Structured Weak Semantic Space Construction for Visual Categorization

Chunjie Zhang, Jian Cheng, and Qi Tian, *Fellow, IEEE*

***Abstract*— Visual features have been widely used for image representation and categorization. However, visual features are often inconsistent with human perception. Besides, constructing explicit semantic space is still an open problem. To alleviate these two problems, in this paper, we propose to construct structured weak semantic space for image representation. Exemplar classifier is first trained to separate each training image from other images for weak semantic space construction. However, each exemplar classifier separates one training image from other images, and it only has limited semantic separability. Besides, the outputs of exemplar classifiers are inconsistent with each other. We jointly construct the weak semantic space using structured constraint. This is achieved by imposing low-rank constraint on the outputs of exemplar classifiers with sparsity constraint. An alternative optimization procedure is used to learn the exemplar classifiers. Since the proposed method does not dependent on the initial image representation strategy, we can make use of various visual features for efficient exemplar classifier training (e.g., fisher vector-based methods and convolutional neural networks-based methods). We apply the proposed structured weak semantic space-based image representation method for categorization. The experimental results on several public image data sets prove the effectiveness of the proposed method.**

***Index Terms*— Exemplar classifier training, image classification, structure learning, visual categorization, weak semantic space.**

I. Introduction

In RECENT years, local features have been widely used for image classification [1], [2], object detection [3], segmentation [4], and retrieval [5]. Local features [3], [6]–[9] are designed to cope with deformations and can also be computed efficiently. To represent images, the bag-of-visual-word scheme is often used.

Manuscript received April 21, 2016; revised April 25, 2017; accepted July 13, 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 Scientific Research Key Program of Beijing Municipal Commission of Education under Grant KZ201610005012, and in part by the National Science Foundation of China under Grant 61429201. The work of Q. Tian was supported in part by ARO under Grant W911NF-15-1-0290 and in part by the Faculty Research Gift Awards by NEC Laboratories of America and Blippar. *(Corresponding author: Chunjie Zhang.)*

C. Zhang is with the Research Center for Brain-Inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: chunjie.zhang@ia.ac.cn).

J. Cheng is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jcheng@nlpr.ia.ac.cn).

Q. Tian is with the Department of Computer Sciences, The University of Texas at San Antonio, San Antonio, TX 78249-1604 USA (e-mail: qitian@cs.utsa.edu).

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/TNNLS.2017.2728060

Although effective, visual feature-based representations lack semantic correspondences with human perception due to the semantic gap [10]. To alleviate this problem, many works [11]–[18] have been done. Some researchers [11]–[14] try to construct semantic space directly using training samples. However, the discriminative power of the corresponding semantic spaces would be degenerated if the training samples are biased. Others [15]–[18] try to annotate images at first. However, the contamination of irrelevant objects hinders its performances. Besides, the training samples are often imbalanced, which restricts the annotation accuracy.

Instead of only using training images collected by human experts, researchers also try to transfer useful information [19]–[24] for image representation and classification. The Internet has abundant information, which has been explored both for annotation [19], [20] and classification [21]. However, the Internet often contains noisy information, which cannot be used directly for classification. To make use of other data sets, researchers propose various transfer learning-based methods [22]–[24]. However, the performances are not satisfactory for images with large interclass variations. To alleviate these problems, attributes [25]–[29] and exemplar classifier-based [30]–[33] methods become popular. However, attributes are often predefined and of limited discriminative power. There are also many images, which cannot be well represented by attributes. Besides, exemplar classifiers are often trained independently, which makes the resulting representation inconsistent with each other.

To solve these problems, in this paper, we propose a novel structured weak semantic space construction method for visual representation. We first train exemplar classifiers to separate each training image from other images of different classes. The output of the learned exemplar classifier has weak semantic meanings as it measures the semantic similarity between the testing image and the exemplar image. Besides, to make the exemplar classifiers consistent with each other, we impose low-rank constraint to the outputs of exemplar classifiers. Sparsity constraint is also used to train exemplar classifiers for final image representation. To evaluate the effectiveness of the proposed method, we apply it for the categorization tasks on several public image data sets. Experimental results prove that the proposed method can achieve superior performances than many visually based methods.

The main contributions of this paper lie in three aspects.

1) First, we propose a novel structured weak semantic space-based image representation method. Images are semantically represented for better categorization.

2) Second, we generate the structured weak semantic space by alternative optimization, which can be computed efficiently.
3) Third, the proposed method can be combined with more discriminative representation strategies [e.g., convolutional neural network (CNN)-based methods] to further improve categorization performances.

The rest of this paper is organized as follows. Related work is given in Section II. The details of the proposed structured weak semantic space-based representation method are given in Section III. To evaluate the effectiveness of the proposed method, we apply it for classification tasks on several public data sets in Section IV. Finally, we conclude in Section V.

II. Related Work

Various visual features [3], [6]–[9] had been proposed in recent years. Scale-invariant feature transform (SIFT) feature was proposed by Lowe [6] and widely used for categorization. To speed up computation, Dalal and Triggs [3] proposed the histograms of oriented gradients feature for human detection. Nguyen *et al.* [7] proposed the support local pattern while Fan *et al.* [8] targeted object matching with aggregated gradient distributions. The Speeded up robust features was proposed by Bay *et al.* [9]. However, the visual features lacked explicit semantic meanings due to the semantic gap [10]. To solve this problem, many works [11]–[18] had been made. Rasiwasia and Vasconcelos [11] tried to classify scene images with low-dimensional semantic spaces while Vogel and Schiele [12] semantically modeled natural scenes for retrieval. Zhang *et al.* [13] proposed to categorize objects in subsemantic spaces while Rasiwasia and Vasconcelos [14] used holistic context models.

Researchers also tried to annotate images [15]–[18]. Duygulu *et al.* [15] viewed object recognition as translating words from a fixed vocabulary by visual similarities. Jeon *et al.* [16] proposed a cross-media relevance model for automatic image annotation. The use of metric learning technique was proposed by Guillaumin *et al.* [17] with nearest neighbor models for annotation. Wang *et al.* [18] simultaneously conducted image classification and annotation. To harvest the information of other sources, many works had been made. Wang *et al.* [19] tried to annotate images by searching and mining technologies [20]. Hierarchical synthetic image classification was conducted using image search and generic features by Wang and Kan [21]. Zhang *et al.* [22] tried to used prelearned codebooks for new application with linear [24] and nonlinear codebook transfer. Zhu *et al.* [23] classified images with heterogeneous transfer learning.

