Image-Specific Classification With Local and Global Discriminations

Chunjie Zhang, Jian Cheng, Changsheng Li, and Qi Tian

***Abstract*—Most image classification methods try to learn classifiers for each class using training images alone. Due to the inter-class and intra-class variations, it would be more effective to take the testing images into consideration for classifier learning. In this brief, we proposes a novel image-specific classification method by combing the local and global discriminations of training images. We adaptively train classifier for each testing image instead of generating classifiers for each class with training images alone. For each testing image, we first select its *k* nearest neighbors in the training set with the corresponding labels for local classifier training. This helps to model the distinctive characters of each testing image. Besides, we also use all the training images for global discrimination modeling. The local and global discriminations are combined for final classification. In this way, we could not only model the specific character of each testing image but also avoid the local optimum by jointly considering all the training images. To evaluate the usefulness of the proposed image-specific classification with local and global discriminations method (ISC-LG), we conduct image classification experiments on several public image datasets. The superior performances over other baseline methods prove the effectiveness of the proposed ISC-LG method.**

***Index Terms*—Image-specific classification, global information, local information, object categorization**

I. Introduction

Image classification refers to automatically predict the classes of images based on the contents. It is often very hard to accurately classify images due to the discrepancies between visual representations and human perception. Local features are widely used with the bag-of-visual-words (BoW) scheme [1] for classification with good performances. Inspired by the BoW model, many works [2-8] have been done. Instead of using human designed local features, automatically learning from image pixels with deep convolutional neural networks (CNN) [9-12] becomes popular in recent years which outperforms traditional local feature based methods on many image datasets. The deep network can capture various correlations of objects and help to represent images more discriminatively.

However, both the BoW and CNN based methods try to learn a number of classifiers for image class prediction. Usually, this is done in the one-vs-all way by separating images of one class from other classes. Although simple and effective, this scheme does not explicitly consider the inter and intra class variances of images. Besides, the discrepancies between visual features and human perception are often encountered when classifying images. Take 'jaguar' for example, it can refer to an animal or a car. Images of car are visually different from images of animal. If we try to learn a classifier to separate the 'jaguar' images from other images, the classification accuracy would be far from satisfactory, especially when there are similar images of different classes. We can try to avoid this problem by training a number of classifiers [13-20], however, how to determine the proper number of classifiers is still an open problem. Another way to alleviate this problem is by using auxiliary information [21-26]. However, the generalization power of these methods is limited. Besides, the independent and identically distributed (iid) assumption cannot always be satisfied. Fortunately, locally nearby samples are probably similar. The exploration of local information can help to alleviate the variations of images to some extent [27-32]. However, it may be biased to only use the local information. It would be more effective to combine the local and global information jointly.

Many pre-learned classifiers are trained to separate training images apart without considering the distinctive properties of testing images. In other words, the classifiers are fixed and do not change when applied to different testing images. We believe this is sub-optimal for two reasons. First, the training images may not be able to fully represent the semantics well. Hence, the learned classifiers are biased. Second, different testing images may have varied characters that should be treated differently. One way to solve this problem is to classify images directly without classifier training [33-35]. However, this strategy costs a lot of computational power. Besides, their performances are inferior compared with classifier training based methods. Moreover, local features are often used by these methods which cannot be easily combined with more discriminative representation strategies (e.g. fisher vector [36-37] and deep learning based methods [38-40]). Hence, it would be more

Chunjie Zhang is with Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China. He is also with University of Chinese Academy of Sciences, 100049, Beijing, China. E-mail: chunjie.zhang@ia.ac.cn.

Jian Cheng is with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O.Box 2728, Beijing, China. He is also with University of Chinese Academy of Sciences, 100049, Beijing, China. E-mail: jcheng@nlpr.ia.ac.cn.

Changsheng Li is with School of Computer Science and Engineering, University of Electronic Science and Technology of China. E-mail: changsheng_li_507@hotmail.com.

Qi Tian is with Department of Computer Sciences, University of Texas at San Antonio. TX, 78249, U.S.A. E-mail: qitian@cs.utsa.edu.

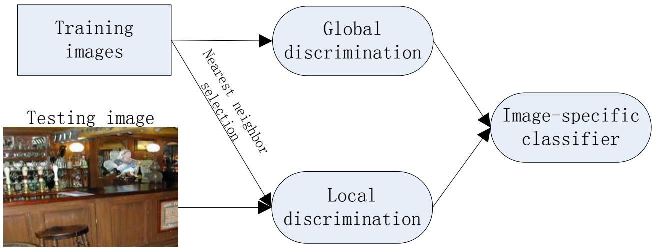


Fig. 1. Flowchart of the proposed global and local discrimination modeling for image-specific classification method.

effective if we can generate image-specific classifiers for testing images and improve the discriminative power by exploring the local and global discriminations.

To solve the problems mentioned above, in this paper, we propose a novel image-specific classification method by considering the local and global discriminations. Instead of learning the classifiers per class using training samples alone, we try to learn the classifier for each testing image. This is achieved by choosing the *k* nearest training samples for one testing image and learn the local classifiers accordingly. Besides, to make use of all the training images and avoid local optimum, we use training images to learn the global classifiers jointly with the local classifiers. In this way, we are able to predict the classes of the testing images by learning image-specific classifiers. We evaluate the effectiveness of the proposed image-specific classification with local and global discrimination method on several public image datasets. The superior performances over other baseline methods prove the usefulness of the proposed method. Figure 1 shows the flowchart of the proposed method.

The main contributions of this paper lie in three aspects.

- First, instead of only using training images, the proposed method learns image-specific classifier for each testing image. The learned classifier can adapt to the testing image for better classification.
- Second, we learn the local classifier using *k* nearest neighbors of each testing image and the global classifier with all the training images. This helps to combine the discriminative power of local and global information simultaneously.
- Third, the proposed method is able to outperform many baseline methods. It can also be combined with other image representation methods (e.g. convolutional neural network) which further improves the classification performances.

II. Related Work

The BoW model [1] was widely used for image classification. Local features were first quantized to form the histogram representation of images and classifiers were then trained to predict image classes. Many BoW based methods were proposed by researchers [2-8]. In recent years, convolutional neural networks (CNN) [11] was used to extract image representation from image pixels. Krizhevsky *et al.* [9] proposed the AlexNet by classifying images with deep convolutional neural networks. Inspired by this, Simonyan and Zisserman [10] explored the discriminative power of very deep convolutional networks and found that adding layers properly can boost the performances. To combine the advantages of fisher vectors and neural networks, Perronnin and Larlus [12] proposed a hybrid architecture for image classification and improved the performance.

