Multiview Label Sharing for Visual Representations and Classifications

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

***Abstract*—Different views represent different aspects of images. It is more effective to combine them for visual classifications. This paper proposes a novel multiview label sharing method to combine the discriminative power of different views for classifications. Especially, we linearly transfer different views into a shared space for representations. The inter-view similarities are kept in the shared space for each view. We also ensure the intra-view similarities of the same class between different views are preserved in the shared space. We jointly learn the classifiers and transformation matrices by minimizing the summed classification loss along with the inter-view and intra-view similarity constraints. In this paper, the inter-view constraints refer to the similarities between images of the corresponding view, whereas the intra-view constraints refer to the similarities between different views of images with the same semantics. Experimental results and analysis on several public datasets show the effectiveness of the proposed multiview label sharing method for visual classifications.**

***Index Terms*—Multi-view learning, linear transformation, shared space, image representation, visual classification.**

I. Introduction

THE Internet has huge volumes of multimedia contents. Visual information plays an important role for efficiently analyzing of these multimedia contents. Hence, how to develop discriminative image representations becomes a problem which needs to be solved. Researchers have proposed many visually based methods. For example, the bag-of-visual-words (BoW) model [1], spatial pyramid matching model [2], the sparse coding spatial pyramid matching model (ScSPM) [3], fisher vector based model (FV) [4] and the convolutional neural network (CNN) based model [5] with many variants [6]–[16].

Due to the varieties of images, we often need multiple views for joint representations. In this paper, the views can refer to the visual representations generated by different methods [1]–[5] and can also refer to image representations using different types of features (e.g. color histogram and SIFT) with varied encoding strategies (sparse coding and fisher vector). Combing the discriminative power of different views can greatly improve the classification accuracies. However, different views cannot be compared directly. Simply concatenating them together may not be able to fully explore the discriminative power of different views.

To make use of the discriminative information of multiple views, researchers have proposed many methods [17]–[25]. Some works [17]–[18] try to combine different views by graph fusion. Others [19]–[21] make transformations either by transfer learning or learning general spaces for representations. on one hand, the similarities of samples within one view are often used while the correlations of different views are ignored or contaminated by noisy correlations. On the other hand, the combinations of different views are independent of the classifier training process.

Similarity preserving strategy is often used to ensure consistency during transformation. This is often based on the assumption that the initially similar samples should be transformed to similar representations. There are mainly two types (feature based [26]–[28] and semantics based [29]–[32]) of similarity preserving methods. The feature based methods assume visually similar features should be transformed into similar representations while the semantic based methods use the semantic similarities during transformation. However, the two types of similarity preserving strategies are often used independently without jointly considering their discriminative abilities. The similarity consistency constraints work well but may occasionally fail on the borders of different classes. The visually similar constraints should be combined with their semantic correlations.

To solve the problems mentioned above, in this paper, we propose a novel multi-view label sharing method (MVLS) for efficient visual classifications. To bridge the gaps of different views, we linearly transform them into a shared space. For each view, we try to preserve the inter-view similarity in the shared space. The intra-view similarities of images of different views

Manuscript received April 19, 2017; revised July 28, 2017 and September 4, 2017; accepted September 4, 2017. Date of publication October 5, 2017; date of current version March 15, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 61303154 and 61332016, and in part by the Scientific Research Key Program of Beijing Municipal Commission of Education under Grant KZ201610005012. The work of Q. Tian was supported in part by ARO Grant W911NF-15-1-0290, in part by the Faculty Research Gift Awards by NEC Laboratories of America and Blippar, and in part by the National Science Foundation of China under Grant 61429201. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abdulmotaleb El Saddik. *(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, and also with the University of Chinese Academy of Sciences, Beijing 100049, 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, with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jcheng@nlpr.ia.ac.cn).

Q. Tian is with the Department of Computer Science, 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/TMM.2017.2759500

are also combined along with the classifier training process. We alternatively optimize over the classifier parameters and the transformation matrixes for visual representations and classifications. Experimental results on several datasets well demonstrate the effectiveness of the proposed MVLS method for image classification.

The proposed method differs with latent models (e.g. Latent Semantic Analysis [33], Probabilistic Latent Semantic Analysis [34] and Latent Dirichlet Allocation [35]). Latent spaces are not explicitly generated while the shared space can be obtained directly. MVLS is also different from Canonical Correlation Analysis [21] as it jointly learn classifiers along with transformation matrixes.

The main contributions of this paper lie in three aspects:

- First, we linearly transfer the information of different views into a shared space for visual representations and classifications.
- Second, we combine the classifier training and transformation matrix learning into a unified process by considering both the inter-view similarities and intra-view similarities.
- Third, we alternatively optimize over the classifier parameters and transformation matrixes. We achieve superior classification performances compared with other baseline methods.

The rest of this paper is organized as follows. We discuss the related work in Section II. The details of the proposed multi-view label sharing method for visual representations and classifications are given in Section III. Image classification experiments on several datasets are given in Section IV. Finally, we conclude in Section V.

II. Related Work

Many methods had been introduced to improve the accuracy of image classification [1]–[16], e.g. the BoW model [1], SPM [2], ScSPM [3], FV [4] and locality constrained linear coding (LLC) [6]. In recent years, the deep convolutional neural network based methods [5], [8], [9] became popular. Krizhevsky *et al.* [8] proposed to use deep convolutional neural networks for large scale image classification. Simonyan and Zisserman [5] increased the depth of convolutional neural networks. Perronnin and Larlus [9] proposed a hybrid architecture by combing the fisher vector with convolutional neural networks. Motivated by these methods, many works [10]–[16] were made. The low-rank sparse coding was used by Zhang *et al.* [10] to generate general and class specific codebooks. Gao *et al.* [11] incorporated similarity preserving term into the sparse coding process. Huang *et al.* [13] targeted the fine-grained classification problem with polygon based classifier learning while Zhao *et al.* [14] made use of diversified attention networks. Wu *et al.* [15] combined active learning with label correlation while Puthenputhussery *et al.* [16] used the complete marginal fisher analysis technique. However, these methods only tried to classify images from single view without exploring the correlations of different views. Graph based fusion [17]–[18] strategies were used to combine various information. Liu *et al.* [17] constructed hypergraph for image clustering and classification. Shao and Fu [18] proposed a novel hierarchical hyperlingual-words algorithm to classify faces with different modalities. The joint consideration of different views helped to boost the performances over single view.

Instead of using the initial representations, transformation based strategies [19]–[22] were also used. It worked by transferring the initial representations to new features which were more discriminative and effective for new applications. Zhang *et al.* [19] implicitly transferred pre-learned codebooks for image classification. Dai *et al.* [20] tried to transfer knowledge among different feature spaces while Gong *et al.* [21] used multi-view embedding space. Wang *et al.* [22] tried to learn compact hash codes for multimodal representation. Chou *et al.* [23] made use of different information for video annotation. Wang and Guo [24] imposed sparse constraint to embed different modalities while Li *et al.* [25] explored the correlations between image and text for news topic analysis. To ensure consistency of the transformed representations, researchers also used various similarity preserving constraints [26]–[39]. Zhang *et al.* [26] contextualized exemplar classifiers while Dong *et al.* [27] learned the subcategory for classification. Yuan and Yan [28] explored multi-task sparse representation. Oquab *et al.* [29] used convolutional neural networks for mid-level image representations. Li *et al.* [30] learned low-rank constrained subspace while Zhang *et al.* [31] classified objects in sub-semantic space. To cope with the data missing problem, Shao *et al.* [32] proposed a novel sparse low-rank fusion based deep features for face recognition while Ding *et al.* [39] made use of the latent low-rank constraint with improved performances. Researchers also tried to learn the transformation implicitly [33]–[35]. Most of these methods only considered the inter-view similarities. Multiple views were needed for efficient representations. However, the correlations between different views were often ignored. This information should also be used for reliable transformations.