Attribute was also used to alleviate the semantic gap problem. Farhadi *et al.* [25] tried to describe objects by their attributes while Parikh and Grauman [26] compared the relativeness of different attributes. Lampert *et al.* [27] proposed to detect unseen object classes by between-class attribute transfer. Patterson *et al.* [28] collected various attributes for scene understanding while Lampert *et al.* [29] targeted zero-shot categorization by attribute-based classification. Exemplar classifier-based method was also widely used. Malisiewicz *et al.* [30] combined exemplar support vector machine (SVM) classifiers for object detection while Zhang *et al.* [31] proposed to use exemplar classifiers for weak semantic representation. Zepeda and Perez [32] viewed exemplar SVMs as visual feature encoders while Modolo *et al.* [33] calibrated ensemble of exemplar classifiers for detection.

Sparse coding [34] had been widely used for visual categorization with good performance. Wang *et al.* [35] combined locality constraint with sparse coding. Gao *et al.* [36] encoded visually similar local features with similar parameters to reduce the quantization error. In order to encode high-order information, fisher vector was used [37], [38]. Cinbis *et al.* [37] categorized images by combining fisher kernels with nonindependent identically distributed image models. Sánchez *et al.* [38] systematically evaluated fisher vector for image classification. The direct usage of local features was also proposed by Boiman *et al.* [39]. The CNN-based methods [40]–[44] went one step further by working on image pixels directly. Donahue *et al.* [40] proposed a deep convolutional activation feature (DeCAF) for generic visual recognition while Liang and Hu [41] used recurrent CNN. Lin *et al.* [42] combined localization, alignment, and classification with the deep learning scheme. Perronnin and Larlus [43] proposed a hybrid classification architecture by combing fisher vectors with neural networks. The feature-sign-search strategy [45] was often used for sparse coding.

Researchers had also proposed many data sets [46]–[49] and algorithms [50]–[68] for efficient categorization. Van de Sande *et al.* [50] extended the SIFT feature with different color channels while Gemert *et al.* [51] explored the hard assignment problem of local feature quantization. Torresani *et al.* [52] proposed to semantically represent objects. Li *et al.* [53] harvested information from the Google to construct the ObjectBank for image representation. Zhang *et al.* [54] used low-rank sparse coding to jointly encode local features while Xie *et al.* [55] jointly considered image classification and retrieval. Yuan and Yan [56] used sparse reconstruction jointly for visual classification while Angelova and Zhu [57] combined object detection and segmentation for classification. Chai *et al.* [58] also combined segmentation with classification. Ito and Kubota [59] explored heterogeneous co-occurrence features while Bosch *et al.* [60] used hybrid approach with generative and discriminative models. Chatfield *et al.* [61], [63] tried to implement different algorithms. Yang *et al.* [62] used multiple kernel learning for object categorization. Zhang *et al.* [64] tried to propagate discriminative information with graphlet path while Li *et al.* [65] used locality constraint with codebook learning for classification. Xiong *et al.* [66] explored the conformal transformation kernel while Li *et al.* [67] tried to learn ordinal distance metrics for image ranking. Li *et al.* [68] also proposed a novel descriptor for efficient image classification.

In recent years, the CNN-based methods [69]–[75] had been widely used for image classification. Krizhevsky *et al.* [69] classified the ImageNet with CNNs and improved the accuracy dramatically. Simonyan and Zisserman [70] increased the depth of the network with improved performances.

Perronnin and Larlus [43] combined the fisher vector and CNNs while Zeiler and Ferugs [71] tried to visualize the learned networks. Donahue *et al.* [72] introduced DeCAF for classification while Wei *et al.* [73] targeted the multilabel classification problem. He *et al.* [74] proposed a spatial pyramid pooling strategy for classification with the convolutional networks while Azizpour *et al.* [75] made classification by exploring the deep representations from generic to specific characters.

III. Visual Categorization by Structured Weak Semantic Space Construction

In this section, we give the details of the proposed structured weak semantic space method for visual categorization.

A. Structured Weak Semantic Space Construction

We first represent images using the sparse coding technique [34]. Formally, let $\boldsymbol{x_n} \in \mathbb{R}^{D\times 1}, n = 1, \ldots, N$, be the visual representation of the nth image with its corresponding label as y_n, where N is the number of training images. We try to train linear exemplar classifier

$$\hat{y_n} = \boldsymbol{w_n x_n} \quad n = 1, \ldots, N \tag{1}$$

to separate each training image from other images of different classes, where $\boldsymbol{w_n} \in \mathbb{R}^{1\times D}$ is the classifier parameter. This is achieved by minimizing the loss of training images as

$$\boldsymbol{w_n} = \underset{\boldsymbol{w_n}}{\operatorname{argmin}} \sum_{i=1, y_i \neq y_n}^{N} \ell(\boldsymbol{w_n x_n}, y_i) \quad n = 1, \ldots, N. \tag{2}$$

We use exponential loss $\ell(\boldsymbol{w_n x_n}, y_j) = e^{-\boldsymbol{w_n x_n} \times y_j}$ in this paper. To avoid overfitting, we add sparsity constraint on the parameter of exemplar classifier as

$$\boldsymbol{w_n} = \underset{\boldsymbol{w_n}}{\operatorname{argmin}} \sum_{i=1, y_i \neq y_n}^{N} \ell(\hat{y}_n, y_i) + \alpha \|\boldsymbol{w_n}\|_1 \tag{3}$$

where α is the parameter that balances the influences of sparsity constraint and the summed loss.

The nth exemplar classifier is trained to separate image x_n from other images of different classes. Suppose we have two other images x_i and x_j. x_i is similar with x_n while x_j is dissimilar. If we classify x_i and x_j using the nth exemplar classifier, the predicted $\hat{y_i}$ would be larger than $\hat{y_j}$. In other words, the exemplar classifier believes image x_i is more semantically similar with x_n than image x_j. However, the exemplar classifier is trained with only one positive sample. The output is relatively weak as it only measures the similarity with one image instead of images of the same class. Although the semantic information is weak for each exemplar classifier, the joint representation using a number of exemplar classifiers would be more discriminative. We use the outputs of exemplar classifiers for image representation as $\boldsymbol{h} = [\hat{y_1}; \ldots; \hat{y_N}]$ and call it weak semantic representation in this paper. If two images are similar, their weak semantic representations would be similar and consistent with each other. In this way, we are able to replace image's visual representation with weak semantic representation, which is more semantically correlated than visual features.

Let $\hat{\boldsymbol{Y}} = [\boldsymbol{h^1}, \ldots, \boldsymbol{h^N}]$, $\boldsymbol{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}]$, and $\boldsymbol{W} = [\boldsymbol{w_1}; \ldots; \boldsymbol{w_N}]$, and we can get the matrix representation of images as

$$\hat{\boldsymbol{Y}} = \boldsymbol{W X}. \tag{4}$$

The exemplar classifiers can be optimized as

$$\boldsymbol{W} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \|\boldsymbol{e^{-W\tilde{X}}}\|_{1,1} + \alpha \|\boldsymbol{W}\|_{1,1} \tag{5}$$

with $\widetilde{\boldsymbol{X}} = \boldsymbol{X.} \times \boldsymbol{Y}$, where $.\times$ is the dot product. The exemplar classifier can be trained one after another. However, due to the interclass and intraclass variances, each independently trained exemplar classifier only tries to separate the corresponding training image from other images, leaving the other training images of the same class unconsidered. Besides, the outputs of different exemplar classifiers have varied degrees of scales. The resulting representations may be dominated by the exemplar classifiers with large scales. It is more reasonable to jointly learn exemplar classifiers.