To cope with class variations, researchers [13-20] tried to train a number of classifiers instead of only using one-vs-all classifier. Sivic *et al.* [13] tried to learn person specific classifiers while Sungwoong *et al.* [14] used task-specific partition technique. Zhang *et al.* [15] modeled the general and class-specific properties of images. Instead of learning class-level classifiers, the use of exemplar classifiers [16-19] also became popular. Weinshapp *et al.* [20] also studied the problem when the learned general and specific classifiers were not consistent.

Another way to alleviate this problem was by using auxiliary information [21-26]. Guillaumin *et al.* [21] leveraged semi-supervised learning technique for image classification while Fernando *et al.* [22] used subspace alignment for domain adaptation without supervision. Li *et al.* [23] harvested the images from Internet. Chen *et al.* [24] used the contextual information of object detection and classification. Liu *et al.* [25] proposed to use the hypergraph to model the high-order relationships for better classification. Fu *et al.* [26] tried to classify images with few training images by transductive multi-view learning. The exploring of neighbor information [27-32] had been proven very effective both for improving the performance and for reducing the computational cost.

The direct classification of images without training classifiers had also been studied [33-35]. Boiman *et al.* [33] used the nearest-neighbor distances for direct classification. Sthitsos *et al.* [34] proposed to use a cascade approximate measurement method for efficient nearest neighbor classification. Mccann and Lowe [35] used local information to speed up the computation of nearest neighbors. Other more effective methods were also proposed [36-40]. Deep convolutional networks were also used for visual tracking [38] and feature learning [39]. It was also used for event recognition by fusing multiple semantic cues by Zhang *et al.* [40].

A number of datasets [41-44] were also collected for classification tasks. Many algorithms [45-55] had been proposed to improve the classification accuracy. There were two main differences between [56] and the proposed method. First, the proposed method targeted the classification task while [56] tried to solve the clustering problem. Second, instead of only using the similarities of images, the proposed method also leveraged the class information for discriminative classifier training. Zhang *et al.* [57] proposed a novel probabilistic topology discovering model to predict the class of scene images. To classify the aerial images, Zhang *et al.* [58] also tried to discover discriminative graphlets which greatly improved the performances. More challenging dataset was also collected [59] which was widely used for evaluating the performances of various convolutional neural network based methods [60-62]. Other efficient methods were also adopted for classification [63-66].

III. Image-Specific Classification Using Local And Global Discriminations

In this section, we give the details of the proposed image-specific classification method using local and global discriminations.

A. Image Representation

We use the sparse coding technique [3] and the CNN technique [9] to represent images in this paper as they have been widely used for image classification in recent years. Note that other image representation methods can also be used as the proposed method does not rely on image representations but focuses on the learning of classifiers.

Let $\boldsymbol{h_i}, i = 1, ..., M$ be the M local features, the sparse coding technique tries to minimize the reconstruction error with L_1 constraint as:

$$\boldsymbol{\beta_i} = argmin_{\boldsymbol{\beta_i}} \|\boldsymbol{h_i} - \boldsymbol{\beta_i^T D}\|^2 + a\|\boldsymbol{\beta_i}\|_1, i = 1, ..., M \quad (1)$$

where $\boldsymbol{D}$ is the codebook, a is the weighting parameter, $\boldsymbol{\beta_i}$ is the encoding parameter. The sparse coding technique choose the max absolute values of each dimension of encoded parameters $\boldsymbol{\beta_i}$ within an image region as:

$$\boldsymbol{l_j} = max(\boldsymbol{\beta_{1,j}}, \boldsymbol{\beta_{2,j}}, ..., \boldsymbol{\beta_{M_1,j}}) \quad (2)$$

where $\boldsymbol{l}$ is the image region's representation, $\boldsymbol{\beta}_{i,j}$ is the j-th element of $\boldsymbol{\beta_i}$. $\boldsymbol{l_j}$ is the j-th element of $\boldsymbol{l}$, M_1 is the number of local features within this region. The final image representation is obtained by concatenating all the region's representation as $\boldsymbol{x_{sc}} = [\boldsymbol{l^1}; ...; \boldsymbol{l^{M_2}}]$, where M_2 is the number of regions. We use the CNN based image representation [9] $\boldsymbol{x_{cnn}}$ and combine it with the sparse coding based representation to form the final image representation as $\boldsymbol{x} = [\boldsymbol{x_{sc}}; \boldsymbol{x_{cnn}}]$. We then train image-specific classifiers using $\boldsymbol{x}$.

B. Image-Specific Local Discrimination Modeling

It is a challenging problem to correctly classify images based on their visual contents. Due to the varieties of images, visually similar images may belong to different classes while images of the same class may have differently visual appearances. It is hard to learn reliable classifiers which can separate images well, especially for the classes with cluttered objects and background. Besides, the state-of-the-art classifiers are often trained to separate images of the same class from other classes, leaving the testing images unconsidered. We believe if we can learn classifiers based on the contents of testing images, it would be more likely to classify images better than simply using training images.

Our motivation is based on the observation that locally nearby images tend to be semantically similar, as observed by many researchers [27-32]. However, we do not only rely on the neighborhood information but go one step further by training discriminative local classifiers. For each testing image, this is done by firstly choosing its k nearest neighbors of training images and then use them to learn the local classifier.

Formally, let $(\boldsymbol{x_n}, y_n), n = 1, ..., N$ be the N training images, $y_n \in 1, 2, ..., C$ are the corresponding labels of images with C is the number of classes. For each testing image $\boldsymbol{x_t}$, we first calculate its similarities with all the training images and choose the top K most similar images. We use $s(\boldsymbol{i}, \boldsymbol{j}) = exp^{-\|\boldsymbol{i}-\boldsymbol{j}\|^2}$ as the similarity measurement in this paper. Let $\{\boldsymbol{x_k}, y_k\}, k = 1, ..., K$ be the K nearest neighbors of training images, we learn the corresponding local classifier $f(\boldsymbol{x_t})$ by minimizing the summed loss over the K training images as:

$$f(.) = argmin_{f(.)} \sum_{k=1}^{K} \ell_{loc}(y_k, f(\boldsymbol{x_k})) + \gamma_1 \mathcal{T}_{loc}(f(.)) \quad (3)$$

where ℓ_{loc} is the local loss function, $\mathcal{T}_{loc}(f(.))$ is the local regularization term. γ_1 is the local weighting parameter.

C. Global Discrimination Modeling

We can directly use the locally trained classifier for image class prediction. However, only using the local information may bias the final classification. Besides, a testing image may be dissimilar with any training images which makes the learned local classifier unreliable. The discriminative power of other training images should also be considered to boost the discriminative power.