To improve the classification performances, researchers made use of many strategies [40]–[55] and greatly boosted the accuracies. The local feature based strategies tried to go beyond the Euclidean distance [40] or make use the detection and segmentation techniques [41]–[42]. Xie *et al.* [40] used bin-ratio similarity for classification. Angelova and Zhu [41] combined object detection and segmentation strategies for fine-grained application while Chai *et al.* [42] proposed a co-segmentation method. Zhang *et al.* [48] proposed a low-rank sparse coding scheme for classification. Boiman *et al.* [43] went beyond local feature encoding process and used local features directly with class-level similarity. To bridge the semantic gap, Li *et al.* [44] collected images from the Internet and used the ObjectBank for semantic representations of images. The CNN based methods were widely used recently [45], [47]. Sohn *et al.* [45] tried to learn sparse convolutional features while Bo *et al.* [46] used a hierarchical strategy. Zeiler and Ferugs [47] studied the details of convolutional networks while Sermanet *et al.* [50] used convolutional networks to integrate localization, detection and classification into a unified framework. Chatfield *et al.* [52] evaluated the implementation details of different convolutional network based methods. Wei *et al.* [53] studied the multi-label classification problem with CNN based methods while Yang *et al.* [54] used bounding box information. One problem

TABLE I
THE SYMBOLS AND THEIR CORRESPONDING DESCRIPTIONS USED IN THIS PAPER

Symbol	Description
M	view number
N	number of training images
D^m	dimension of the m-th view
$\mathcal{Z}$	shared space
d	dimension of $\mathcal{Z}$
$\boldsymbol{x}_n^m$	representation of n-th image of m-th view
y_n	label of the n-th image
$\boldsymbol{W}^m$	transformation matrix of m-th view
$f(\boldsymbol{z})$	linear classifier for prediction
$\ell(*,*)$	hinge loss function
$\boldsymbol{\alpha}^T$	transpose of $\boldsymbol{\alpha}$
$s_{i,j}^m$	similarity between the i-th and j-th images of m-th view
$\hat{s}_{i,j}^{m,\tilde{m}}$	similarity between the i-th image of m-th view and the j-th image of $\tilde{m}$-th view
$\lambda, \lambda_1, \lambda_2$	balancing parameters

with these methods was that they only used the visual information and worked well on carefully collected datasets. However, combining multiple types of information was more proper for real applications [56]–[67]. Rasiwasia *et al.* [56] used both images and text descriptions for cross-modal retrieval. CCA [57] was used by Yang *et al.* for face matching [58]. Wang *et al.* [59] combined the deep semantic network for multimodal representations while Sharma *et al.* [60] used generalized multi-view analysis technique. The cross-modal correlation problem was also explored by feature selection with subspace learning [61] and deep metric learning [62]. However, they treated the combination of multi-views and classifier training separately. It would be more effective to incorporate the discriminative power of multi-views along with classifier training into a unified framework.

III. Multi-View Label Sharing Based Visual Representations and Classifications

In this section, we give the details of the proposed multi-view label sharing method for visual representations and classifications (MVLS). The symbols and their descriptions used in this paper are given in Table I.

A. Linear Transformation

Let M be the view number, D^m is the dimension of the m-th view with $m = 1, \ldots, M$. Suppose we have N training images with their multi-view representations and labels as $(\boldsymbol{x}_n^m, y_n), m = 1, \ldots, M, n = 1, \ldots, N$. Different views cannot be directly compared. To combine the discriminative power of each view, we try to linearly transform each view into a shared space $\mathcal{Z} \in \mathbb{R}^{d\times 1}$ as:

$$\boldsymbol{z}_n^m = \boldsymbol{W}^m \times \boldsymbol{x}_n^m \tag{1}$$

where $\boldsymbol{W}^m \in \mathbb{R}^{d\times D^m}$ is the linear transformation matrix for the m-th view. Linear transformation is used because it can be efficiently combined for visual representations. It is also differentiable for efficient optimization.

B. Inter-View and Intra-View Similarity Measurements

we can use the linearly transferred representations for classification directly as:

$$\widehat{y} = f(\boldsymbol{z}) \tag{2}$$

where f is the corresponding classifier to be learned. We can learn this classifier by minimizing the summed loss of training images with constraints as:

$$f = \mathrm{argmin}_f \sum_{n=1}^{N}\sum_{m=1}^{M} \ell(f(\boldsymbol{z}_n^m), y_n) + \lambda\Omega(f(*)) \tag{3}$$

where $\ell(*,*)$ is the loss function to be learned, $\Omega(*)$ is the regularization term. λ is the parameter for balancing the influences of summed loss and the regularization term. N is the number of training images and M is the number of views. Usually, the hinge loss is used:

$$\ell(f(\boldsymbol{z}_n^m), y_n) = \max(0, 1 - f(\boldsymbol{z}_n^m) \times y_n) \tag{4}$$

In this paper, we use the linear classifier $f(\boldsymbol{z}) = \boldsymbol{\alpha}^T\boldsymbol{z}$ with L_2 regularization. Problem 3 can be rewritten as:

$$\boldsymbol{\alpha} = \mathrm{argmin}_{\boldsymbol{\alpha}} \sum_{n=1}^{N}\sum_{m=1}^{M} \max(0, 1 - \boldsymbol{\alpha}^T\boldsymbol{z}_n^m \times y_n) + \lambda\|\boldsymbol{\alpha}\|_2 \tag{5}$$

We can learn the classifiers by solving Problem 5 directly. However, it does not fully explore the similarity information. We transform the representations of different views into the shared space. However, the transformed representations of different views have varied semantic meanings. Hence, it is necessary to impose some constraints on the transformation process to ensure semantically consistent transformation. We leverage two types of constraints (inter-view similarity and intra-visual similarity) in this paper. On one hand, visually similar images should have similar representations in the shared space for each view. This inter-view similarity constraint is widely used by researchers [3], [6], [7], [10], [11], [20]. On the other hand, one image may be represented with different views. The representations with the same semantics should also be transformed with correlated representations in the shared space. Due to the semantic gap, the similarity constraint may fail on some images and lead to the over-fitting problem. However, we can try to minimize a proper loss function to statistically ensure efficient combinations of inter-view and intra-view similarities. Balancing parameters can also be used to avoid the over-usage of similarities.

To ensure inter-view similarity, we add an inter-view similarity preserving term to Problem 5 with weighting parameter λ_1 as:

$$\boldsymbol{\alpha} = \mathrm{argmin}_{\boldsymbol{\alpha}} \sum_{n=1}^{N}\sum_{m=1}^{M} \max(0, 1 - \boldsymbol{\alpha}^T\boldsymbol{z}_n^m \times y_n) + \lambda\|\boldsymbol{\alpha}\|_2 + \lambda_1 \sum_{m=1}^{M}\sum_{i,j=1}^{N} s_{i,j}^m \|\boldsymbol{z}_i^m - \boldsymbol{z}_j^m\|_2^2 \tag{6}$$

The similarity between $\boldsymbol{x}_i^m$ and $\boldsymbol{x}_j^m$ is defined as:

$$s_{i,j}^m = \exp^{-\|\boldsymbol{x}_i^m - \boldsymbol{x}_j^m\|_2/\sigma} \tag{7}$$

where σ is the scaling parameter.