We impose low-rank constraint on the outputs $\hat{\boldsymbol{Y}}$ with the problem as

$$\boldsymbol{W} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \|e^{-\boldsymbol{W}\widetilde{\boldsymbol{X}}}\|_{1,1} + \alpha \|\boldsymbol{W}\|_{1,1} + \beta \|\boldsymbol{WX}\|_* \tag{6}$$

where β is the parameter, which controls the influence of low-rank constraint. We use the low-rank constraint for two reasons. First, the consistency of exemplar classifiers means they have similar predicted values for similar images. Second, the low-rank constraint can model the intrinsic correlations and suits our problem well. In this way, we can jointly learn the exemplar classifiers for structured weak semantic space construction.

The proposed method can alleviate the semantic discrepancy between visual features and human perception for two reasons. First, we use the outputs of exemplar classifiers for image representation instead of using visual features. Each exemplar classifier tries to predict the semantic relatedness with the exemplar image. Since exemplar classifier is trained with one positive sample, the corresponding semantic representation is weak. Second, we jointly model the structured information of exemplar classifiers with low-rank and sparse constraint. The structured modeling strategy ensures exemplar classifiers to consistently representing images with the same semantic meaning. In this way, we are able to get semantically consistent image representations and alleviate the visual and semantic discrepancy to some extent.

Note that this strategy can be combined with other visual feature-based representation methods (e.g., sparse coding, fisher vector, and CNN). This is because the proposed method first trains exemplar classifiers for semantic representations. The exemplar classifier training process is independent of the initial image representation. After exemplar classifiers are learned, we can construct weak semantic spaces for representation. Since the neural network-based methods can achieve more accurate classification performances, the learned semantic representations are more discriminative and semantically

consistent with human perception. Finally, we train linear SVM classifiers to predict image categories. The procedures of the proposed structured weak sematic space construction method for visual categorization are given in Algorithm 1.

Algorithm 1 Procedures of the Proposed Structured Weak Semantic Representation Method for Visual Categorization

Input:

Training images $(\boldsymbol{x_n},\ y_n)$, n= 1, . . . ,N and the testing images.

Output:

The predicted classes of test images;

1: Represent images using the sparse coding strategy with local features;
2: Construct the weak semantic space by optimizing over Problem 6 and represent images accordingly;
3: Train SVM classifiers using the weak semantic spaces based representation for categorization;
4: **return** The predicted classes of testing images.

B. Optimization

Problem 6 is hard to optimize as both the nuclear norm and the sparsity constraint are nonsmooth. To handle this problem, we introduce two slack variables with equality constraint as

$$\begin{aligned}[\boldsymbol{W}, \boldsymbol{W_1}, \boldsymbol{W_2}] = \underset{\boldsymbol{W},\boldsymbol{W_1},\boldsymbol{W_2}}{\operatorname{argmin}} & \|e^{-\boldsymbol{W_2}\widetilde{\boldsymbol{X}}}\|_{1,1} \\ & + \alpha\|\boldsymbol{W_1}\|_{1,1} + \beta\|\boldsymbol{WX}\|_* \\ \text{s.t. } & \boldsymbol{W} = \boldsymbol{W_1},\ \boldsymbol{W_1} = \boldsymbol{W_2}. \end{aligned} \tag{7}$$

This problem can be solved by introducing augmented lagrange multipliers as

$$\begin{aligned} L(\boldsymbol{W}, \boldsymbol{W_1}, \boldsymbol{W_2}) = & \|e^{-\boldsymbol{W_2}\widetilde{\boldsymbol{X}}}\|_{1,1} + \alpha\|\boldsymbol{W_1}\|_{1,1} + \beta\|\boldsymbol{WX}\|_* \\ & + \operatorname{tr}\big[\boldsymbol{Y_1^T}(\boldsymbol{W} - \boldsymbol{W_1})\big] + \frac{u_1}{2}\|\boldsymbol{W} - \boldsymbol{W_1}\|_F^2 \\ & + \operatorname{tr}\big[\boldsymbol{Y_2^T}(\boldsymbol{W_1} - \boldsymbol{W_2})\big] + \frac{u_2}{2}\|\boldsymbol{W_1} - \boldsymbol{W_2}\|_F^2 \end{aligned} \tag{8}$$

using the Inexact Augmented Lagrange Multiplier method [44]. Problem 8 can be solved by alternatively optimizing over $\boldsymbol{W}, \boldsymbol{W_1}, \boldsymbol{W_2}$ while keeping the other two parameters fixed.

1) Optimizing $\boldsymbol{W_1}$: When $\boldsymbol{W}$ and $\boldsymbol{W_2}$ are fixed, the optimization over $\boldsymbol{W_1}$ can be achieved by

$$\begin{aligned} \boldsymbol{W_1} = \underset{\boldsymbol{W_1}}{\operatorname{argmin}} & \alpha\|\boldsymbol{W_1}\|_{1,1} + \operatorname{tr}\big[\boldsymbol{Y_1^T}(\boldsymbol{W} - \boldsymbol{W_1})\big] \\ & + \frac{u_1}{2}\|\boldsymbol{W} - \boldsymbol{W_1}\|_F^2 + \operatorname{tr}\big[\boldsymbol{Y_2^T}(\boldsymbol{W_1} - \boldsymbol{W_2})\big] \\ & + \frac{u_2}{2}\|\boldsymbol{W_1} - \boldsymbol{W_2}\|_F^2. \end{aligned} \tag{9}$$

Let $L_1 = \operatorname{tr}[\boldsymbol{Y_1^T}(\boldsymbol{W} - \boldsymbol{W_1})] + (u_1/2)\|\boldsymbol{W} - \boldsymbol{W_1}\|_F^2 + \operatorname{tr}[\boldsymbol{Y_2^T}(\boldsymbol{W_1} - \boldsymbol{W_2})] + (u_2/2)\|\boldsymbol{W_1} - \boldsymbol{W_2}\|_F^2$, and Problem 9 can be solved using the feature-sign-search [45] strategy with

$$\frac{\partial L_1}{\partial \boldsymbol{W_1}} = -\boldsymbol{Y_1} + u_1(\boldsymbol{W_1} - \boldsymbol{W}) + \boldsymbol{Y_2} + u_2(\boldsymbol{W_1} - \boldsymbol{W_2}). \tag{10}$$

