Information Sciences 376 (2017) 125–135

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Image classification by search with explicitly and implicitly semantic representations

Chunjie Zhang[a,b], Guibo Zhu[c,*], Qingming Huang[a,b,d], Qi Tian[e]

[a] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[b] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, China
[c] Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
[d] Key Lab of Intell. Info. Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[e] Department of Computer Sciences, University of Texas at San Antonio TX, 78249, U.S.A

ARTICLE INFO

*Article history:*
Received 20 June 2016
Revised 20 September 2016
Accepted 9 October 2016
Available online 11 October 2016

*Keywords:*
Explicit representation
Implicit representation
Semantic modeling
Image classification

ABSTRACT

Image classification refers to the task of automatically classifying the categories of images based on the contents. This task is typically solved using visual features with the histogram based classification scheme. Although effective, this strategy has two drawbacks. On one hand, histogram based representation often disregards the object layout which is very important for classification. On the other hand, visual features are unable to fully separate different images due to the semantic gap. To solve these two problems, in this paper, we propose a novel image classification method by explicitly and implicitly representing the images with searching strategy. First, to make use of object layouts, we randomly select a number of regions and then use these regions for image representations. Second, we generate the explicitly semantic representations using a number of pre-learned semantic models. Third, we measure the visual similarities with the Internet images and use the text information for implicitly semantic representations. Since Internet images are contaminated with noise, the resulting representations only implicitly reflect the contents of images. Finally, both the explicitly and implicitly semantic representations are jointly modeled for image classifications by training bi-linear classifiers. We evaluate the effectiveness of the proposed image classification by search with explicitly and implicitly semantic representations method (EISR) on the Scene-15 dataset, the MIT-Indoor dataset, the UIUC-Sports dataset and the PASCAL VOC 2007 dataset. The experimental results prove the usefulness of the proposed method.

## 1. Introduction

To effectively classify images, we need to first discriminatively represent the images. However, this is not an easy task as visual features do not have explicitly semantic correspondences with the semantic concepts. To bridge this gap, both the exploration of discriminative visual representations and semantic descriptions of images are studied.

To explore the visual information, some researchers try to design discriminative features [3,18] while others try to make various transformations and combinations of these features. The well-designed features can cope with various

[*] Corresponding author.
*E-mail addresses:* zhangcj@ucas.ac.cn, zhangchunjie1983@gmail.com (C. Zhang), guibo.zhu@ia.ac.cn (G. Zhu), qmhuang@ucas.ac.cn (Q. Huang), qitian@cs.utsa.edu (Q. Tian).

http://dx.doi.org/10.1016/j.ins.2016.10.019
0020-0255/© 2016 Elsevier Inc. All rights reserved.

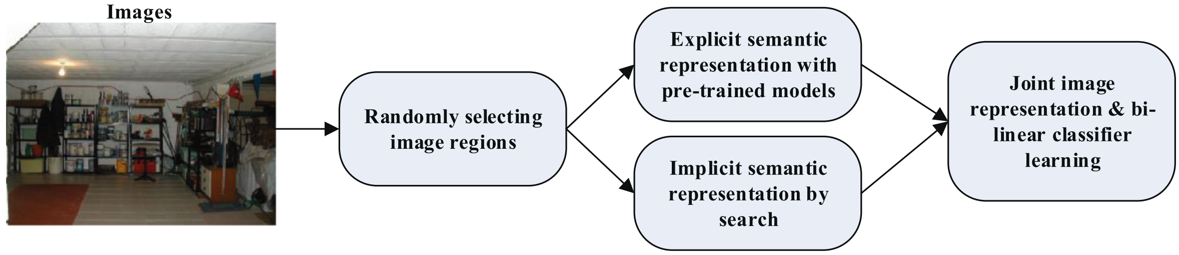
Images
Randomly selecting image regions
Explicit semantic representation with pre-trained models
Implicit semantic representation by search
Joint image representation & bi-linear classifier learning

**Fig. 1.** Flowchart of the proposed image classification method by search with explicitly and implicitly semantic representations (EISR).

image deformations and are widely used. However, the designing of discriminative and robust features is very hard which remains a question that needs to be solved. Instead of using visual features directly, the transformation of features becomes another choice to boost the classification performances. The spatial layout and context correlation are widely explored [8,10,12,14,30,36,38,40–44,46,48]. Besides, the combinations of different types of features [4,20,35,50] are also very popular. Different features are combined for image representations [20,35,50] and classifier training [4]. One problem with the visually based strategy is the semantic inconsistence between visual information and semantics. Especially when images are visually similar but have different semantics.