To model the intra-class similarity, we add another term by considering the intra-view similarity to Problem 6:

$$\begin{aligned}\boldsymbol{\alpha} = \operatorname{argmin}_{\boldsymbol{\alpha}} & \sum_{n=1}^{N}\sum_{m=1}^{M} \max(0, 1 - \boldsymbol{\alpha}^T \boldsymbol{z}_n^m \times y_n) \\ & + \lambda\|\boldsymbol{\alpha}\|_2 + \lambda_1 \sum_{m=1}^{M}\sum_{i,j=1}^{N} s_{i,j}^m \|\boldsymbol{z}_i^m - \boldsymbol{z}_j^m\|_2^2 \\ & + \lambda_2 \sum_{m=1}^{M}\sum_{\widetilde{m}=1}^{M}\sum_{i,j=1}^{N} \widehat{s}_{i,j}^{m,\widetilde{m}} \|\boldsymbol{z}_i^m - \boldsymbol{z}_j^{\widetilde{m}}\|_2^2\end{aligned} \tag{8}$$

with λ_2 is the weighting parameter. The intra-view similarity $\widehat{s}_{i,j}^{m,\widetilde{m}}$ is defined as:

$$\widehat{s}_{i,j}^{m,\widetilde{m}} = \begin{cases} 1, \text{if} \quad y_i == y_j \\ 0, \text{otherwise} \end{cases} \tag{9}$$

It is used to ensure that we only consider the intra-view similarities of images with the same semantics.

For fixed transformation matrixes $\boldsymbol{W}^m, m = 1, \ldots, M$, the second and third terms of Problem 8 have no influences on $\boldsymbol{\alpha}$. However, we do not know the transformation matrixes in advance. $\boldsymbol{W}^m, m = 1, \ldots, M$ should also be learned from training images as:

$$\begin{aligned}[\boldsymbol{\alpha}, \boldsymbol{W}^m] = \operatorname{argmin}_{\boldsymbol{\alpha}, \boldsymbol{W}^m} & + \lambda\|\boldsymbol{\alpha}\|_2 \\ & + \sum_{n=1}^{N}\sum_{m=1}^{M} \max(0, 1 - \boldsymbol{\alpha}^T \boldsymbol{z}_n^m \times y_n) \\ & + \lambda_1 \sum_{m=1}^{M}\sum_{i,j=1}^{N} s_{i,j}^m \|\boldsymbol{z}_i^m - \boldsymbol{z}_j^m\|_2^2 \\ & + \lambda_2 \sum_{m=1}^{M}\sum_{\widetilde{m}=1}^{M}\sum_{i,j=1}^{N} \widehat{s}_{i,j}^{m,\widetilde{m}} \|\boldsymbol{z}_i^m - \boldsymbol{z}_j^{\widetilde{m}}\|_2^2 \\ & \forall \boldsymbol{z}_i^m = \boldsymbol{W}^m \boldsymbol{x}_i^m, m = 1, \ldots, M\end{aligned} \tag{10}$$

C. Alternative Optimization for Visual Representations and Classifications

It is hard to learn the optimal classifier parameter $\boldsymbol{\alpha}$ and the transformation matrixes $\boldsymbol{W}^m, m = 1, \ldots, M$ simultaneously. To solve Problem 10, we alternatively optimize over the classifier parameter $\boldsymbol{\alpha}$ and the transformation matrixes $\boldsymbol{W}^m, m = 1, \ldots, M$. When $\boldsymbol{W}^m, m = 1, \ldots, M$ are fixed, Problem 10 equals to:

$$\begin{aligned}\boldsymbol{\alpha} = \operatorname{argmin}_{\boldsymbol{\alpha}} & \sum_{n=1}^{N}\sum_{m=1}^{M} \max(0, 1 - \boldsymbol{\alpha}^T \boldsymbol{z}_n^m \times y_n) \\ & + \lambda\|\boldsymbol{\alpha}\|_2\end{aligned} \tag{11}$$

which is a standard support vector machine learning problem with hinge loss and L_2 constraint. It can be solved efficiently with the state-of-the-art SVM solver.

When $\boldsymbol{\alpha}$ is fixed, Problem 10 equals to:

$$\begin{aligned}\boldsymbol{W}^m = \operatorname{argmin}_{\boldsymbol{W}^m} & \lambda_1 \sum_{m=1}^{M}\sum_{i,j=1}^{N} s_{i,j}^m \|\boldsymbol{z}_i^m - \boldsymbol{z}_j^m\|_2^2 \\ & + \sum_{n=1}^{N}\sum_{m=1}^{M} \max(0, 1 - \boldsymbol{\alpha}^T \boldsymbol{z}_n^m \times y_n) \\ & + \lambda_2 \sum_{m=1}^{M}\sum_{\widetilde{m}=1}^{M}\sum_{i,j=1}^{N} \widehat{s}_{i,j}^{m,\widetilde{m}} \|\boldsymbol{z}_i^m - \boldsymbol{z}_j^{\widetilde{m}}\|_2^2 \\ & \forall \boldsymbol{z}_i^m = \boldsymbol{W}^m \boldsymbol{x}_i^m, m = 1, \ldots, M\end{aligned} \tag{12}$$

We try to optimize over each transformation matrix by fixing the other matrixes. By ignoring the fixed terms, for the m-th transformation matrix, Problem 12 equals to the following optimization problem as:

$$\begin{aligned}\boldsymbol{W}^m = \operatorname{argmin}_{\boldsymbol{W}^m} & \sum_{n=1}^{N} \max(0, 1 - \boldsymbol{\alpha}^T \boldsymbol{z}_n^m \times y_n) \\ & + \lambda_1 \sum_{i,j=1}^{N} s_{i,j}^m \|\boldsymbol{W}^m(\boldsymbol{x}_i^m - \boldsymbol{x}_j^m)\|_2^2 \\ & + \lambda_2 \sum_{\widetilde{m}=1}^{M}\sum_{i,j=1}^{N} \widehat{s}_{i,j}^{m,\widetilde{m}} \|\boldsymbol{W}^m \boldsymbol{x}_i^m - \boldsymbol{W}^{\widetilde{m}} \boldsymbol{x}_j^{\widetilde{m}}\|_2^2\end{aligned} \tag{13}$$

which can be solved using gradient descent. Let

$$\begin{aligned}\Phi(\boldsymbol{W}^m) = & \sum_{n=1}^{N} \max(0, 1 - \boldsymbol{\alpha}^T \boldsymbol{z}_n^m \times y_n) \\ & + \lambda_1 \sum_{i,j=1}^{N} s_{i,j}^m \|\boldsymbol{W}^m(\boldsymbol{x}_i^m - \boldsymbol{x}_j^m)\|_2^2 \\ & + \lambda_2 \sum_{\widetilde{m}=1}^{M}\sum_{i,j=1}^{N} \widehat{s}_{i,j}^{m,\widetilde{m}} \|\boldsymbol{W}^m \boldsymbol{x}_i^m - \boldsymbol{W}^{\widetilde{m}} \boldsymbol{x}_j^{\widetilde{m}}\|_2^2\end{aligned} \tag{14}$$