2) Optimizing $\boldsymbol{W_2}$: When $\boldsymbol{W}$ and $\boldsymbol{W_1}$ are fixed, the optimization over $\boldsymbol{W_2}$ can be done as

$$\begin{aligned} \boldsymbol{W_2} = \operatorname{argmin}_{\boldsymbol{W_2}} & \|e^{-\boldsymbol{W_2}\widetilde{\boldsymbol{X}}}\|_{1,1} + \operatorname{tr}\big[\boldsymbol{Y_2^T}(\boldsymbol{W_1} - \boldsymbol{W_2})\big] \\ & + \frac{u_2}{2}\|\boldsymbol{W_1} - \boldsymbol{W_2}\|_F^2. \end{aligned} \tag{11}$$

This problem can be solved with gradient descent. Let $L_2 = \|e^{-\boldsymbol{W_2}\widetilde{\boldsymbol{X}}}\|_{1,1} + \operatorname{tr}[\boldsymbol{Y_2^T}(\boldsymbol{W_1} - \boldsymbol{W_2})] + \frac{u_2}{2}\|\boldsymbol{W_1} - \boldsymbol{W_2}\|_F^2$; we can have

$$\frac{\partial L_2}{\partial \boldsymbol{W_2}} = -\|e^{-\boldsymbol{W_2}\widetilde{\boldsymbol{X}}}\|_{1,1}\boldsymbol{W_2}e^{-\boldsymbol{W_2}\widetilde{\boldsymbol{X}}} - \boldsymbol{Y_2} + u_2(\boldsymbol{W_2} - \boldsymbol{W_1}). \tag{12}$$

3) Optimizing $\boldsymbol{W}$: When $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ are fixed, the optimization over $\boldsymbol{W}$ can be equally rewritten as

$$\boldsymbol{W} = \operatorname{argmin}_{\boldsymbol{W}} \frac{\beta}{u_1}\|\boldsymbol{WX}\|_* + \frac{1}{2}\left\|\boldsymbol{W} - \left(\boldsymbol{W_1} - \frac{\boldsymbol{1}}{\boldsymbol{u_1}}\boldsymbol{Y_1}\right)\right\|_F^2. \tag{13}$$

Let $\boldsymbol{C} = \boldsymbol{WX}$; we have $\boldsymbol{W} = \boldsymbol{CX^+}$, where $\boldsymbol{X}^+$ is the pseudoinverse of matrix $\boldsymbol{X}$. Since $\boldsymbol{X}$ is known, the optimization over Problem 13 can be rewritten as

$$\boldsymbol{C} = \operatorname{argmin}_{\boldsymbol{C}} \frac{\beta}{u_1}\|\boldsymbol{C}\|_* + \frac{1}{2}\left\|\boldsymbol{CX^+} - \left(\boldsymbol{W_1} - \frac{\boldsymbol{1}}{\boldsymbol{u_1}}\boldsymbol{Y_1}\right)\right\|_F^2. \tag{14}$$

Let $(\boldsymbol{W_1} - \frac{\boldsymbol{1}}{\boldsymbol{u_1}}\boldsymbol{Y_1})\boldsymbol{X} = \boldsymbol{D}$; we can have

$$\boldsymbol{C} = \operatorname{argmin}_{\boldsymbol{C}} \frac{\beta}{u_1}\|\boldsymbol{C}\|_* + \frac{1}{2}\|(\boldsymbol{C} - \boldsymbol{D})\boldsymbol{X^+}\|_F^2. \tag{15}$$

Since $\|(\boldsymbol{C} - \boldsymbol{D})\boldsymbol{X^+}\|_F^2$ is always no larger than $\|\boldsymbol{X^+}\|_F^2\|(\boldsymbol{C} - \boldsymbol{D})\|_F^2$, we can optimize over an upper bound of the objective of Problem 15 as

$$\begin{aligned} \boldsymbol{C} &= \operatorname{argmin}_{\boldsymbol{C}} \frac{\beta}{u_1}\|\boldsymbol{C}\|_* + \frac{1}{2}\|\boldsymbol{X^+}\|_F^2\|(\boldsymbol{C} - \boldsymbol{D})\|_F^2 \\ &\geq \frac{\beta}{u_1}\|\boldsymbol{C}\|_* + \frac{1}{2}\|(\boldsymbol{C} - \boldsymbol{D})\boldsymbol{X^+}\|_F^2 \end{aligned} \tag{16}$$

which can be easily solved as

$$\boldsymbol{C} = \mathcal{T}_{\frac{\beta}{u_1\|\boldsymbol{X^+}\|_F^2}}(\boldsymbol{D}) \tag{17}$$

with $\mathcal{T}_{(1/u_1)}(\boldsymbol{D}) = \boldsymbol{U_D}\mathcal{S}_\lambda(\sum_{\mathbf{D}})\boldsymbol{V_D^T}$ is the soft-threshold singular value operator and $\boldsymbol{U_D}\sum_{\mathbf{D}}\boldsymbol{V_D^T}$ is the singular value decomposition of $\boldsymbol{D}$.

4) Updating Parameters: u_1, u_2, $\boldsymbol{Y_1}$, and $\boldsymbol{Y_2}$ can be updated accordingly as

$$\begin{aligned} \boldsymbol{Y_1} &= \boldsymbol{Y_1} + u_1(\boldsymbol{W} - \boldsymbol{W_1}), \quad u_1 = \epsilon u_1 \\ \boldsymbol{Y_2} &= \boldsymbol{Y_2} + u_2(\boldsymbol{W_1} - \boldsymbol{W_2}), \quad u_2 = \epsilon u_2,\ \epsilon > 1. \end{aligned} \tag{18}$$

We alternatively optimizing over $\boldsymbol{W}, \boldsymbol{W_1}, \boldsymbol{W_2}$ and update the parameters $\boldsymbol{Y_1}, \boldsymbol{Y_2}, u_1, u_2$ for a number of times. In this way, we can jointly learn the exemplar classifiers with structure and sparsity constraint to get the semantic representations of images. Compared with independently training exemplar classifiers, we can speed up the process for about five times. We give the procedures for solving Problem 8 in Algorithm 2. We then train linear SVM classifiers to predict the categories of images.