Let $g(\boldsymbol{x_t})$ be the global classifier which is to be learned using all the training images. Similar as the local classifier, $g(\boldsymbol{x_t})$ can also be learned by minimizing the training loss as:

$$g(.) = argmin_{g(.)} \sum_{i=1}^{N} \ell_{glob}(y_i, g(\boldsymbol{x_i})) + \gamma_2 \mathcal{T}_{glob}(g(.)) \quad (4)$$

where ℓ_{glob} is the global loss function, $\mathcal{T}_{glob}(g(.))$ is the global regularization term. γ_2 is the global weighting parameter.

We can use different loss functions ℓ_{loc} and ℓ_{glob} to incorporate different information for classification. Similarly, the regularization term $\mathcal{T}_{loc}(f(.))$ and $\mathcal{T}_{glob}(g(.))$ can also be chosen differently. In this way, we can combine various types of information for efficient classification.

D. Local And Global Discrimination Combination

Instead of using local classifier or global classifier individually, we combine them together for classification. The local classifier models the neighbor information of testing image while the global classifier models all the training images. We use linear combination $\alpha f(.) + (1 - \alpha)g(.)$ to predict the testing image's class. α is the parameter which balances the relative influences of local and global classifiers. We jointly learn local classifier and global classifier by minimizing the summed loss with regularization as:

$$\begin{aligned} f, g =& argmin_{f,g} \sum_{k=1}^{K} \ell_{loc}(y_k, \alpha f(\boldsymbol{x_k}) + (1-\alpha) g(\boldsymbol{x_k})) \\ & + \gamma_1 \mathcal{T}_{loc}(f(.)) + \beta(\sum_{i=1}^{N} \ell_{glob}(y_i, g(\boldsymbol{x_i})) \\ & + \gamma_2 \mathcal{T}_{glob}(g(.))) \end{aligned} \tag{5}$$

where β is the parameter which controls the influences of local and global training images. The traditional class-level classifier learning strategy can be viewed as a special case of the proposed method where only the global classifier is learned. Besides, if we set K to 1 and β to zero, Eq. 5 would degenerate to the exemplar classifier based scheme. The proposed ISC-LG method is able to combine the discriminative power of both local and global information for better classification.

We use linear classifiers

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{w_{loc}^T x} + b_{loc} \\ g(\boldsymbol{x}) &= \boldsymbol{w_{glob}^T x} + b_{glob} \end{aligned} \tag{6}$$

where $\boldsymbol{w_{loc}}, b_{loc}$ are the parameters of local classifier and $\boldsymbol{w_{glob}}, b_{glob}$ are the parameters of global classifier respectively. After the parameters $\boldsymbol{w_{loc}}, b_{loc}, \boldsymbol{w_{glob}}, b_{glob}$ are learned, we can predict the class of testing image $\boldsymbol{x_t}$ accordingly as:

$$\widehat{y_t} = \alpha f(\boldsymbol{x_t}) + (1-\alpha) g(\boldsymbol{x_t}) \tag{7}$$

Note that the proposed method can also use other types of classifiers. The hinge loss is often used for the classification tasks, however, it is not smooth. To alleviate this problem, we use the quadric hinge loss for ℓ_{loc} and ℓ_{glob} in this paper as:

$$\begin{aligned} \ell_{loc}(y_{loc}, y) &= \frac{1}{2} max(0, 1 - y_{loc} \times y)^2 \\ \ell_{glob}(y_{loc}, y) &= \frac{1}{2} max(0, 1 - y_{glob} \times y)^2 \end{aligned} \tag{8}$$

Besides, we use L_2 norm for the regularization of $\mathcal{T}_{loc}$ and $\mathcal{T}_{glob}$ as:

$$\begin{aligned} \mathcal{T}_{loc}(f(.)) &= \|\boldsymbol{w_{loc}}\|^2 + b_{loc}^2 \\ \mathcal{T}_{glob}(g(.)) &= \|\boldsymbol{w_{glob}}\|^2 + b_{glob}^2 \end{aligned} \tag{9}$$

In this way, we can rewrite Eq.5 as:

$$\begin{aligned} & [\boldsymbol{w_{loc}}, b_{loc}, \boldsymbol{w_{glob}}, b_{glob}] = argmin_{[\boldsymbol{w_{loc}}, b_{loc}, \boldsymbol{w_{glob}}, b_{glob}]} \\ & \sum_{k=1}^{K} \frac{1}{2} max(0, 1 - \alpha y_k \times (\boldsymbol{w_{loc}^T x_k} + b_{loc}) - \\ & (1-\alpha)(\boldsymbol{w_{glob}^T x_k} + b_{glob}))^2 + \gamma_1(\|\boldsymbol{w_{loc}}\|^2 + b_{loc}^2) \\ & + \beta(\sum_{i=1}^{N} \frac{1}{2} max(0, 1 - y_i \times (\boldsymbol{w_{glob}^T x_i} + b_{glob}))^2 \\ & + \gamma_2(\|\boldsymbol{w_{glob}}\|^2 + b_{glob}^2)) \end{aligned} \tag{10}$$

We alternatively optimize for the local classifier and the global classifier by fixing the other. When learning the local classifier while fixing the global classifier, let $\xi_{loc,k} = 1 - (1-\alpha)(\boldsymbol{w_{glob}^T x_k} + b_{glob})$, $\overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}} = [\boldsymbol{w_{loc}}; b_{loc}]$, Eq. 10 equals to the following problem:

$$\begin{aligned} & \overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}} = argmin_{\overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}}} \gamma_1 \|\overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}}\|^2 + \\ & \sum_{k=1}^{K} \frac{1}{2} max(0, \xi_{loc,k} - \alpha y_k \times \overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}}{}^T [\boldsymbol{x_k}; 1])^2 \end{aligned} \tag{11}$$

This problem can be solved with the gradient descent method. Let

$$\begin{aligned} & \varphi = \gamma_1 \|\overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}}\|^2 + \\ & \sum_{k=1}^{K} \frac{1}{2} max(0, \xi_{loc,k} - \alpha y_k \times \overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}}{}^T [\boldsymbol{x_k}; 1])^2 \end{aligned} \tag{12}$$

We can have:

$$\begin{aligned} & \frac{\partial \varphi}{\partial \overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}}} = 2\gamma_1 \overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}} - \\ & \sum_{k=1}^{K} \mu_{1,k} \alpha y_k max(0, \xi_{loc,k} - \alpha y_k \times \overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}}{}^T [\boldsymbol{x_k}; 1])[\boldsymbol{x_k}; 1] \end{aligned} \tag{13}$$

with $\mu_{1,k} = 1$, if $\xi_{loc,k} - \alpha y_k \times \overrightarrow{\boldsymbol{w}}_{\boldsymbol{loc}}{}^T [\boldsymbol{x_k}; 1] > 0$, otherwise, $\mu_{1,k} = 0$.