we can calculate the gradient of $\Phi(\boldsymbol{W}^m)$ as:

$$\begin{aligned}\frac{\partial\Phi}{\partial\boldsymbol{W}^m} = & \sum_{n=1}^{N} \frac{\partial \max(0, 1 - \boldsymbol{\alpha}^T \boldsymbol{z}_n^m \times y_n)}{\partial\boldsymbol{W}^m} \\ & + 2\lambda_1 \sum_{i,j=1}^{N} s_{i,j}^m \boldsymbol{W}^m(\boldsymbol{x}_i^m - \boldsymbol{x}_j^m)(\boldsymbol{x}_i^m - \boldsymbol{x}_j^m)^T \\ & + 2\lambda_2 \sum_{i,j=1}^{N} \widehat{s}_{i,j}^{m,\widetilde{m}} (\boldsymbol{W}^m \boldsymbol{x}_i^m - \boldsymbol{W}^{\widetilde{m}} \boldsymbol{x}_j^{\widetilde{m}})(\boldsymbol{x}_i^m)^T\end{aligned} \tag{15}$$

with $\frac{\partial \max(0,1-\boldsymbol{\alpha}^T \boldsymbol{z}_n^m \times y_n)}{\partial\boldsymbol{W}^m} = 0$, if $1 - y_n\boldsymbol{\alpha}^T\boldsymbol{W}^m\boldsymbol{x}_n^m \leq 0$ and $\frac{\partial \max(0,1-\boldsymbol{\alpha}^T \boldsymbol{z}_n^m \times y_n)}{\partial\boldsymbol{W}^m} = -y_n\boldsymbol{\alpha}(\boldsymbol{x}_n^m)^T$, if $1 - y_n\boldsymbol{\alpha}^T\boldsymbol{W}^m\boldsymbol{x}_n^m > 0$.

Algorithm 1: The Alternative Optimization Strategy for Solving Problem 10

```
Input:
    The parameters: $\lambda$, $\lambda_1$, $\lambda_2$; training images of $M$ views
    $(\boldsymbol{x}_n^m, y_n)$, $n = 1, \ldots, N, m = 1, \ldots, M$, maximum
    iteration number $M_{\text{iter}}$, the stopping threshold $\theta$.
Output:
    The classifier parameter $\boldsymbol{\alpha}$ and the transformation
    matrixes $\boldsymbol{W}^m, m = 1, \ldots, M$;
1: for iter = 1:$M_{\text{iter}}$
2:     Search for the optimal classifier parameter $\boldsymbol{\alpha}$ while
       keeping the transformation matrixes $\boldsymbol{W}^m, m = 1, \ldots,$
       $M$ fixed using the state-of-the-art SVM classifiers;
3:     Search for the optimal transformation matrix $\boldsymbol{W}^m$
       for the $m$-th view iteratively with fixed classifier
       parameter $\boldsymbol{\alpha}$ and the other transformation matrixes
       by solving Problem 13;
4:     Check the decrement of the objective value
       of Eq. (10).
         If the decrement falls below $\theta$, stop, go to step 5.
         else go to step 2.
5: end for.
6: return The learned $\boldsymbol{\alpha}$ and $\boldsymbol{W}^m, m = 1, \ldots, M$.
```

We alternatively optimize over the parameter $\boldsymbol{\alpha}$ and the transformation matrixes $\boldsymbol{W}^m, m = 1, \ldots, M$ for a pre-defined iterations or the changes of the objective value of Problem 10 falls below a threshold. Algorithm 1 gives the procedure of the alternative optimization method. After the transformation matrixes and the classifier parameters are learned, we can use them for representations and classifications of images. We assign the images with the class whose corresponding classifier has the largest response.

IV. Experiments

To evaluate the performances of the proposed method, we conduct image classification experiments on several datasets: the Flower-102 dataset [36], the Caltech-256 dataset [37], the PASCAL VOC 2007 dataset and the PASCAL VOC 2012 dataset [38]. Fig. 1 shows some example images of these datasets. We also test the effectiveness of the proposed method on the wiki dataset [56].

A. Experimental Setup

We extract different types of features for image presentations and treat each type of representations as one view. Specially, we use both the local feature based methods and convolutional neural network based methods for representations. As to the local feature based methods, the SIFT and color-SIFT [39] are used and encoded with the BoW model, the ScSPM model and the FV model respectively. We extract local features densely by grid search with overlap. The overlap is set to 6 pixels with the minimum local region of 16 × 16 pixels. The codebook size is set to 1,000 for the BoW model and the ScSPM model. 200 is used for the FV model. We use the code provided by [3]–[6] to get the image representations of different views. This results in a total of six views for local feature based representations. We use three views of convolutional neural network based strategies. Specially, we adopt the same structure as [5] and [8] for image representations with 4,096 dimensions. We use the pre-trained AlexNet [8], VGG [5] and ResNet [55] with 50 layers on the ILSVRC 2014 dataset for initialization. We then finely tune the networks with the training images of each dataset. Finally, we remove the last fully-connected layer and use the 4,096 dimensions of the penultimate layer as image representations of the corresponding network (view). Hence, we can get a total of nine views for image representations. The maximum iteration number in Algorithm 1 is set to 50. We only use the six local feature based views for classifications on the Flower-102 dataset (MVLS (6 views)) as this dataset is relatively easier to classify. The classification rate is used for performance evaluations on the Flower-102 dataset and the Caltech-256 dataset. Mean average precision (mAP) is used on the PASCAL VOC 2007 dataset and the PASCAL VOC 2012 dataset for evaluations.

We follow the experimental setup as other baseline methods and compare with the reported results directly for fair comparison. The baseline methods are chosen for three reasons. First, the proposed method combines multiple views for classification. Hence, we compare with the performances of different views [3], [4], [5], [6], [11], [51] to show the effectiveness of multi-view combination. Second, we transfer different views for joint classification. Hence, we also compare with some of the transformation based methods [10], [19], [31], [40], [44], [46]. Third, we test the proposed method on several public datasets. Many works have tested their performances on these datasets. We compare with several the state-of-the-art methods [28], [36], [37], [41]–[43], [47]–[54] which use local features and convolutional neural networks. We give the performances of MVLS (6 views, no intra), MVLS (6 views, no inter+intra), MVLS (6 views) and MVLS (9 views) on the Flower-102 dataset and the Caltech-256 dataset to show the detailed effectiveness of the proposed method. Since the combination of inter-view and intra-view information of different views can consistently improve the classification performances, we only give the performances of MVLS (9 views) on the PASCAL VOC 2007 and 2012 datasets.

B. The Flower-102 Dataset

The Flower-102 dataset has 102 classes of 8,189 images with varied numbers for each class. The images are divided into 10/10/rest for train/validate/test respectively [36]. We follow this setup and conduct image classification experiments accordingly. Table II gives the performances of the proposed method along with other baseline methods. We also give the performance of the proposed method when no intra-view similarity is added (MVLS (6 views, no intra)). This corresponds to set λ_2 in Problem 10 to zero. The performance of MVLS with no inter-view and intra-view similarities is also given in Table II (MVLS (6 views, no inter+intra)). Its performance can be obtained by setting λ_1 and λ_2 to zero in Problem 10.