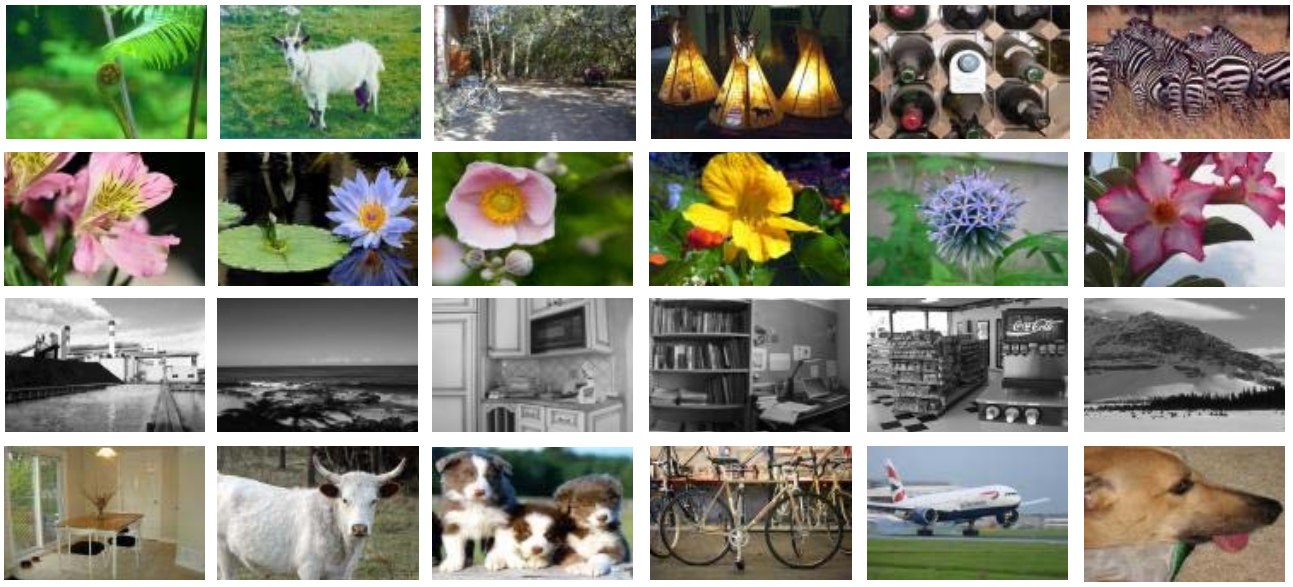

Fig. 1. Example images of the four data sets (one data set for each row, from top to bottom: Caltech-256 data set, Flower-102 data set, Scene-15 data set, and PASCAL VOC 07 data set).

Algorithm 2 Procedures for Solving Problem 8

Input:

The initial parameters $\boldsymbol{W}$, $\boldsymbol{W_1}$, $\boldsymbol{W_2}$, $\boldsymbol{Y_1}$, $\boldsymbol{Y_2}$, α, β, u_1, u_2, ϵ; the regularization parameters α and β; training images ($\boldsymbol{x_n}$, y_n), n= $1, \ldots,$N; maximum iteration number M.

Output:

The learned exemplar classifiers $\boldsymbol{W}$;

1: **for** $m = 1, 2, \ldots, M$
2: Solve for the optimal $\boldsymbol{W_1}$ while keeping $\boldsymbol{W}$, $\boldsymbol{W_2}$ fixed by solving Problem 9;
3: Solve for the optimal $\boldsymbol{W_2}$ while keeping $\boldsymbol{W}$, $\boldsymbol{W_1}$ fixed by solving Problem 11;
4: Solve for the optimal $\boldsymbol{W}$ while keeping $\boldsymbol{W_1}$, $\boldsymbol{W_2}$ fixed by solving Problem 16 as the upper bound of Problem 15;
5: Update the parameters $\boldsymbol{Y_1}$, $\boldsymbol{Y_2}$, u_1, u_2 using Eq. 18;
6: **end for**.
7: **return** The learned exemplar classifiers $\boldsymbol{W}$.

IV. Experiments

To evaluate the effectiveness of the proposed structured weak semantic space-based visual representation method (SWSS), we conduct categorization tasks on several public image data sets: the Caltech-256 data set [46], the Flower-102 data set [47], the Scene-15 data set [48], and the PASCAL VOC 07 data set [49]. Fig. 1 gives some example images of the four data sets.

A. Experimental Setup

We densely extract SIFT features with multiscales and overlap. The minimum size of local features is set to 16×16 pixels. We normalize the extracted local features with L_2 norm. The local features are encoded with sparse coding and pooled with maximum values for initial image representation [34]. For the Flower-102 data set, multiple color SIFT features [50] (red, green, blue-SIFT, hue, saturation, value-SIFT, C-invariant SIFT, and the opponent SIFT) are extracted. The codebook size is set to 1000. Spatial pyramid matching with three pyramids is used ($2^s \times 2^s$, $s = 0, 1, 2$), for the Caltech-256 data set, the Flower-102 data set, and the MIT-Indoor data set. The mean of per-class classification rate is used for performance evaluation. For the PASCAL VOC 07 data set, we use the train/validation/test images provided by [49] and leverage the average precision for evaluation. We also combine the proposed method with CNN-based methods [70], [72] (SWSS-VGG and SWSS-DeCAF) by using the image representations of [70] and [72] for exemplar classifier training. We compare with the performances reported by other researchers directly for fair comparisons.

B. Caltech-256 Data Set

There are 29 780 images of 256 classes in the Caltech-256 data set. Each class has at least 80 images. For fair comparison, we follow the experimental setup as [46] did and randomly select 15/30/45/60 training images per class for training. The random selection process is repeated for ten times.

Table I gives the performance comparisons of SWSS with other methods [13], [31], [34]–[36], [38]–[39], [46], [51]–[54]. Since SWSS can also be combined with other visual representations, we also give the performances of SWSS using the Laplacian sparse coding and Fisher vector, respectively (SWSS-LSc and SWSS-FV).

SWSS is able to outperform many visually based methods [34]–[36]. Specially, SWSS and SWSS-LSc can improve over sparse coding, spatial pyramid matching ScSPM [34] and Laplacian sparse coding, spatial pyramid matching LScSPM [36] dramatically. Besides, compared with other exemplar-based methods, SWSS jointly learns exemplar

TABLE I
PERFORMANCE COMPARISONS ON THE CALTECH-256 DATA SET

Methods	15 images	30 images	45 images	60 images
S^3R[13]	37.85 ± 0.48	43.52 ± 0.44	46.86 ± 0.63	–
WSR-EC[31]	35.28 ± 0.65	42.01 ± 0.47	45.82 ± 0.54	–
KSPM [34]	23.34 ± 0.42	29.51 ± 0.52	–	–
ScSPM [34]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
LLC [35]	27.74 ± 0.32	32.07 ± 0.24	35.09 ± 0.44	37.79 ± 0.42
LScSPM [36]	30.00 ± 0.14	35.74 ± 0.10	38.47 ± 0.51	40.32 ± 0.32
FV [38]	38.50 ± 0.20	47.40 ± 0.10	52.10 ± 0.40	54.80 ± 0.40
NBNN(1 Desc)[39]	30.45	38.18	–	–
KSPM [46]	–	34.10	–	–
KC [51]	–	27.17 ± 0.46	–	–
Classemes [52]	– 36.00	–	–	–
ObjectBank [53]	– 39.00	–	–	–
LRSC[54]	–	41.04 ± 0.23	–	–
FV+L^2EMG[68]	45.00 ± 0.20	53.60 ± 0.30	58.20 ± 0.30	61.80 ± 0.40
VUCN [71]	65.70 ± 0.20	70.60 ± 0.20	72.70 ± 0.40	74.20 ± 0.30
SWSS	39.53 ± 0.42	46.28 ± 0.40	49.19 ± 0.44	52.61 ± 0.38
SWSS-LSc	41.28 ± 0.55	47.53 ± 0.48	50.13 ± 0.50	52.70 ± 0.42
SWSS-FV	42.46 ± 0.38	49.85 ± 0.42	54.66 ± 0.47	56.52 ± 0.41
SWSS-DeCAF	61.52 ± 0.39	67.68 ± 0.65	69.77 ± 0.53	72.83 ± 0.44
SWSS-VGG	69.37 ± 0.46	73.56 ± 0.51	74.83 ± 0.39	76.25 ± 0.47

