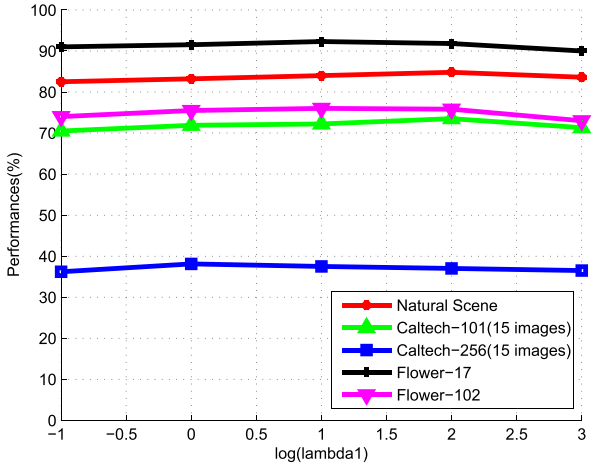


Fig. 5. Performances change with λ_1 on the five data sets.

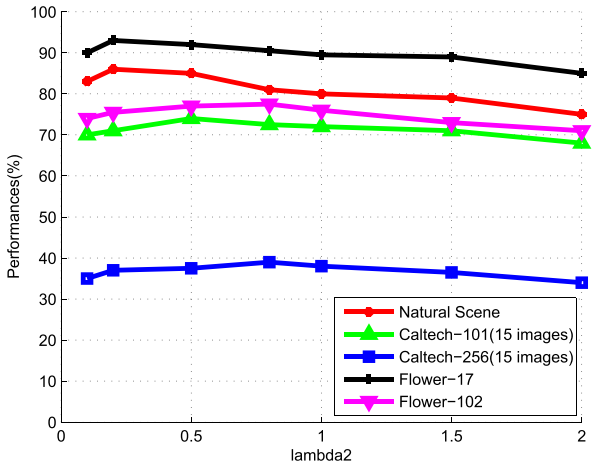


Fig. 6. Performances change with λ_2/λ_3 on the five data sets.

G. SUN-397 Data Set

We also evaluate the proposed method on the SUN-397 data set [60]. This data set has 397 classes of

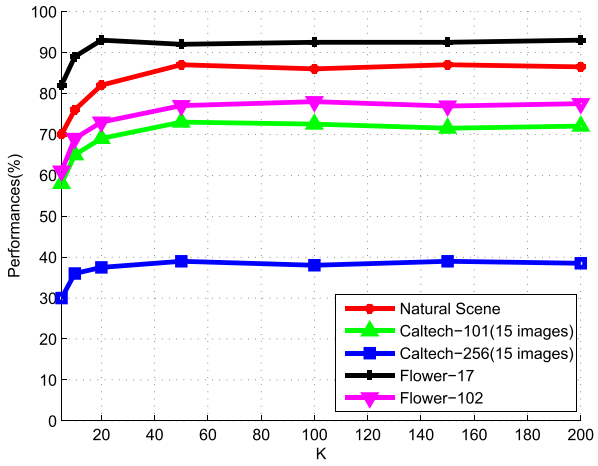


Fig. 7. Performances change with K on the five data sets.

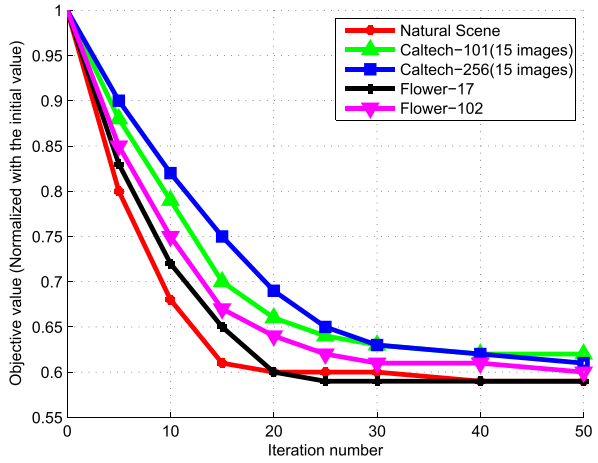


Fig. 8. Objective value changes of (14) with the iteration number on the five data sets (the objective value is normalized using the initial value to show the relative decrease of objective values).

approximately 100K images. We follow the same experimental setup in [60] did and use 5/50 images per class for training/testing, respectively, for ten times. We use the combined feature matrix provided in [60] to make use of various types of features (GIST, HoG, Dense SIFT, and so on). The proposed CEC/CE achieves 16.2%/15.1% accuracy, which outperforms [60] by 1.7%/0.6%, respectively.