Fig. 1. Example images of (a) the Flower-102 dataset, (b) the Caltech-256 dataset, (c) the PASCAL VOC 2007 dataset and (d) the PASCAL VOC 2012 dataset.

TABLE II
CLASSIFICATION RESULT COMPARISONS (%) ON THE FLOWER-102 DATASET

Methods	Classification rate
LR-GCC [10]	75.7
ICT [19]	77.3
KMTJSRC-CG [28]	74.1
S^3 R [31]	85.3
Nilsback [36]	72.8
Xie [40]	86.8
Det+Seg [41]	80.7
TriCoS [42]	85.2
MVLS (6 views, no inter+intra)	86.3
MVLS (6 views, no intra)	89.5
MVLS (6 views)	91.6

We can see from Table II that the proposed method can improve the classification performances over baseline methods. Specially, MVLS (6 views) is able to outperform both visual and semantic based methods. MVLS (6 views) also improves over KMTJSRC-CG [28] which automatically learns the combinations from images. Yuan and Yan [28] make use of joint sparse reconstruction minimization for image classification with improved performances. Moreover, the proposed method works better than [41] which makes use of the region level annotation while the proposed method only uses image level labels. Finally, by combining the inter-view similarity and the intra-view similarity, we can consistently improve the classification performances over MVLS (6 views, no inter+intra) and MVLS (6 views, no intra). These results on the Flower-102 dataset prove the effectiveness of the proposed multi-view label sharing method for image classification.

C. The Caltech-256 Dataset

There are 29,780 images of 256 classes in the Caltech-256 dataset with at least 80 images for each class. We follow the experimental setup as [37] and randomly select 15/30/45/60 training images per class for training and use the other images for testing. The random selection process is repeated for ten times and the mean and standard variation of the results are used for performance evaluations.

Table III shows the experimental results of MVLS with different view settings along with other methods. We can have three conclusions from Table III. First, the combination of different views is useful for reliable classification. MVLS is able to improve over ScSPM [3], FV [4] and CNN based methods [45], [47] dramatically. Second, the performances of MVLS can be boosted by considering more views. Hence, MVLS (9 views) improves the performances dramatically over MVLS (6 views). Third, the inter-view and intra-view information can help to represent images discriminatively as the performances can be consistently improved with the adding of inter-view and intra-view information.

D. The PASCAL VOC 2007 Dataset

There are twenty classes (*aeroplane, bicycle, boat, bottle, bus, bird, car, cat, cow, chair, dining table, dog, horse, person, sheep, motorbike, train, potted plant, sofa* and *tv/monitor*) of 9,963 images in the PASCAL VOC 2007 dataset. There are often multiple objects with different poses and occlusions. Images are divided with train/validate/test sets [38].

Table IV gives the average precision comparisons of MVLS (9 views) with other baseline methods on the PASCAL VOC 2007 dataset. We can have three conclusions from Table IV. First, the proposed MVLS (9 views) method improves the performances over baseline methods. By combining the discriminative information of different views, we are able to improve over each view [4]–[6], [38], [51]. Second, MVLS (9 views) also improves over detection based methods [49], [50] which use more information than class level annotations. Third, MVLS (9 views) has larger improvements over non-rigid classes than rigid classes. We use

TABLE III
PERFORMANCE COMPARISONS (%) ON THE CALTECH-256 DATASET

Methods	15 images	30 images	45 images	60 images
ScSPM [3]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
FV [4]	38.50 ± 0.20	47.40 ± 0.10	52.10 ± 0.40	54.80 ± 0.40
VGG [5]	–	–	–	86.20 ± 0.30
LLC [6]	27.74 ± 0.32	32.07 ± 0.24	35.09 ± 0.44	37.79 ± 0.42
LR-GCC [10]	39.21 ± 0.48	45.87 ± 0.41	–	–
LScSPM [11]	30.00 ± 0.14	35.74 ± 0.10	38.47 ± 0.51	40.32 ± 0.32
S^3 R [31]	37.85 ± 0.48	43.52 ± 0.44	46.86 ± 0.63	–
KSPM [37]	23.34 ± 0.42	29.51 ± 0.52	–	–
NBNN(1 Desc) [43]	30.45	38.18	–	–
ObjectBank [44]	39.00	–	–	–
SDC [45]	35.10	42.10	45.70	47.90
MSC [46]	40.50 ± 0.40	48.00 ± 0.20	51.90 ± 0.20	55.20 ± 0.30
VUCN [47]	65.70 ± 0.20	70.60 ± 0.20	72.70 ± 0.40	74.20 ± 0.30
LRSC [48]	–	41.04 ± 0.23	–	–
MVLS (6 views, no inter+intra)	63.58 ± 0.72	69.40 ± 0.81	71.28 ± 0.65	72.56 ± 0.62
MVLS (6 views, no intra)	66.76 ± 0.80	71.25 ± 0.74	73.51 ± 0.68	74.63 ± 0.66
MVLS (6 views)	72.53 ± 0.84	75.68 ± 0.61	77.45 ± 0.72	78.82 ± 0.69
MVLS (9 views)	77.39 ± 0.62	84.23 ± 0.58	87.72 ± 0.56	89.95 ± 0.63

TABLE IV
PERFORMANCE COMPARISONS (%) ON THE PASCAL VOC 2007 DATASET

object class	LLC [6]	Best07 [38]	FV [4]	INRIA [49]	Overfeat [50]	DeCAF [51]	VGG [5]	Chatfield [52]	HCP-VGG [53]	MVLS (9 views)
airplane	74.8	77.5	80.0	77.2	88.5	87.4	–	95.3	98.6	99.2
bicycle	65.2	63.6	67.4	69.3	81.0	79.3	–	90.4	97.1	98.5
bird	50.7	56.1	51.9	56.2	83.5	84.1	–	92.5	98.0	98.8
boat	70.9	71.9	70.9	66.6	82.0	78.4	–	89.6	95.6	97.7
bottle	28.7	33.1	30.8	45.5	42.0	42.3	–	54.4	75.3	80.2
bus	68.8	60.6	72.2	68.1	72.5	73.7	–	81.9	94.7	96.4
car	78.5	78.0	79.9	83.4	85.3	83.7	–	91.5	95.8	96.7
cat	61.7	58.8	61.4	53.6	81.6	83.7	–	91.9	97.3	98.8
chair	54.3	53.5	56.0	58.3	59.9	54.3	–	64.1	73.1	78.1
cow	48.6	42.6	49.6	51.1	58.5	61.9	–	76.3	90.2	92.4
table	51.8	54.9	58.4	62.2	66.5	70.2	–	74.9	80.0	84.2
dog	44.1	45.8	44.8	45.2	77.8	79.5	–	89.7	97.3	97.9
horse	76.6	77.5	78.8	78.4	81.8	85.3	–	92.2	96.1	97.3
motorbike	66.9	64.0	70.8	69.7	78.8	77.2	–	86.9	94.9	96.7
person	83.5	85.9	85.0	86.1	90.2	90.5	–	95.2	96.3	98.2
plant	30.8	36.3	31.7	52.4	54.8	51.1	–	60.7	78.3	81.2
sheep	44.6	44.7	51.0	54.4	71.1	73.8	–	82.9	94.7	97.3
sofa	53.4	50.9	56.4	54.3	62.6	57.0	–	68.0	76.2	80.1
train	78.2	79.2	80.2	75.8	87.4	86.4	–	95.5	97.9	98.5
tv	53.5	53.2	57.5	62.1	71.8	68.0	–	74.4	91.5	93.3
mAP	59.3	59.4	61.7	63.5	73.9	73.4	89.3	82.4	90.9	93.1

both inter-view similarity and intra-view similarity jointly for discriminative representations and classifications. The results on PASCAL VOC 2007 dataset prove the usefulness of using multi-view label sharing for classification.