TABLE II
PERFORMANCE COMPARISONS ON THE FLOWER-102 DATA SET

Methods	Classification rate
Nilsback and Zisserman [47]	72.8
Xie [55]	86.8
KMTJSRC-CG [56]	74.1
Det+Seg [57]	80.7
TriCoS [58]	85.2
CoHoG [59]	74.8
WSR-EC[31]	82.5
S^3R[13]	85.3
Deep Optimized[75]	91.3
SWSS	87.8
SWSS-LSc	89.1
SWSS-FV	91.4
SWSS-DeCAF	94.5
SWSS-VGG	96.7

classifiers with consistency. In this way, we are able to improve the categorization accuracy over [13] and [31]. Moreover, SWSS can also outperform ObjectBank [53]. The automatically collected Internet images are often contaminated with noise, which hinders the final accuracy.

The fisher vector [38] can encode more information compared with sparse coding. The proposed SWSS can also be combined with fisher vector by using fisher vector for initial image representations. In this way, we are able to improve over FV on the Caltech-256 data set. Besides, by combining SWSS with DeCAF and VGG, we can further improve the performance. These experimental results prove the usefulness of the proposed SWSS method for categorization.

C. Flower-102 Data Set

The Flower-102 data set has 8189 images of 102 classes with 40–250 images per class. Some of the flower images are visually similar, which are hard to classify even for humans. We follow the same experimental setup as [47] did by using the data splits for fair comparison. Table II gives the performance comparisons of SWSS with [13], [31], [47], and [55]–[59] on the Flower-102 data set.

We can have similar conclusions as on the Caltech-256 data set. First, compared with visually based methods [47], [55]–[59], modeling the semantic information can help to improve the discriminative power. Second, detection and segmentation of objects can help to improve the performance. However, the detection and segmentation strategies are not always satisfactory, especially when objects and background are cluttered. Detecting and segmenting objects [57], [58] also cost a lot of computational time. The proposed SWSS method can improve over [57] and [58] and also does not need to detect and segment objects. Third, by jointly learning exemplar classifiers, we are able to improve over weak semantic representation using examplar classifier [31] and S^3R [13]. By encoding local features using fisher vector strategy, we can get finer initial representations of images over sparse coding. Hence, the combination of SWSS with fisher vector can improve over simple SWSS by about 3.6%. The performance can be further improved by more than 5% when combining SWSS with VGG. The results again show the usefulness of the proposed method.

D. Scene-15 Data Set

This data set has 200 to 400 images with the 15 classes as: *store, office, tallbuilding, street, opencountry, mountain, insidecity, highway, forest, coast, livingroom, kitchen, industrial, suburb, and bedroom*. We randomly select 100 images per class for performance evaluation as [48] did and use the rest images for evaluation. This process is repeated for ten times.

Table III gives the performance comparison on the Scene-15 data set. We compare with visually based methods [34], [38], [51] and semantic-based methods [11], [13], [14], [31], [53], [60]. SWSS is able to outperform both visual and semantic-based methods. Since visual features

TABLE III
PERFORMANCE COMPARISONS ON THE SCENE-15 DATA SET

Algorithms	Classification Rate
pLSA[60]	72.7
LDA[11]	59.0
ScSPM[34]	80.28 ± 0.93
SPM[48]	81.40 ± 0.50
KC [51]	76.67 ± 0.39
Semantic Space[11]	73.95 ± 0.74
Contextual Models[14]	77.20 ± 0.39
ObjectBank [53]	80.9
WSR-EC[31]	81.54 ± 0.59
S^3R[13]	83.72 ± 0.78
LScSPM [36]	89.78 ± 0.40
SWSS	86.49 ± 0.63
SWSS-LSc	90.51 ± 0.45
SWSS-FV	91.83 ± 0.57
SWSS-DeCAF	94.26 ± 0.43
SWSS-VGG	96.72 ± 0.39

TABLE IV
MAP COMPARISONS ON THE PASCAL VOC 07 DATA SET

Algorithms	mAP
LLC [35]	59.3
FV [38]	61.7
Best'07[49]	59.4
NScSPM[2]	61.5
SV[61]	58.2
GS-MKL[62]	62.2
Chatfield[63]	82.4
DeCAF[72]	73.4
VGG[70]	89.3
HCP-VGG[73]	90.9
He[74]	82.4
SWSS	60.2
SWSS-FV	64.1
SWSS-DeCAF	75.3
SWSS-VGG	90.5

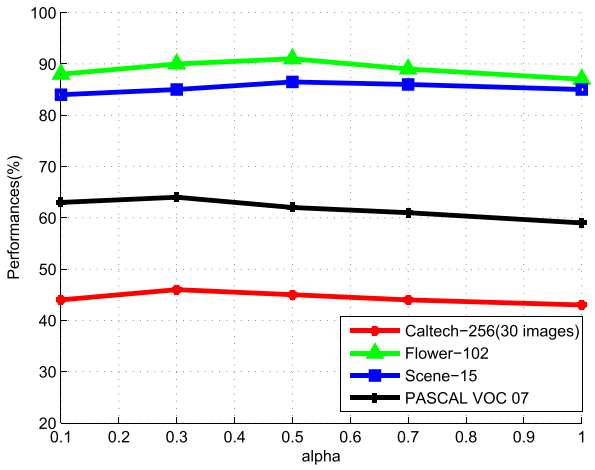

Fig. 2. Performance changes with alpha on the four data sets (SWSS for the Caltech-256 data set with 30 images per class and the Scene-15 data set, and SWSS-FV for the Flower-102 data set and the PASCAL VOC 07 data set).