To alleviate the semantic discrepancy, the use of semantic representations of images becomes popular [5,16,21,22,28,44]. This helps to represent images in an understandable way and is more consistent with human perceptions than visual features. Usually, the semantic space is generated by pre-definitions [37] or by harvesting from training images [5,16,22,28,44]. Using pre-defined semantics is often labor intensive and requires domain knowledge. Besides, harvesting information from training images is often limited by the collected images. The use of Internet information helps to cope with the dataset bias. However, it can only implicitly describe the semantics with noisy information. Moreover, the spatial layouts of different semantics are often unconsidered by the semantically based methods. The combination of spatial layout with semantic representations can help to improve the discriminative power of the final image representations.

To solve the problems mentioned above, in this paper, we propose a novel image classification method by search with explicitly and implicitly semantic representations (EISR). We first randomly select image regions as the basic elements for image representations. For each region, we try to represent it with explicitly and implicitly semantic representations. The explicitly semantic representation is obtained using the pre-trained models which are learned using images collected by domain experts. Since images of each class depict the same concept and contain no irrelevant images, the trained model can help to classify the corresponding concept well. Hence, we use it for explicitly semantic representations. The implicitly semantic representations are obtained by harvesting from the Internet. This is achieved by searching the visually similar images for each image region and using the corresponding text information for implicitly semantic representations. Since web images are often contaminated with irrelevant information, these text information only reflect implicit semantics. We combine the explicitly and implicitly semantic representations for final image representations and train bi-linear classifiers to predict image classes. We evaluate the effectiveness of the proposed EISR method on the Scene-15 dataset, the MIT-Indoor dataset, the UIUC-Sports dataset and the PASCAL VOC 2007 dataset. Experimental results prove the usefulness of the proposed method. Fig. 1 shows the flowchart of the proposed method.

The main contributions of this paper lie in three aspects:

- First, we propose a novel image classification method by explicitly and implicitly modeling the semantic representations. This helps to alleviate the discrepancy between visual information and human understandings to some extent.
- Second, we model the explicitly and implicitly semantic representations jointly by exploring the information of training images and the images from the Internet. This helps to harvest the information from various sources and boost the discriminative power of the final image representations.
- Third, by representing images more semantically, we can improve the classification accuracy over many visually based methods.

Compared with search based methods for image annotation and classification [29,31,32,47], the improvements of the proposed method lie in two aspects. First, instead of only using the Internet information, EISR also combines the explicitly semantic representation using training images for joint representations. Second, we use the matrix based representations to jointly model the correlations among image regions which is more discriminative than histogram based representations.

The rest of this paper is organized as follows. We give the related work in Section 2. In Section 3, we systematically describe the details of the proposed image classification method by search with explicitly and implicitly semantic representations. The experimental results and analysis are given in Section 4. Finally, we give the conclusions in Section 5.

## 2. Related work

With the explosion of visual information, how to efficiently analyze the visual content became popular. Many discriminatively designed features were proposed [3,18]. SIFT feature was proposed by Lowe [18] and widely used for various visual applications. Dalal and Triggs [3] proposed a simplified version of SIFT as histograms of oriented gradients and used it for object detection.

Although effective, the well designed visual features cannot cope with all types of deformations. To increase the discriminative power, the spatial layout and context correlation were also explored [8,10,14,30,36,40–45,48]. Lazebnik et. al. [14] proposed the spatial pyramid matching technique which was widely used by researchers. Gemert et. al. [10] explored the distance information for local feature encoding which reduced the quantization loss. Yang et. al. [36] used sparse coding instead of nearest neighbor assignment for local feature quantization. Wang et. al. [30] restricted the sparse coding process with locality constraint and improved the efficiency. Zhang et. al. [45] proposed to use non-negative sparse coding while Gao et. al. [8] used laplacian sparse coding to reduce quantization loss. Zhang et. al. [41] made harr-like transformation over local features. The low-rank constraint was used to model the correlations of images of the same class [39]. Zhang et. al. [43] also proposed a bi-linear model for object recognition while Zheng et. al. [48] used graph regularized sparse coding. Zhang et. al. [39] proposed to use the orientation and location information along with the visual information of local features for better classification.

Instead of using one type of feature, the combinations of different features [4,20,35,50] were also used. Xiao et. al. [35] proposed a kernel reconstruction technique for sparse representation. Zhou et. al. [50] proposed to encode local features with super-vector while Rao et. al. [20] tried to recognize actions with view-invariant representations. Darrell et. al. [4] used interpolated views for specific tasks.