E. The PASCAL VOC 2012 Dataset

This dataset has the same twenty classes as the PASCAL VOC 2007 dataset. There are 22,531 images with the train+val/test splits as 11,540/10,991. We follow the same experimental setup as [38] and train classifiers accordingly for image class prediction.

We give the performance comparisons with other methods in Table V. We can see that MVLS (9 views) outperforms many convolutional neural networks based methods [5], [29], [47], [52], [54] by combining the discriminative information of different views. On analysis of the per-class performances, we can have similar conclusions as on the PASCAL VOC 2007 dataset. MVLS (9 views) is able to have larger improvements on non-rigid classes over baseline methods because of the preserving of inter-view and intra-view similarities. The proposed method is also able to improve over HCP-VGG which uses object segment hypothesis to assist the classification task. We only use the PASCAL VOC 2012 dataset instead of using learned architectures from other sources. The classification improvements on the PASCAL VOC 2012 dataset prove the effectiveness of the proposed method again.

TABLE V
PERFORMANCE COMPARISONS (%) ON THE PASCAL VOC 2012 DATASET

object class	VUCN [47]	NUS-PSL [47]	PRE-1000C [29]	Chatfield [52]	VGG [5]	Yang [54]	PRE-1000C [29]	HCP-VGG [53]	MVLS (9 views)
airplane	96.0	97.3	93.5	96.8	99.0	98.9	94.6	99.1	99.4
bicycle	77.1	84.2	78.4	82.5	88.8	93.1	82.9	92.8	95.3
bird	88.4	80.8	87.7	91.5	95.9	96.0	88.2	97.4	98.7
boat	85.5	85.3	80.9	88.1	93.8	94.1	84.1	94.4	96.1
bottle	55.8	60.8	57.3	62.1	73.1	76.4	60.3	79.9	82.8
bus	85.8	89.9	85.0	88.3	92.1	93.5	89.0	93.6	95.2
car	78.6	86.8	81.6	81.9	85.1	90.8	84.4	89.8	92.6
cat	91.2	89.3	89.4	94.8	97.8	97.9	90.7	98.2	98.9
chair	65.0	75.4	66.9	70.3	79.5	80.2	72.1	78.2	83.4
cow	74.4	77.8	73.8	80.2	91.1	92.1	86.8	94.9	96.5
table	67.7	75.1	62.0	76.2	83.3	82.4	69.0	79.8	83.6
dog	87.8	83.0	89.5	92.9	97.2	97.2	92.1	97.8	98.9
horse	86.0	87.5	83.2	90.3	96.3	96.8	93.4	97.0	99.1
motorbike	85.1	90.1	87.6	89.3	94.5	95.7	88.6	93.8	96.7
person	90.9	95.0	95.8	95.2	96.9	98.1	96.1	96.4	97.2
plant	52.2	57.8	61.4	57.4	63.1	73.9	64.3	74.3	79.5
sheep	83.6	79.2	79.0	83.6	93.4	93.6	86.6	94.7	95.6
sofa	61.1	73.4	54.3	66.4	75.0	76.8	62.3	71.9	77.8
train	91.8	94.5	88.0	93.5	97.1	97.5	91.1	96.7	98.4
tv	76.1	80.7	78.3	81.9	87.1	89.0	79.8	88.6	92.1
mAP	79.0	82.2	78.7	83.2	89.0	90.7	82.8	90.5	92.9

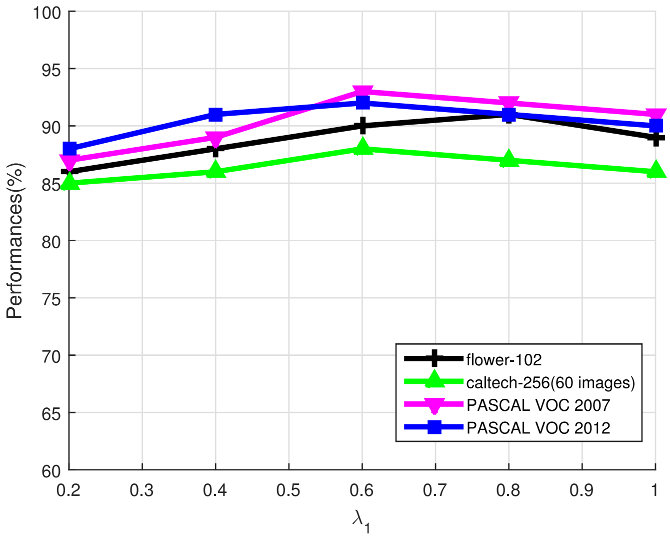

Fig. 2. Influences of λ_1 on (a) the Flower-102 dataset, (b) the Caltech-256 dataset(60 training images), (c) the PASCAL VOC 2007 dataset and (d) the PASCAL VOC 2012 dataset.

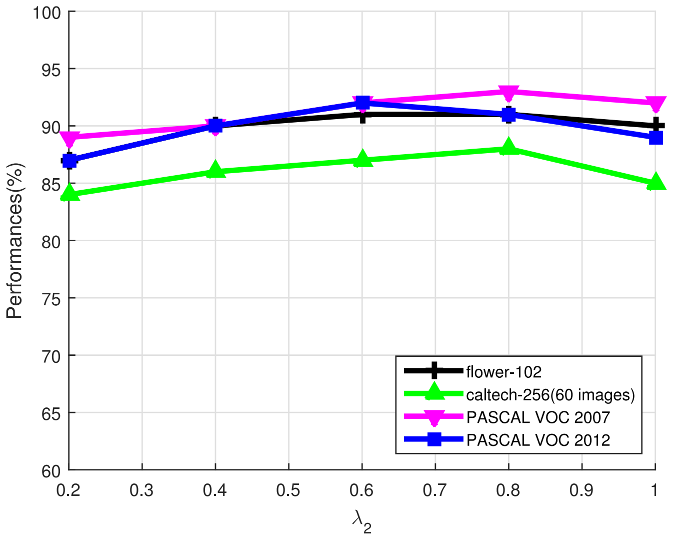

Fig. 3. Influences of λ_2 on (a) the Flower-102 dataset, (b) the Caltech-256 dataset(60 training images), (c) the PASCAL VOC 2007 dataset and (d) the PASCAL VOC 2012 dataset.

F. Parameter Influences

λ_1 and λ_2 are the two parameters which control the influences of the inter-view and intra-view similarity. To show their influences, we plot the performance changes with λ_1 and λ_2 on the Flower-102 dataset, the Caltech-256 dataset (60 training images), the PASCAL VOC 2007 dataset and the PASCAL VOC 2012 dataset in Figs. 2 and 3 respectively. We can see from Figs. 2 and 3 that adding the inter-view and intra-view similarities can improve the classification performances. The performances decrease if we place too much emphasis on the similarity terms by setting λ_1 and λ_2 to 1. We can see from Fig. 2 that setting λ_1 and λ_2 to 0.4$\sim$ 0.8 is a better choice.