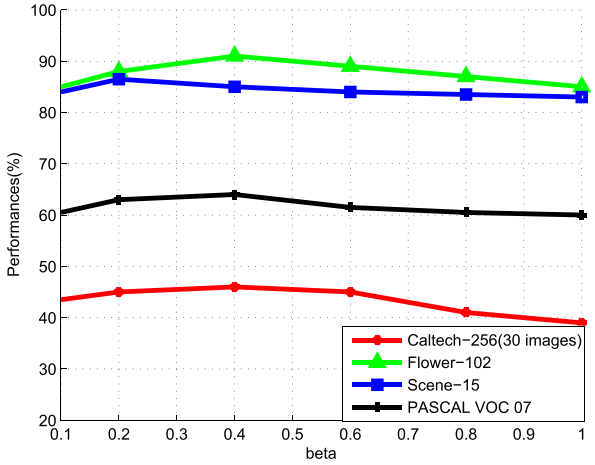

Fig. 3. Performance changes with beta on the four data sets (SWSS for the Caltech-256 data set with 30 images per class and the Scene-15 data set, and SWSS-FV for the Flower-102 data set and the PASCAL VOC 07 data set).

are quite different from words, directly using the word processing techniques [11], [60] cannot cope with image categorization tasks well. It is more reasonable and efficient to model the relationships with visual features. By encoding local features more finely (SWSS-LSc/SWSS-FV), we can improve the categorization accuracy. The performance can be further improved with more discriminative representations (SWSS-DeCAF/SWSS-VGG). The experimental results on the Scene-15 data set again prove the usefulness of the proposed method.

E. PASCAL VOC 07 Data Set

This data set has 20 classes (*person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, soft* and *tv/monitor*) of about 10 000 images. We use the train/validation/test data set provided by [49]. The validation samples are used for parameter selection. After the parameters are learned, we combine the train and validation samples together for testing.

Table IV gives the mean average precision (mAP) comparisons of different methods on the PASCAL VOC 07 data set. We also combine SWSS with fisher vector and CNN-based strategies (SWSS-FV/SWSS-DeCAF/ SWSS-VGG/SWSS-Hypotheses-CNN-Pooling) for categorization. SWSS-FV outperforms FV by 2.4%. Besides, the CNN-based strategy is able to outperform well-designed features by exploring deep correlations among image pixels. By combing CNN for initial image representation, we can further improve the categorization accuracy by jointly modeling the weak semantic correspondences among training samples. Specially, SWSS-DeCAF improves over DeCAF by about 2%. When combined with VGG, the mAP can be further improved by 15%.

We also give the per-class average precision of different methods in Table V. SWSS is able to improve the categorization performances over other methods when combined with fisher vector and CNN-based visual representation strategies. Besides, the relative improvements of different classes vary for rigid and nonrigid objects. Since nonrigid objects often have large interclass variations, it is hard to model them properly with visual features. The use of semantic representation would

TABLE V
PER-CLASS PERFORMANCE COMPARISONS ON THE PASCAL VOC 07 DATA SET

class	LLC[35]	Best07[49]	FV [38]	SV[61]	NScSPM[2]	GS-MKL[62]	DeCAF[72]	Chatfield[63]	SWSS-FV	SWSS-DeCAF	SWSS-VGG
airplane	74.8	77.5	80.0	74.3	80.2	79.4	87.4	95.3	82.7	88.7	98.2
bicycle	65.2	63.6	67.4	63.8	67.1	62.4	79.3	90.4	70.1	81.3	95.1
bird	50.7	56.1	51.9	47.0	52.7	58.5	84.1	92.5	55.4	86.4	97.8
boat	70.9	71.9	70.9	69.4	71.3	70.2	78.4	89.6	72.6	80.1	94.2
bottle	28.7	33.1	30.8	29.1	31.5	46.6	42.3	54.4	35.3	45.3	73.5
bus	68.8	60.6	72.2	66.5	71.9	62.3	73.7	81.9	74.5	74.8	95.6
car	78.5	78.0	79.9	77.3	80.4	75.6	83.7	91.5	81.7	85.6	94.3
cat	61.7	58.8	61.4	60.2	61.8	54.9	83.7	91.9	62.5	84.3	95.8
chair	54.3	53.5	56.0	50.2	55.7	63.8	54.3	64.1	60.6	55.9	75.6
cow	48.6	42.6	49.6	46.5	49.6	40.7	61.9	76.3	52.2	63.2	85.4
table	51.8	54.9	58.4	51.9	56.2	58.3	70.2	74.9	59.8	71.5	83.1
dog	44.1	45.8	44.8	44.1	44.7	51.6	79.5	89.7	47.5	80.6	96.6
horse	76.6	77.5	78.8	77.9	79.1	79.2	85.3	92.2	80.2	87.6	97.5
motor	66.9	64.0	70.8	67.1	69.3	68.1	77.2	86.9	73.5	78.3	93.7
person	83.5	85.9	85.0	83.1	84.8	87.1	90.5	95.2	86.2	91.2	98.4
plant	30.8	36.3	31.7	27.6	31.9	49.5	51.1	60.7	34.8	53.8	77.1
sheep	44.6	44.7	51.0	48.5	48.6	48.8	73.8	82.9	52.7	76.9	96.4
sofa	53.4	50.9	56.4	51.1	56.6	56.4	57.0	68.0	57.9	58.1	77.6
train	78.2	79.2	80.2	75.5	79.9	75.9	86.4	95.5	82.1	86.7	97.3
tv	53.5	53.2	57.5	52.3	56.3	54.4	68.0	74.4	60.6	71.8	87.2

be more appropriate. Moreover, CNN is able to improve the performance by deeply exploring the visual information of images. By combining CNN for initial image representation, we can train more discriminative structured weak exemplar classifiers for better categorization.

F. Influences of Parameters

α controls the degree of sparsity of exemplar classifiers. β influences the structure consistency of exemplar classifiers. We give the performance changes with α and β on the four data sets in Figs. 2 and 3, respectively. The performances vary for different parameters on different data sets. We can see from Figs. 2 and 3 that if we set α and β too large, the performances decrease. α should be set to smaller values for data sets with larger class variances. Besides, since the visual representations of the Scene-15 data set are more consistent with each other and hence relatively easier to classify than the other three data sets, the optimal β can be set to smaller values.

V. CONCLUSION

In this paper, we proposed a novel structured weak semantic space construction method and applied it for categorization tasks. Each exemplar classifier was trained to separate one image from the other images of different classes. We jointly learn the exemplar classifiers with structured constraint by restricting the outputs of exemplar classifiers to be low rank. We also imposed sparsity constraint on the exemplar classifiers to improve the discriminative power. We alternatively optimized for the optimal parameters and evaluated the effectiveness of the image representations with categorization tasks. Since the exemplar classifier could be trained with various image representation strategies, the proposed method could make use of more discriminatively image representations for efficient classification. Experimental results on several public data sets proved the usefulness 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] 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.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[4] 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.

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

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

[7] 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.

[8] B. Fan, F. Wu, and Z. Hu, "Aggregating gradient distributions into intensity orders: A novel local image descriptor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2377–2384.

[9] 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.

[10] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

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

[12] 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, 2007.

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

[14] 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.

[15] P. Duygulu, K. Barnard, J. F. G. de Fretias, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 97–112.

[16] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2003, pp. 119–126.

[17] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schimid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 309–316.

[18] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1903–1910.

[19] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Scalable search-based image annotation," *Multimeida Syst.*, vol. 14, no. 4, pp. 205–220, 2008.