After the local classifier is learned, we fixed it to learn the global classifier. Let $\xi_{glob,k} = 1 - \alpha y_k \times (\boldsymbol{w_{loc}^T x_k} + b_{loc})$, $\overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}} = [\boldsymbol{w_{glob}}; b_{glob}]$, Eq.10 equals to the following problem:

$$\begin{aligned} & \overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}} = argmin_{\overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}} \\ & \sum_{k=1}^{K} \frac{1}{2} max(0, \xi_{glob,k} - (1-\alpha) \overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}{}^T [\boldsymbol{x_k}; 1])^2 \\ & + \beta(\sum_{i=1}^{N} \frac{1}{2} max(0, 1 - y_i \times \overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}{}^T [\boldsymbol{x_i}; 1])^2 \\ & + \gamma_2 \|\overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}\|^2 \end{aligned} \tag{14}$$

Let

$$\begin{aligned} \phi =& \sum_{k=1}^{K} \frac{1}{2} max(0, \xi_{glob,k} - (1-\alpha) \overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}{}^T [\boldsymbol{x_k}; 1])^2 \\ & + \beta(\sum_{i=1}^{N} \frac{1}{2} max(0, 1 - y_i \times \overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}{}^T [\boldsymbol{x_i}; 1])^2 \\ & + \gamma_2 \|\overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}\|^2 \end{aligned} \tag{15}$$

The derivative of ϕ over $\overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}$ can then be calculated as:

$$\begin{aligned} & \frac{\partial \phi}{\partial \overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}} = 2\gamma_2 \overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}} - \\ & \sum_{k=1}^{K} \mu_{2,k} max(0, \xi_{glob,k} - (1-\alpha) \overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}{}^T [\boldsymbol{x_k}; 1]) \\ & \times (1-\alpha)[\boldsymbol{x_k}; 1] \\ & - \beta \sum_{i=1}^{N} \mu_{3,i} max(0, 1 - y_i \times \overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}{}^T [\boldsymbol{x_i}; 1]) y_i [\boldsymbol{x_i}; 1] \end{aligned} \tag{16}$$

with $\mu_{2,k} = 1$ if $\xi_{glob,k} - (1-\alpha)\overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}^{T}[\boldsymbol{x_k};1] > 0$, otherwise $\mu_{2,k} = 0$. $\mu_{3,i} = 1$ if $1 - y_i \times \overrightarrow{\boldsymbol{w}}_{\boldsymbol{glob}}^{T}[\boldsymbol{x_i};1] > 0$, otherwise $\mu_{3,i} = 0$.

We alternatively optimize for the local classifier/global classifier while keeping the other fixed for a number of times or the decrement of the objective value of Eq.10 falls below a pre-defined threshold θ. After the local and global classifiers are learned, we can use them to jointly predict the testing image's class using Eq.7.

IV. Experiments

To evaluate the effectiveness of the proposed ISC-LG method, we conduct image classification experiments on several public image datasets: the Caltech-256 dataset [41], the Flower-17 dataset [42], the Flower-102 dataset [43] and the PASCAL VOC 2007 dataset [44]. The classification performance is evaluated using the classification rates for the Caltech-256 dataset, the Flower-17 dataset and the Flower-102 dataset. For the PASCAL VOC 2007 dataset, the mean average precision (mAP) is used for performance comparison. We compare with the results reported by other methods directly using the same experimental setup instead of re-implementing them for fair comparison.

A. Experimental Setup

We combine the BoW scheme and the CNN scheme for image representation. For the BoW based image representation, we first densely extract SIFT features with multi-scales. The minimum scale is set to 16×16 pixels with an overlap of 6 pixels. Spatial pyramid matching (SPM) [2] is also used with three pyramids ($L = 0, 1, 2$) to make use of the location information of local features. We use the sparse coding code provided by Yang *et al.* [3] and set the codebook size to 1,000. Color-SIFT features [45] are also extracted. We generate the BoW representations for each type of local features and then concatenate them together. As to the CNN base image representation, we use the code provided by [46] with seven layers networks. Each image is represented with a vector of 4,096 dimensions. The BoW and CNN based image representations are then concatenated for final image representations which are then used for image-specific classifier training. γ_1 is set to be equal with γ_2. The maximum iteration number in Algorithm 1 is set to 40.

The dataset is divided into the training set and the testing set without overlap. For each testing image, the proposed method tries to learn one classifier by combining the local and global information. We only use the visual representation of the testing image without using the class information. Only the labels of training images are used. The traditional scheme first learns the classifier using training images and then predicts the class information of each testing image. The proposed method uses the same information as traditional schemes but differs from the learning of classifiers. The experimental comparison is fair compared with traditional schemes.

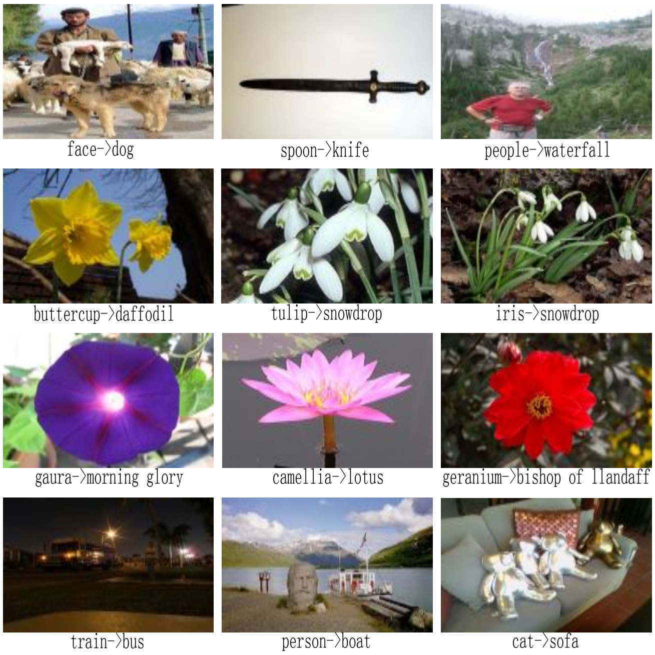

face->dog
spoon->knife
people->waterfall
buttercup->daffodil
tulip->snowdrop
iris->snowdrop
gaura->morning glory
camellia->lotus
geranium->bishop of llandaff
train->bus
person->boat
cat->sofa

Fig. 2. Example images that are misclassified by the global classifier but corrected by the proposed method. From top to bottom (the Caltech-256 dataset, the Flower-17 dataset, the Flower-102 dataset and the PASCAL VOC 2007 dataset).

We also give some example images that are misclassified by the global classifier but are predicted with the right categories using the proposed ISC-LG in Figure 2 on the four datasets. By generating image-specific classifiers, the proposed method is able to correctly predict the classes of images. Due to class variations, globally trained classifier may not be able to separate images well. For example, face and people are often misclassified as other classes (e.g. dog and segway) because they often appear on the same image. However, we can separate them apart by combining the local information.