λ controls the influence of the L_2 regularization term while σ is the scaling parameter for measuring the visual similarity. We also plot the performance changes with λ and σ jointly on the Flower-102 dataset, the Caltech-256 dataset (60 training images), the PASCAL VOC 2007 dataset and the PASCAL VOC 2012 dataset in Figs. 4 and 5 respectively. We can see from Figs. 4 and 5 that their influences are relatively stable compared with λ_1 and λ_2. We believe this is because λ and σ measure the similarities between images, their influences are consistent as long as the similarities between images are preserved.

G. Wiki Dataset

We also conduct experiments on the Wiki image-text dataset [56] for retrieval in a classification way by sorting the classification results in a descending order. This dataset has 2,866 image-text pairs with each pair having one image and text of 10

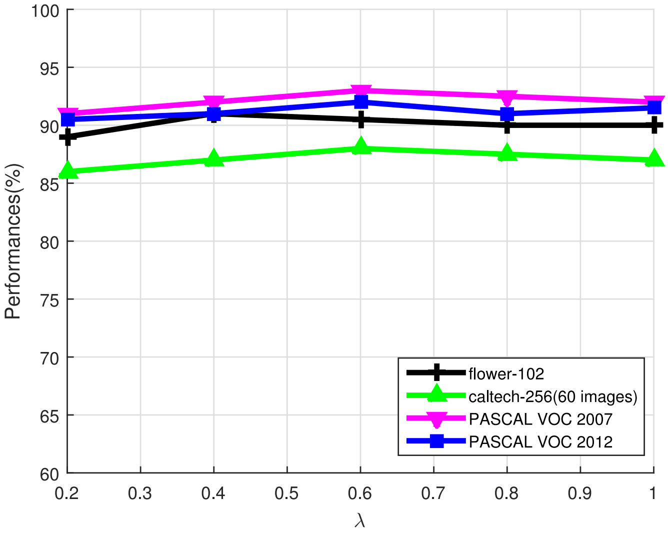

Fig. 4. Influences of λ on (a) the Flower-102 dataset, (b) the Caltech-256 dataset(60 training images), (c) the PASCAL VOC 2007 dataset and (d) the PASCAL VOC 2012 dataset.

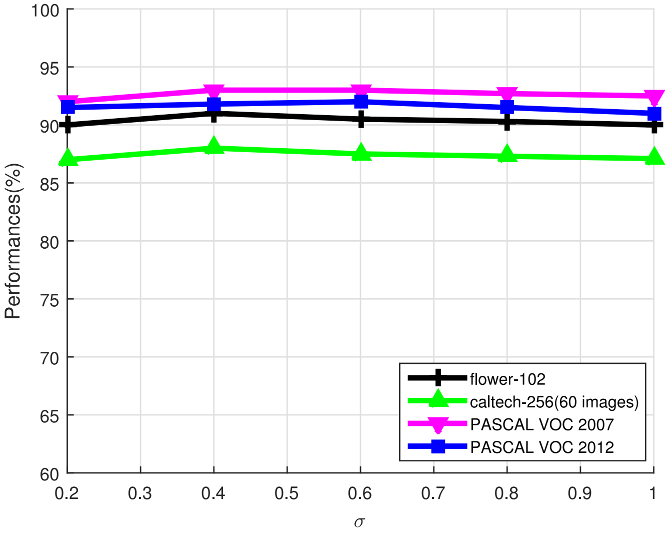

Fig. 5. Influences of σ on (a) the Flower-102 dataset, (b) the Caltech-256 dataset(60 training images), (c) the PASCAL VOC 2007 dataset and (d) the PASCAL VOC 2012 dataset.

semantic classes. We treat the 128-dimensional visual features and the 10-dimensional text features as two different views. The effectiveness of combining of different views has been proven on the Flower-102 dataset, the Caltech-256 dataset, the PASCAL VOC 2007 and 2012 datasets. Hence, we simply extract the convolutional neural network based features (VGG [5]) in the same way as on the four image datasets described above and regard it as the third view. In this way, a total of three views are used for image-text pair retrieval. More views can be used to further improve the performances. We use the same split setup as [56] by using 2,173 pairs for training and 693 pairs for testing. Mean average precision (mAP) is used to evaluate the performances in Table VI. We can have three conclusions from Table VI. First, the proposed method can improve the mAP over other baseline methods. Second, compared with local feature based methods [56]–[60], the usages of convolutional neural network based image representations are more discriminative. Third, MVLS also improves over latent space based methods [57], [60] by jointly learning the transformation and correlations.

TABLE VI
mAP COMPARISONS (%) OF THE PROPOSED METHOD AND OTHER BASELINE METHODS ON THE WIKI DATASET

Methods	mAP
CCA [57]	24.6
SCM [56]	27.7
MvDA [58]	16.2
RE-DNN [59]	35.3
GMMFA [60]	27.8
JFSSL [61]	42.8
DCML [62]	55.4
MVLS (3 views)	62.7

V. CONCLUSION

This paper proposed a multi-view label sharing method for visual representations and classifications. We transferred the visual representations of each view linearly to a shared space. The inter-view similarity and the intra-view similarity were jointly considered with the learning of classifiers. The transformation matrixes and classifiers were obtained by minimizing the summed classification loss along with the inter-view and intra-view similarities. This was achieved by alternatively optimizing for the transformation matrixes and classifier parameters. We conducted image classification experiments on several public datasets and the results proved the usefulness of the proposed method.

REFERENCES

[1] C. Zhang *et al.*, “Image classification by non-negative sparse coding, low-rank and sparse decomposition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1673–1680.
[2] 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.*, 2006, pp. 2169–2178.
[3] 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.*, 2009, pp. 1794–1801.
[4] J. Sanchez, 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, Dec. 2013.
[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arxiv.1409.1556, 2014.
[6] J. Wang *et al.*, “Locality-constrained linear coding for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3360–3367.
[7] C. Zhang, J. Cheng, and Q. Tian, “Incremental codebook adaptation for visual representation and categorization,” *IEEE Trans. Cybern.*, 2017, to be published, doi: 10.1109/TCYB.2017.2726079.
[8] 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.
[9] F. Perronnin and D. Larlus, “Fisher vectors meet neural networks: A hybrid classification architecture,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3743–3752.

[10] C. Zhang et al., "Fine-grained image classification via low-rank sparse coding With general and class-specific codebooks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1550–1559, Jul. 2017.

[11] S. Gao, I. Tsang, and L. 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.

[12] C. Zhang et al., "Image class prediction by joint object, context and background modeling," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2016.2613125.

[13] C. Huang, H. Li, Y. Xie, Q. Wu, and B. Luo, "PBC: Polygon-based classifier for fine-grained categorization," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 673–684, Apr. 2017.

[14] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1245–1256, Jun. 2017, doi: 10.1109/TMM.2017.2648498.

[15] J. Wu et al., "Weak labeled active learning with conditional label dependence for multi-label image classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1156–1169, Jun. 2017, doi:10.1109/TMM.2017.2652065.