[20] X. J. Wang, L. Zhang, F. Jing, and W. Y. Ma, "Image annotation using search and mining technologies," in *Proc. WWW*, 2006, pp. 1045–1046.

[21] F. Wang and M.-Y. Kan, "NPIC: Hierarchical synthetic image classification using image search and generic features," in *Proc. CIVR*, 2006, pp. 473–482.

[22] C. Zhang, J. Cheng, J. Liu, J. Pang, Q. Huang, and Q. Tian, "Beyond explicit codebook generation: Visual representation using implicitly transferred codebooks," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5777–5788, Dec. 2015.

[23] Y. Zhu *et al.*, "Heterogeneous transfer learning for image classification," in *Proc. 25th AAAI Conf. Artifi. Intell.*, 2011, pp. 1–6.

[24] C. Zhang *et al.*, "Undo the codebook bias by linear transformation for visual applications," in *Proc. ACM Multimedia*, 2013, pp. 533–536.

[25] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1778–1785.

[26] D. Parikh and K. Grauman, "Relative attributes," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 503–510.

[27] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 951–958.

[28] G. Patterson, C. Xu, H. Su, and J. Hays, "The SUN attribute database: Beyond categories for deeper scene understanding," *Int. J. Comput. Vis.*, vol. 108, pp. 59–81, Jan. 2014.

[29] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[30] 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.

[31] 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.

[32] J. Zepeda and P. Perez, "Exemplar SVMs as visual feature encoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3052–3060.

[33] D. Modolo, A. Vezhnevets, O. Russakovsky, and V. Ferrari, "Joint calibration of ensemble of exemplar SVMs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3955–3963.

[34] 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.*, USA, Jun. 2009, pp. 1794–1801.

[35] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[36] 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.

[37] R. G. Cinbis, J. Verbeek, and C. Schmid, "Image categorization using fisher kernels of non-iid image models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2184–2191.

[38] 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.

[39] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[40] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.

[41] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3367–3375.

[42] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1666–1674.

[43] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3743–3752.

[44] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233–2246, Nov. 2012.

[45] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. NIPS*, 2006, pp. 801–808.

[46] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Dept. Vis., CalTech, Pasadena, CA, USA, Tech. Rep. CaltechAUTHORS:CNS-TR-2007-001, 2007.

[47] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. ICCVGIP*, Dec. 2008, pp. 722–729.

[48] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, Jun. 2006, pp. 2169–2178.

[49] M. Everingham, A. Zisserman, C. Williams, and L. Van Gool, "The PASCAL visual object classes challenge 2007 (VOC 2007) results," Pascal Challenge, Tech. Rep., 2007. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[50] 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.

[51] 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.

[52] 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.

[53] 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. Processing Syst.*, Vancouver, BC, Canada, 2010, pp. 1–9.

[54] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja, "Low-rank sparse coding for image classification," in *Proc. ICCV*, Dec. 2013, pp. 281–288.

[55] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image classification and retrieval are one," in *Proc. ICMR*, 2015, pp. 3–10.

[56] X.-T. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *Proc. CVPR*, Jun. 2010, pp. 3493–3500.

[57] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proc. CVPR*, Dec. 2013, pp. 811–818.

[58] Y. Chai, E. Rahtu, V. S. Lempitsky, L. J. V. Gool, and A. Zisserman, "TriCoS: A tri-level class-discriminative cosegmentation method for image classification," in *Proc. ECCV*, 2012, pp. 794–807.

[59] S. Ito and S. Kubota, "Object classification using heterogeneous co-occurrence features," in *Proc. ECCV*, 2010, pp. 209–222.

[60] 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.

[61] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in *Proc. BMVC*, 2011, pp. 1–12.

[62] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao, "Group-sensitive multiple kernel learning for object categorization," in *Proc. ICCV*, Sep./Oct. 2009, pp. 436–443.

[63] K. Chatfield, K. Simnyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC*, 2014, pp. 1–12.

[64] L. Zhang, R. Hong, Y. Gao, R. Ji, Q. Dai, and X. Li, "Image categorization by learning a propagated graphlet path," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 3, pp. 674–685, Mar. 2016.

[65] Z. Li, Z. Lai, Y. Xu, J. Yang, and D. Zhang, "A locality-constrained and label embedding dictionary learning algorithm for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 2, pp. 278–293, Feb. 2017, doi: 10.1109/TNNLS.2015.2508025.

[66] H. Xiong, W. Yu, X. Yang, M. N. S. Swamy, and Q. Yu, "Learning the conformal transformation kernel for image recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 1, pp. 149–163, Jan. 2017, doi: 10.1109/TNNLS.2015.2504538.

[67] C. Li, Q. Liu, J. Liu, and H. Lu, "Ordinal distance metric learning for image ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1551–1559, Jul. 2015.

[68] P. Li, Q. Wang, H. Zeng, and L. Zhang, "Local log-Euclidean multivariate Gaussian descriptor and its application to image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 803–817, Apr. 2017.
[69] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[70] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[71] M. Zeiler and R. Ferugs. (2013). "Visualizing and understanding convolutional networks." [Online]. Available: https://arxiv.org/abs/1311.2901
[72] J. Donahue *et al.* (2013). "DeCAF: A deep convolutional activation feature for generic visual recognition." [Online]. Available: https://arxiv.org/abs/1310.1531
[73] Y. Wei *et al.*, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Dec. 2016.
[74] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[75] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 36–45.

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, Beijing, where he joined as an Assistant Professor. He is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include image processing, machine learning, pattern recognition, and computer vision.

Jian Cheng received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 1998 and 2001, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2004.

From 2004 to 2006, he held a post-doctoral position with the Nokia Research Center, Beijing. He has been with the National Laboratory of Pattern Recognition, Beijing, since 2006. His current research interests include machine learning methods and their applications for image processing and social network analysis.

Qi Tian (F'15) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 1992, the M.S. degree in ECE from Drexel University, Philadelphia, PA, USA, in 1996, and the Ph.D. degree in ECE from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2002.

He was a tenure-track Assistant Professor from 2002 to 2008 and a tenured Associate Professor from 2008 to 2012. From 2008 to 2009, he was on Faculty Leave at Microsoft Research Asia, Beijing, as a Lead Researcher with the Media Computing Group. He is currently a Full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA), San Antonio, TX, USA. He authored over 360 refereed journal and conference papers. His current research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics.

Dr. Tian received the 2017 UTSA President's Distinguished Award for Research Achievement, the 2016 UTSA Innovation Award, the 2014 Research Achievement Award from the College of Science, UTSA, the 2010 Google Faculty Award, and the 2010 ACM Service Award. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *ACM Transactions on Multimedia Computing, Communications, and Applications*, and the *Multimedia System Journal*. He is on the Editorial Board of the *Journal of Multimedia* and the *Journal of Machine Vision and Applications*. He is a Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and the *Journal of Computer Vision and Image Understanding*.