B. The Caltech-256 Dataset

This dataset has 29,780 images of 256 classes with 80 images for each class. The Caltech-256 dataset is widely used for performance evaluations of the classification algorithms. For fair comparison, we follow the procedures of other methods by randomly selecting 15, 30, 45 and 60 images per class for training and use the rest images for testing. The random selection process is repeated for ten times with the mean and standard derivation of classification rates are used for evaluation. We give the classification results on the Caltech-256 dataset in Table 1.

The proposed ISC-LG is able to outperform many baseline methods. First, compared with traditional one-vs-all methods [3, 5-7, 15, 47], ISC-LG can consider the local information of testing images for image-specific modeling. Besides, the use of more discriminative local features [47] can help to improve the performance over sparse coding based strategy. However, ISC-LG is still able to outperform [47]. Second, compared with exemplar classifier

TABLE I
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
KC [5]	–	27.17 ± 0.46	–	–
LLC [6]	27.74 ± 0.32	32.07 ± 0.24	35.09 ± 0.44	37.79 ± 0.42
LScSPM [7]	30.00 ± 0.14	35.74 ± 0.10	38.47 ± 0.51	40.32 ± 0.32
LR-GCC-FV [15]	41.39 ± 0.36	49.13 ± 0.32	–	–
WSR-EC [18]	35.28 ± 0.65	42.01 ± 0.47	45.82 ± 0.54	–
ObjectBank [23]	– 39.00	–	–	–
NBNN [33]	35.20	42.80	–	–
KSPM [41]	–	34.10	–	–
FV [47]	38.50 ± 0.20	47.40 ± 0.10	52.10 ± 0.40	54.80 ± 0.40
PM [57]	38.40	43.40	47.90	49.80
ISC-LG	43.14 ± 0.67	50.62 ± 0.53	53.27 ± 0.56	55.76 ± 0.48

TABLE II
PERFORMANCE COMPARISON ON THE FLOWER-17 DATASET.

Algorithm	Performance
ICT [4]	91.37± 0.72
LR-GCC [15]	91.52± 1.24
WSR-EC [18]	85.26± 0.94
Nilsback [42]	71.76± 1.76
Varma [48]	82.55± 0.34
Xie [49]	87.45± 1.13
KMTJSRC-CG [50]	88.90± 2.30
CSDL[51]	72.65± 1.79
ISC-LG	92.46± 0.88

based methods [18, 23], ISC-LG can also improve classification accuracy by considering the discriminat information of all images. Since exemplar classifier ba method only considers one sample at a time, the classi is often biased. However, by considering the neighborhood information along with the global distribution of samples, we can generate more discriminative classifiers. Third, the discriminatively trained classifiers also outperform [33] which avoids classifier training by directly computing the image-to-class distance. It is more effective to use discriminative classifiers.

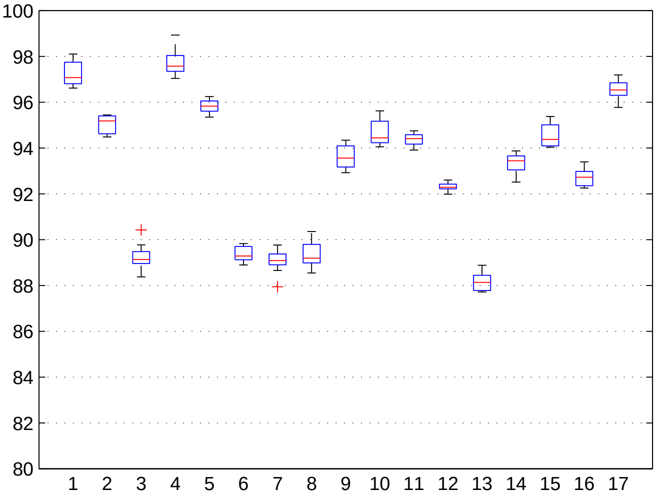


Fig. 3. Boxplot of the performance on the Flower-17 dataset(%). The numbers in the horizon row indicate daffodil, lily valley, snowdrop, iris, bluebell, crocus, tigerlily, fritillary, tulip, daisy, sunflower, dandelion, colts' foot, cowslip, buttercup, windflower, pansy respectively.

C. The Flower-17 Dataset

There are seventeen different types of flowers in the Flower-17 dataset as: *Buttercup, Colts foot, Daffodil, Daisy, Dandelion, Fritillary, Iris, Pansy, Sunflower, Windflower, Snowdrop, Lily valley, Bluebell, Crocus, Tigerlily, Tulip* and *Cowslip*. Images of this dataset are evenly distributed with 80 images for each class. We use the train/validata/test splits provided by [42] with 40/20/20 images respectively for comparison.

The classification results are given in Table 2. The boxplot of the performance of ISC-LG on the Flower-17 data set is also given in Figure 3. We can see that the proposed method again outperforms many baseline methods. ISC-LG jointly combines the local and global information of each testing image by training discriminative classifier. This helps to improve over global classifier [15, 42, 48-51] and exemplar classifier [18] based methods. Besides, ISC-LG also improves over KMTJSRC-CG [50] which classifies images using reconstructed errors without classifier training. Moreover, it is able to beat ICT [4] which leverages auxiliary information from other datasets. These experimental results again prove the usefulness of the proposed ISC-LG method.

D. The Flower-102 Dataset

As an extension of the Flower-17 dataset, the Flower-102 dataset has more classes of images. There are 102 classes of 8,189 images with varied number for each class. The images are divided into 10/10/rest for train/validate/test respectively by Nilsback and Zisserman [43]. We follow this setup and report the classification performance comparisons in Table 3.

We can have similar conclusions as on the Caltech-256 dataset and the Flower-17 dataset. ISC-LG again improves over global classifier [15, 42, 49, 50], local classifier [18] and transfer learning [4] based methods. The image representation is more discriminative by combing detection and segmentation [52, 53], however, ISC-LG still classifies flower images better than [52, 53] due to the combinations of local and global information. Besides, the relative improvements of ISC-LG over baseline methods on the Flower-17 and Flower-102 datasets are larger than that on the Caltech-256 dataset. We believe this is because flower images are visually similar, only using global classifiers cannot separate them apart while the proposed ISC-LG can model this more properly.

TABLE III
PERFORMANCE COMPARISONS ON THE FLOWER-102 DATASET.