[16] A. Puthenputhussery, Q. Liu, and C. Liu, "A sparse representation model using the complete marginal fisher analysis framework and its applications to visual recognition," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1757–1770, Aug. 2017, doi:10.1109/TMM.2017.2685179.

[17] Q. Liu, Y. Sun, C. Wang, T. Liu, and D. Tao, "Elastic net hypergraph learning for image clustering and semi-supervised classification," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 452–463, Jan. 2017.

[18] M. Shao and Y. Fu, "Cross-modality feature learning through generic hierarchical hyperlingual-words," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 2, pp. 451–462, Feb. 2017.

[19] C. Zhang et al., "Beyond explicit codebook generation: Visual representation using implicitly transferred codebooks," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5777–5788, Dec. 2015.

[20] W. Dai, Y. Chen, G. Xue, Q. Yang, and Y. Yu, "Translated learning: Transfer learning across different feature spaces," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 353–360.

[21] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2013.

[22] D. Wang, P. Cui, M. Ou, and W. Zhu, "Learning compact hash codes for multimodal representations using orthogonal deep structure," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1404–1416, Sep. 2015.

[23] C. Chou, H. Chen, and S. Lee, "Multimodal video-to-near-scene annotation," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 354–366, Feb. 2017.

[24] S. Wang and W. Guo, "Sparse multi-graph embedding for multimodal feature representation," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1454–1466, Jul. 2017, doi:10.1109/TMM.2017.2663324.

[25] W. Li, J. Joo, H. Qi, and S. Zhu, "Joint image-text news topic detection and tracking by multimodal topic and-or graph," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 367–381, Feb. 2017.

[26] C. Zhang, Q. Huang, and Q. Tian, "Contextual exemplar classifier based image representation for classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1691–1699, Aug. 2017, doi:10.1109/TCSVT.2016.2527380.

[27] J. Dong et al., "Subcategory-aware object classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 827–834.

[28] X. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3493–3500.

[29] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 1717–1724.

[30] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2160–2173, Nov. 2016.

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

[32] M. Shao, Z. Ding, and Y. Fu, "Sparse low-rank fusion based deep features for missing modality face recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2015, pp. 1–6.

[33] T. Landauer, P. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Process.*, vol. 25, pp. 259–284, 1998.

[34] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. Uncertainty Artif. Intell.*, 1999.

[35] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[36] M. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, 2008, pp. 722–729.

[37] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, USA, CalTech Tech. Rep. 7694, 2007.

[38] "The PASCAL visual object classes." [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/

[39] Z. Ding, M. Shao, and Y. Fu, "Missing modality transfer learning via latent low-rank constraint," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4322–4334, Nov. 2015.

[40] 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.*, 2010, pp. 2313–2319.

[41] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 811–818.

[42] 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. Eur. Conf. Comput. Vis.*, 2012, pp. 794–807.

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

[44] L. Li, H. Su, E. Xing, and Li. Fei-Fei, "ObjectBank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2010, pp. 1378–1386.

[45] K. Sohn, D. Jung, H. Lee, and A. Hero, "Efficient learning of sparse, distributed, convolutional feature representations for object recognition," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2643–2650.

[46] L. Bo, X. Ren, and D. Fox, "Multipath sparse coding using hierarchical matching pursuit," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 660–667.

[47] M. Zeiler and R. Ferugs, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[48] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja, "Low-rank sparse coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 281–288.

[49] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 237–244.

[50] P. Sermanet et al., "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2014.

[51] J. Donahue et al., "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. 31st Int. Conf. Mach. Learn*, 2014.

[52] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014. [Online]. Available: http://dx.doi.org/10.5244/C.28.6.

[53] 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, Sep. 2016.

[54] H. Yang et al., "Exploit bounding box annotations for multi-label object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 280–288.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[56] N. Rasiwasia, J. Pereira, E. Coviello, and G. Doyle, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[57] W. Yang, D. Yi, Z. Lei, J. Sang, and S. Z. Li, "2D-3D face matching using CCA," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2008, pp. 1–6.

[58] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 808–821.

[59] C. Wang, H. Yang, and C. Meinel, "A deep semantic framework for multimodal representation learning," *Multimedia Tools Appl.*, vol. 75, pp. 9255–9276, 2016.

[60] A. Sharma, A. Kumar, H. Daume, and D. Jacobs, "Generalized multi-view analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2160–2167.

[61] K. Wang, R. He, L. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.
[62] V. Liong, J. Lu, Y. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1234–1244, Jun. 2017.
[63] C. Zhang, J. Sang, G. Zhu, and Q. Tian, "Bundled local features for image representation," *IEEE Trans. Circuits Syst. Video Technol.*, 2017, to be published, doi:10.1109/TCSVT.2017.2694060.
[64] C. Zhang, J. Cheng, and Q. Tian, "Incremental codebook adaptation for visual representation and categorization," *IEEE Trans. Cybern.*, 2017, to be published, doi:10.1109/TCYB.2017.2726079.
[65] C. Zhang, J. Cheng, and Q. Tian, "Image-specific classification with local and global discriminations," *IEEE Trans. Neural Netw. Learn. Syst.*, 2017, to be published, doi: 10.1109/TNNLS.2017.2748952.
[66] C. Zhang *et al.*, "Image classification by non-negative sparse coding, correlation constrained low-rank and sparse decomposition," in *Proc. Comput. Vis. Image Understanding*, vol. 123, Jun. 2014, pp. 14–22.
[67] C. Zhang *et al.* "Image classification using spatial pyramid robust sparse coding," *Pattern Recognit. Lett.*, vol. 34, no. 9, pp. 1046–1052, 2013.

Chunjie Zhang received the B.E. degree from 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 worked as an Engineer in the Henan Electric Power Research Institute during 2011–2012. He was a Postdoc at the School of Computer and Control Engineering, University of Chinese Academy of Sciences, where he became an Assistant Professor. He is currently an Assistant Professor in 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 was a Postdoctor in Nokia Research Center, China. Since 2006, he has been with the National Laboratory of Pattern Recognition. 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 is currently a Full Professor in the Department of Computer Science, the University of Texas at San Antonio (UTSA), San Antonio, TX, USA. He was a tenured Associate Professor from 2008 to 2012 and a tenure-track Assistant Professor from 2002 to 2008. During 2008–2009, he took one-year Faculty Leave at Microsoft Research Asia as a Lead Researcher in the Media Computing Group. He has published more than 390 refereed journal and conference papers. His research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics. Dr. Tian was the coauthor of a Best Paper in ACM ICMR 2015, a Best Paper in PCM 2013, a Best Paper in MMM 2013, a Best Paper in ACM ICIMCS 2012, a Top 10% Paper Award in MMSP 2011, and a Best Student Paper in ICASSP 2006, and the coauthor of a Best Student Paper Candidate in ICME 2015, and a Best Paper Candidate in PCM 2007. His research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, CIAS, Akiira Media Systems, HP, Blippar, and UTSA. He received the 2017 UTSA President's Distinguished Award for Research Achievement, the 2016 UTSA Innovation Award, the 2014 Research Achievement Awards from College of Science, UTSA, the 2010 Google Faculty Award, and the 2010 ACM Service Award. He is the 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*, *Multimedia System Journal*, and on the Editorial Board of the *Journal of Multimedia* and the *Journal of Machine Vision and Applications*. He is the Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the *Journal of Computer Vision and Image Understanding*, etc.