V. Conclusion

In this paper, we proposed a novel contextual semantic space-based method for image representation and classification. Exemplar classifiers were trained to separate each training image from the other images of different classes. We densely selected image regions to use the spatial and structure information of local features. Each image region was represented as the responses of the learned exemplar classifiers and then refined with the mixture Dirichlet model for semantic representations. A bilayer model was used to predict image's classes by modeling the linear- and high-order relationships. We conducted experiments on several public data sets, and the results proven the effectiveness of the proposed method.

References

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nice, France, Oct. 2003, pp. 1470–1477.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, Apr. 2008, Art. ID 5.

[3] L. Grady and E. L. Schwartz, "Isoperimetric graph partitioning for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 469–475, Mar. 2006.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, Jun. 2005, pp. 886–893.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[7] B. Fan, F. Wu, and Z. Hu, "Aggregating gradient distributions into intensity orders: A novel local image descriptor," in *Proc. CVPR*, Jun. 2011, pp. 2377–2384.

[8] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE features," in *Proc. ECCV*, 2012, pp. 214–227.

[9] M. A. Rao, D. Vázquez, and A. M. López, "Color contribution to part-based person detection in different types of scenarios," in *Proc. 14th Int. Conf. Comput. Anal. Images Patterns*, 2011, pp. 463–470.

[10] V. D. Nguyen, D. D. Nguyen, T. T. Nguyen, V. Q. Dinh, and J. W. Jeon, "Support local pattern and its application to disparity improvement and texture classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 263–276, Feb. 2014.

[11] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *Proc. CVPR*, Jun. 2011, pp. 1673–1680.

[12] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, Jun. 2007.

[13] B. Ni, S. Yan, M. Wang, A. A. Kassim, and Q. Tian, "High-order local spatial context modeling by spatialized random forest," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 739–751, Feb. 2013.

[14] C. Zhang, J. Liu, C. Liang, Q. Huang, and Q. Tian, "Image classification using Harr-like transformation of local features with coding residuals," *Signal Process.*, vol. 93, no. 8, pp. 2111–2118, Aug. 2013.

[15] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate Web image search," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 25–32.

[16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2006, pp. 2169–2178.

[17] C. Zhang, J. Liu, Q. Tian, Y. Han, H. Lu, and S. Ma, "A boosting, sparsity-constrained bilinear model for object recognition," *IEEE MultiMedia*, vol. 19, no. 2, pp. 58–68, Feb. 2012.

[18] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei, "Object-centric spatial pooling for image classification," in *Proc. ECCV*, 2012, pp. 1–15.

[19] N. Rasiwasia and N. Vasconcelos, "Scene classification with low-dimensional semantic spaces and weak supervision," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–6.

[20] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *Int. J. Comput. Vis.*, vol. 72, no. 2, pp. 133–157, Apr. 2007.

[21] C. Zhang, J. Liu, Q. Tian, C. Liang, and Q. Huang, "Beyond visual features: A weak semantic image representation using exemplar classifiers for classification," *Neurocomputing*, vol. 120, pp. 318–324, Nov. 2013.

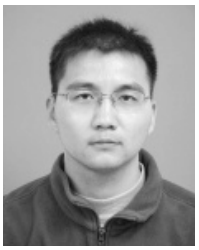
[22] C. Zhang *et al.*, "Object categorization in sub-semantic space," *Neurocomputing*, vol. 142, pp. 248–255, Oct. 2014.

[23] N. Rasiwasia and N. Vasconcelos, "Holistic context models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 902–917, May 2012.

[24] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, May 2001.