Methods	Classification rate
ICT [4]	77.3
LR-GCC [15]	75.7
WSR-EC[18]	82.5
Nilsback [42]	72.8
Xie [49]	86.8
KMTJSRC-CG [50]	74.1
Det+Seg [52]	80.7
TriCoS [53]	85.2
ISC-LG	87.9

E. The PASCAL VOC 2007 Dataset

The 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 images in the PASCAL VOC 2007 dataset are relatively more difficult to classify than the other three datasets. This is not only because there are cluttered objects on each image but also because the inter-class variations are larger. Besides, there are often multiple classes of objects on one image. There are more than 10,000 images with pre-defined division of train/validate/test splits [44]. The validation set is first used to tune the parameters and then combined with the training set for jointly learning of classifiers.

We give the average precision comparisons of the proposed ISC-LG method with other baseline methods in Table 4. ISC-LG outperforms both local feature based methods [6, 44, 47, 54] and CNN [55] based method which again shows the effectiveness of the proposed method. On analyzing the per-class performances, we can have three conclusions. First, ISC-LG is able to outperform other baseline methods on most of the classes. Second, the relative improvements of ISC-LG are larger for non-rigid classes (e.g. cat and dog) than rigid classes (e.g. airplane and bus). This is because non-rigid classes have varied objects whose visual representations are more diverse than rigid objects. This degenerates the discriminative power of global classifiers while the proposed ISC-LG method can model the diversity more effectively. Third, image-specific classifier helps to model the distinctive characters of each image better than general classifier.

TABLE IV
AVERAGE PRECISION COMPARISONS ON THE PASCAL VOC 07 DATASET.

class	LLC [6]	Best07[44]	FV [47]	GS-MKL[54]	DeCAF[55]	ISC-LG
airplane	74.8	77.5	80.0	79.4	87.4	89.1
bicycle	65.2	63.6	67.4	62.4	79.3	81.5
bird	50.7	56.1	51.9	58.5	84.1	85.9
boat	70.9	71.9	70.9	70.2	78.4	81.4
bottle	28.7	33.1	30.8	46.6	42.3	46.7
bus	68.8	60.6	72.2	62.3	73.7	73.9
car	78.5	78.0	79.9	75.6	83.7	85.2
cat	61.7	58.8	61.4	54.9	83.7	83.8
chair	54.3	53.5	56.0	63.8	54.3	60.1
cow	48.6	42.6	49.6	40.7	61.9	61.5
table	51.8	54.9	58.4	58.3	70.2	70.3
dog	44.1	45.8	44.8	51.6	79.5	79.4
horse	76.6	77.5	78.8	79.2	85.3	86.3
motor	66.9	64.0	70.8	68.1	77.2	78.7
person	83.5	85.9	85.0	87.1	90.5	91.0
plant	30.8	36.3	31.7	49.5	51.1	52.9
sheep	44.6	44.7	51.0	48.8	73.8	75.8
sofa	53.4	50.9	56.4	56.4	57.0	57.6
train	78.2	79.2	80.2	75.9	86.4	86.8
tv	53.5	53.2	57.5	54.4	68.0	70.4
mAP	59.3	59.4	61.7	62.2	73.4	74.9

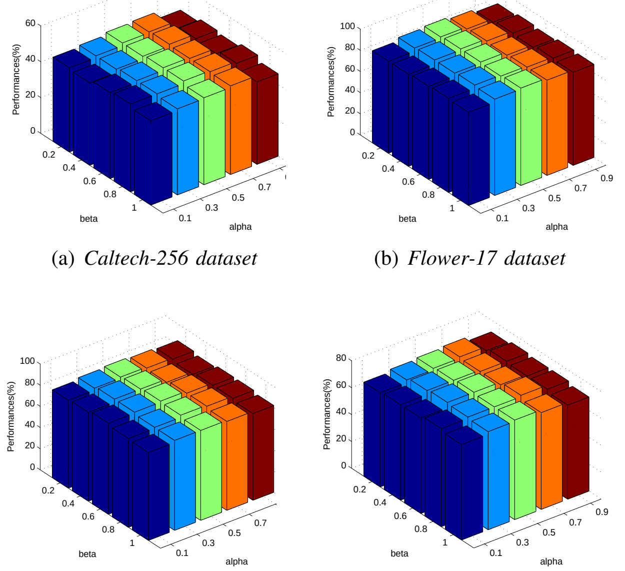
F. Parameter Influences

To show the joint influences of α and β, we plot the performance changes with α and β in Figure 4 on the Caltech-256 dataset, the Flower-17 dataset, the Flower-102 dataset and the PASCAL VOC 2007 dataset. We can see that the relative weights of local and global classifiers have varied influences on the final classification performances. Combining the local discrimination information for image-specific classifier training helps to improve the performances over global classifiers. Besides, during the training process, the optimal β also varies from different datasets. For the flower datasets, the images are more visually similar than images of the Caltech-256 dataset. Hence, more emphasis should be placed on the local discrimination. Moreover, the performance decreases with the increment of classes as more classifiers are needed.

γ_1 and γ_2 control the influences of the regularization terms. In this paper, γ_1 is set to be equal with γ_2. We plot the influences of γ_1/γ_2 on the Caltech-256 dataset, the Flower-17 dataset, the Flower-102 dataset and the PASCAL VOC 2007 dataset in Figure 5. We can see from Figure 5 that the performances increase with γ_1/γ_2 at first. However, if γ_1/γ_2 are too large, the performance would decrease as we pay too much attention to the regularization terms. We can see from Figure 7 that setting γ_1/γ_2 as 5 to 10 can achieve satisfactory performances. K is another parameter which influences the classification performances. We also give the performance changes with K in Figure 5. The performances are less satisfactory with a small K. The classification accuracies increase with K. However, if we set the K too large, the performances also decrease.

G. The ILSVRC-2012 ImageNet dataset

We also evaluate the proposed method on the ILSVRC-2012 ImageNet dataset [59]. There are 1,000 classes of images which are splited into three sets (1.3M training images, 50K validation images and 100K testing images). The top-1 and top-5 error are used to evaluated the classification performance. We follow the experimental setup as [60] and use the validation set for testing. The proposed method can be combined with more discriminative image representation



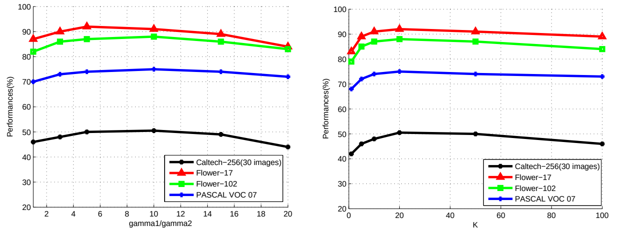
(a) *Caltech-256 dataset*

(b) *Flower-17 dataset*

(c) *Flower-102 dataset*

(d) *PASCAL VOC 2007 dataset*

Fig. 4. Performance changes with α and β jointly on the (a) Caltech-256 dataset, (b) the Flower-17 dataset, (c) the Flower-102 dataset and (d) the PASCAL VOC 2007 dataset.



(a) $\gamma 1/\gamma 2$

(b) K

Fig. 5. Performance changes with (a) γ_1, γ_2 (γ_1=γ_2) and (b) K on the Caltech-256 dataset, the Flower-17 dataset, the Flower-102 dataset and the PASCAL VOC 2007 dataset.

methods for local and global classifiers learning. We use the image representation as [60] and learn the classifiers for class prediction (ISC-LG-VGG). We also give the performance of using CNN [9] for classification (ISC-LG-CNN). Table 5 shows the error rates of different methods on this dataset. We can see from Table 5 that the proposed method can consistently improve the classification performances. Besides, by combining VGG for discriminative image representations, we can improve over CNN based method. The experimental results on the ILSVRC-2012 ImageNet dataset proves the effectiveness of the proposed method again.

V. Conclusion

In this paper, we propose a novel image classification method by classifying each image with local and global discriminations. We first select the k nearest neighbors of one testing image for local classifier training. All the training images are used for global classifier training. The local and global classifiers are jointly learned to not only consider the characters of the testing image but also help to avoid the local optimum when only local information is used. We have tested the proposed image-specific classification with local and global discriminations method on several public image datasets with improved performances.

TABLE V
The error rates of different methods on the ILSVRC-2012 ImageNet dataset.

Methods	top-1 error	top-5 error
CNN [9]	36.7	15.4
VGG [60]	25.5	8.0
Zeiler [61]	36.0	14.7
GoogLeNet [62]	-	9.2
ISC-LG-CNN	33.6	13.1
ISC-LG-VGG	23.8	7.4

VI. Acknowledgments

This work is supported by National Natural Science Foundation of China: 61303154 and 61332016, the Scientific Research Key Program of Beijing Municipal Commission of Education (KZ201610005012). This work is supported in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar, National Science Foundation of China: 61429201.

References

[1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos", In *Proc. ICCV*, pp.1470-1477, 2003.

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

[3] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification", In *Proc. CVPR*, pp.1794-1801, USA, 2009.

[4] C. Zhang, J. Cheng, J. Liu, J. Pang, Q. Huang and Q. Tian, "Beyond Explicit Codebook Generation: Visual Representation Using Implicitly Transferred Codebooks", *IEEE Transactions on Image Processing*, 24(12):5777-5788, 2015.

[5] J. Gemert, C. Veenman, A. Smeulders and J. Geusebroek. "Visual word ambiguity", In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7), pages 1271-1283, 2010.

[6] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification", In *Proc. CVPR*, pp.3360-3367, 2010.

[7] S. Gao, I. Tsang, and L. Chia, "Laplacian sparse coding, hypergraph laplacian sparse coding, and applications", In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):92-104, 2013.

[8] 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*, pp.1673-1680, 2011.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", In *NIPS*, pp.1097-1105, 2012.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv:1409.1556[cs.CV].

[11] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision", In *Proc. Int. Symp. Circuits Syst.*, pp.253-256, 2010.

[12] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture", In *Proc. CVPR*, pp.3743-3752, 2015.

[13] J. Sivic, M. Everingham and A. Zisserman, ”Who are you? Learning person specific classifiers from video”, In *Proc. CVPR*, pp.1145-1152, 2009.

[14] K. Sungwoong, N. Sebastian, K. Pushmeet, and DY Chang, ”Task-specific image partitioning”, *IEEE Transactions on Image Processing*, 22(2):488-500, 2013.

[15] C. Zhang, C. Liang, L. Li, J. Liu, Q. Huang, and Q. Tian, ”Fine-grained image classification via low-rank sparse coding with general and class-specific codebooks”, *IEEE Transactions on Neural Network and Learning Systems*, DOI:10.1109/TNNLS.2016.2545112, 2016.

[16] T. Malisiewicz, A. Gupta and A. Efros, ”Ensemble of exemplar-SVMs for object detection and beyond”, In *Proc. ICCV*, Pages:89-96, 2011.

[17] D. Modolo, A. Vezhnevets, O. Russakovsky and V. Ferrari, ”Joint calibration of ensemble of exemplar SVMs”, In *Proc. CVPR*, pp.3955-3963, 2015.

[18] 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*, 120:318-324, 2013.

[19] C. Zhang, J. Cheng, J. Liu, J. Pang, C. Liang, Q. Huang, and Q. Tian, ”Object categorization in sub-semantic space”, *Neurocomputing*, 142:248-255, 2014.

[20] D. Weinshapp, A. Zweig, H. Hermansky, S. Kombrink, F. Ohl, J. Anemuller, J. Bach, L. Gool, F. Nater, T. Pajdla, M. Havlena, and M. Pavel, ”Beyond novelty detection: Incongruent events, when general and specific classifiers disagree”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1886-1901, 2012.

[21] M. Guillaumin, J. Verbeek and C. Schmid, ”Multimodal semi-supervised learning for image classification”, In *Proc. CVPR*, pp.902-909, 2010.

[22] B. Fernando, A. Habrard, M. Sebban and T. Tuytelaars, ”Unsupervised visual domain adaptation using subspace alignment”, In *Proc. ICCV*, pp.2960-2967, 2013.

[23] L. Li, H. Su, E. Xing, and Li. Fei-Fei, ”ObjectBank: A high-level image representation for scene classification & semantic feature sparsification”, In *NIPS*, Vancouver, Canada, pp.1378-1386, 2010.

[24] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, and S. Yan, ”Contextualizing object detection and classification”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):13-27, 2015.

[25] Q. Liu, Y. Sun, C. Wang, T. Liu, and D. Tao, ”Elastic net hypergraph learning for image clustering and semi-supervised classification”, CoRR abs/1603.01096, 2016.

[26] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, ”transductive multi-view zero-shot learning”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2332-2345, 2015.

[27] Y. Boureau, N. Roux, F. Bash, J. Ponce and Y. LeCun, ”Ask the locals: Multi-way local pooling for image recognition”, In *Proc. ICCV*, pp.2651-2658, 2011.

[28] A. Shabou and H. Le-Borgne, ”Locality-constrained and spatially regularized coding for scene categorization”, In *Proc. CVPR*, pp.3618-3625, 2012.

[29] C. Zhang, G. Zhu, C. Liang, Y. Zhang, Q. Huang, and Q. Tian, ”Image class prediction by joint object, context and background modeling”, *IEEE Transactions on Circuits and Systems for Video Technology*, DOI: 10.1109/TCSVT.2016.2613125.

[30] H. Zhang, A. Berg, M. Maire, J. Malik, ”SVM-KNN: Discriminative nearest neighbor classification for visual category recognition”, In *Proc. CVPR*, pp.2126-2136, 2006.