[25] N. Inoue and K. Shinoda, "A fast and accurate video semantic-indexing system using fast MAP adaptation and GMM supervectors," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1196–1205, Aug. 2012.

[26] Y. Yang, Z.-J. Zha, Y. Gao, X. Zhu, and T.-S. Chua, "Exploiting Web images for semantic video indexing via robust sample-specific loss," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1677–1689, Oct. 2014.
[27] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 958–966, Aug. 2007.
[28] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.
[29] J. Tang, Q. Chen, M. Wang, S. Yan, T.-S. Chua, and R. Jain, "Towards optimizing human labeling for interactive image tagging," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 9, no. 4, p. 29, Aug. 2013.
[30] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.
[31] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, 2010, pp. 776–789.
[32] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1778–1785.
[33] D. Parikh and K. Grauman, "Relative attributes," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 503–510.
[34] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 951–958.
[35] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-SVMs for object detection and beyond," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 89–96.
[36] J. Marin, D. Vazquez, A. M. Lopez, J. Amores, and B. Leibe, "Random forests of local experts for pedestrian detection," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2592–2599.
[37] T. Wang, S. Wang, and X. Ding, "Detecting human action as the spatio-temporal tube of maximum mutual information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 277–290, Feb. 2014.
[38] D. Vázquez, A. López, D. Ponsa, and J. Marin, "Cool world: Domain adaptation of virtual and real worlds for human detection using active learning," in *Proc. Adv. Neural Inf. Process. Syst.-Workshops*, 2011, pp. 1–6.
[39] J. Xu, S. Ramos, D. Vazquez, and A. M. Lopez, "Domain adaptation of deformable part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2367–2380, Dec. 2014.
[40] J. Xu, D. Vazquez, S. Ramos, A. M. Lopez, and D. Ponsa, "Adapting a pedestrian detector by boosting LDA exemplar classifiers," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 688–693.
[41] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPR)*, Jun. 2004, p. 178.
[42] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Dept. Vis., Pasadena, CA, USA, Tech. Rep. TR-2007-001, 2007.
[43] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2006, pp. 1447–1454.
[44] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. ICVGIP*, Dec. 2008, pp. 722–729.
[45] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1794–1801.
[46] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 712–727, Apr. 2008.
[47] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.
[48] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. CVPR*, Jun. 2008, pp. 1–8.
[49] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. CVPR*, Jun. 2010, pp. 3360–3367.
[50] X.-T. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *Proc. CVPR*, Jun. 2010, pp. 3493–3500.
[51] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. CVPR*, Jun. 2006, pp. 2126–2136.
[52] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 92–104, Jan. 2013.
[53] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.
[54] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.
[55] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.
[56] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[57] N. Xie, H. Ling, W. Hu, and X. Zhang, "Use bin-ratio information for category and scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2313–2319.
[58] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 221–228.
[59] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image classification and retrieval are ONE," in *Proc. ICMR*, 2015, pp. 3–10.
[60] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.



Chunjie Zhang received the B.E. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He was an Engineer with the Henan Electric Power Research Institute, Zhengzhou, China, from 2011 to 2012. He held a post-doctoral position with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, where he is currently an Assistant Professor with the School of Computer and Control Engineering. His current research interests include image processing, machine learning, pattern recognition, and computer vision.

Qingming Huang received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1994.

He was a Post-Doctoral Fellow with the National University of Singapore, Singapore, from 1995 to 1996, and was a member of the Research Staff with the Institute for Infocomm Research, Singapore, from 1996 to 2002. He joined the Chinese Academy of Sciences, Beijing, China, in 2003. He is currently a Professor with the University of Chinese Academy of Sciences, Beijing. His current research interests include image and video analysis, video coding, pattern recognition, and computer vision.

Qi Tian (F'16) received the B.E. degree from Tsinghua University, Beijing, China, in 1992, the M.S. degree from Drexel University, Philadelphia, PA, USA, in 1996, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2002.

He is currently a Professor with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA, and an Adjunct Professor with Zhejiang University, Hangzhou, China, and Xidian University, Xi'an, China. His current research interests include multimedia information retrieval, computational systems biology, biometrics, and computer vision.