[31] N. Armanfard, J. Reilly, and M. Komeili, ”Local feature selection for data classification”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI:10.1109/TPAMI.2015.2478471.

[32] N. Inoue, and K. Shinoda, ”Fast coding of feature vectors using neighbor-to-neighbor search”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI:10.1109/TPAMI.2015.2481390.

[33] O. Boiman, E. Shechtman, and M. Irani, ”In defense of nearest-neighbor based image classification”, In *Proc. CVPR*, pp.1-8, 2008.

[34] V. Athitsos, J. Alon, and S. Sclaroff, ”Efficient nearest neighbor classification using a cascade of Approximate Similarity Measures”, In *Proc. CVPR*, pp.486-493, June 2005.

[35] S. Mccann, and D. Lowe, ”Local Naive Bayes Nearest Neighbor for Image Classification”, In *Proc. CVPR*, pp.3650-3656, 2012.

[36] R. Cinbis, J. Verbeek and C. Schmid, ”Image categorization using fisher kernels of non-iid image models”, In *Proc. CVPR*, pp.2184-2191, 2012.

[37] K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman, The devil is in the details: An evaluation of recent feature encoding methods, In *BMVC*, 2011, pages 1-12.

[38] K. Zhang, Q. Liu, Y. Wu, and M. Yang, ”Robust Visual Tracking via Convolutional Networks Without Training”, *IEEE Transactions on Image Processing*, 25(4):1779-1792, 2016.

[39] M. Gong, J. Liu, H. Li, Q. Cai, and L. Su, ”A multiobjective sparse feature learning model for deep neural networks”, *IEEE Transactions on Neural Networks and Learning Systems*, 26(12):3263-3277, 2015.

[40] X. Zhang, H. Zhang, Y. Zhang, Y. Yang, M. Wang, H. Luan, J. Li, and T. Chua, ”Deep Fusion of Multiple Semantic Cues for Complex Event Recognition”, *IEEE Transactions on Image Processing*, 25(3):1033-1046, 2016.

[41] G. Griffin, A. Holub, and P. Perona, Caltech-256 object category dataset, Technical report, CalTech, 2007.

[42] M. Nilsback and A. Zisserman, ”A visual vocabulary for flower classification”, In *Proc. CVPR*, pp.1447-1454, 2006.

[43] M. Nilsback and A. Zisserman, ”Automated flower classification over a large number of classes”, In *Proc. ICCVGIP*, pp.722-729, 2008.

[44] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, The PASCAL visual object classes challenge 2007 (VOC 2007) results, Technical report, Pascal Challenge, 2007.

[45] K. Sande, T. Gevers and C. Snoek, ”Evaluating color descriptors for object and scene recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582-1596, 2010.

[46] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, ”Caffe: Convolutional Architecture for Fast Feature Embedding”, In *Proceedings of the 22nd ACM international conference on Multimedia*, pp.675-678, 2014.

[47] J. Sanchez, F. Perronnin, T. Mensink and J. Verbeek, ”Image classification with the fisher vector: Theory and practice”, *International Journal of Computer Vision*, December 2013, 105(3):222-245.

[48] M. Varma and D. Ray, ”Learning the discriminative power invariance trade-off”, In *Proc. ICCV*, pp.1-8, 2007.

[49] N. Xie, H. Ling, W. Hu and X. zhang, ”Use bin-ratio information for category and scene classification”, In *Proc. CVPR*, pp.2313-2319, 2010.

[50] X. Yuan and S. Yan, ”Visual classification with multi-task joint sparse representation”, In *Proc. CVPR*, pp.3493-3500, 2010.

[51] S. Gao, I. Tsang and Y. Ma, ”Learning category-specific dictionary and shared dictionary for fine-grained image classification”, *IEEE Transactions on Image Processing*, 23(2):623-634, 2014.

[52] A. Angelova and S. Zhu, ”Efficient object detection and segmentation for fine-grained recognition”, In *Proc. CVPR*, pp.811-818, 2013.

[53] 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*, pages 794-807, 2012.

[54] J. Yang, Y. Li, Y. Tian, L. Duan and W. Gao, ”Group-sensitive multiple kernel learning for object categorization”, In *Proc. ICCV*, pp.436-443, 2009.

[55] K. Chatfield, K. Simnyan, A. Vedaldi and A. Zisserman, ”Return of the devil in the details: Delving deep into convolutional nets”, In *BMVC*, pp.1-12, 2014.

[56] Y. Yang, D. Xu, F. Nie, S. Yan and Y. Zhuang, ”Image clustering using local discriminant models and global integration”, *IEEE Transactions on Image Processing*, 19(10):2761-2773, 2010.

[57] L. Zhang, R. Ji, Y. Xia, Y. Zhang, and X. Li, ”Learning a probabilistic topology discovering model for scene categorization”, *IEEE Transactions on Neural Networks and Learning Systems*, 26(8):1622-1635, 2015.

[58] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, and Q. Tian, ”Discovering discriminative graphlets for aerial image categories recognition”, *IEEE Transactions on Image Processing*, 22(12):5071-5084, 2013.

[59] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and Li Fei-Fei, ”ImageNet large scale visual recognition challenge”, *International Journal of Computer Vision*, 115(3):211-252, 2015.

[60] K. Simonyan, and A. Zisserman, ”Very deep convolutional networks for large-scale image reocgnition”, arXiv technical report, arxiv.1409.1556, 2014.

[61] M. Zeiler, and R. Fergus, ”Visualizing and understanding convolutional networks”, In *Proc. ECCV*, pp.818-833, 2014.

[62] C. Szegedy,W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ”Going deeper with convolutions”, In *Proc. CVPR*, pp.1-9, 2015.

[63] H. Xiong, W. Yu, X. Yang, M. Swamy, and Q. Yu, ”Learning the conformal transformation kernel for image recognition”, *IEEE Transactions on Neural Networks and Learning Systems*, 28(1):149-163, 2017.

[64] J. Raitoharju, S. Kiranyaz, and M. Gabbouj, ”Training radial basis function neural networks for classification via class-specific clustering”, *IEEE Transactions on Neural Networks and Learning Systems*, 27(12):2458-2471, 2016.

[65] Y. Li, S. Wang, Q. Tian, and X. Ding, ”A boosting approach to exploit instance correlations for multi-instance classification”, *IEEE Transactions on Neural Networks and Learning Systems*, 27(12):2740-2747, 2016.

[66] C. Deng, J. Xu, K. Zhang, D. Tao, X. Gao, and X. Li, ”Similarity constraints-based structured output regression machine: An approach to image super-resolution”, *IEEE Transactions on Neural Networks and Learning Systems*, 27(12):2472-2485, 2016.