However, only using visual information cannot alleviate the semantic gap. Hence, the use of semantic representations became popular [5,16,21,22,28,44]. Rasiwasia and Vasconcelos [21] used holistic context models by mining the semantical correlations of different concepts. Rasiwasia and Vasconcelos [22] also used low-dimensional semantic spaces for scene classification while Torresani et. al. [28] used classesmes for object recognition with improved accuracy. The ObjectBank was proposed by Li et. al. [16] by using the information from the Internet images. Zhang et. al. [44] proposed to classify images with weak semantic representation by using the discriminative information of exemplar classifier. Dixit et. al. [5] combined the semantic representation with fisher vector encoding for scene classification. The harvesting of Internet information for image annotation was also widely studied [29,31,32,47]. Wang et. al. [31] tried to annotate images by search. However, there were a lot of noisy information which hindered the performance. Hence, the use of duplicate information for semantic transfer became popular. Wang et. al. [32] improved the annotation with duplicate search while Wang et. al. [29] targeted the face annotation to incorporate task-specific information. Zhao et. al. [47] also tried to annotate web videos by near-duplicate search to cope with variations. However,the contamination of noisy information degraded the reliability of this strategy.

## 3. Search based explicitly and implicitly semantic representations for image classification

In this section, we give the details of the proposed image classification by search with explicitly and implicitly semantic representation method.

### *3.1. Randomly selecting image regions*

To make classification of images, we need to represent them first. Instead of using histogram based methods, we use image region as the basic element for image representation. We randomly select a number of regions to cover the whole image. These selected image regions contain the spatial layouts of objects to some extent and may overlap with each other. We use the random selection strategy for three reasons. First, objects may appear on various places of images. Second, locating the exact positions of objects is computationally expensive and inaccurate. Besides, we are not able to detect every objects reliably. Third, by randomly selection, we can extract the image regions very quickly.

After extracting the image regions, we use them for joint image representations by concatenating each region's representation together using the selection order in a matrix form. Formally, let $\boldsymbol{x}_n$ be the representation of the $n$th region. Suppose we select $N$ regions in total for each image, the final image representation is obtained as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$. To represent each image region, we combine the explicitly and implicitly semantic representations together.

### *3.2. Explicitly semantic representation*

Images of the same class collected by human experts concentrate on the same concept and are free of noise. This information can be used for explicitly semantic representation. We use these images to train semantic prediction models. 'explicitly semantic' means the learned models are specific and free of noisy information.

Formally, let $\phi_m(*)$ be the learned semantic prediction model for the $m$th concept, $M$ is the number of semantics (or image classes). The explicitly semantic representation for the $n$th image region $\boldsymbol{x}_n$ can then be obtained as:

$$\boldsymbol{x}_{n,ex} = [\phi_1(\boldsymbol{h}_n); \phi_2(\boldsymbol{h}_n); \ldots; \phi_M(\boldsymbol{h}_n)] \tag{1}$$

where $\boldsymbol{h}_n$ is the visual information of the $n$th region. Since there are many semantic prediction models, we can use them directly in this paper. The proposed explicitly semantic representation scheme can also use different image region representation methods, e.g. local feature based bag-of-visual-word model, convolutional neural network based model. Besides, we can even use different semantic prediction models for different regions. Since each learned model predicts the relevance of the image region with the corresponding concept, the resulting representation $\boldsymbol{x}_{n,\,ex}$ reflects the semantic likelihood accordingly.

### 3.3. Implicitly semantic representation

Only using the explicitly semantic representation for image classification is not enough. The images collected by experts are biased which cannot fully represent different semantics well. It is necessary to harvest information from other sources for more discriminative representations. The Internet provides plenty of such information which can be used.

We use the Internet images by search for implicitly semantic representations. Specially, for each selected image region, we use the Google image search engine in this paper and select the top 1000 images. Each web link of the corresponding image contains some text description. We use these text information for implicitly semantic representation $\boldsymbol{x}_{n,im} \in \mathbb{R}^{\widetilde{M}\times 1}$.

We call this representation as implicitly semantic representation for two reasons. First, the selected images are not always semantically correlated with the image region to be searched. Due to the semantic gap, visually similar images do not always belong to the same class. Hence, the retrieved images are only partially semantically correlated with the image region. Second, for each selected image, its corresponding text descriptions are also contaminated with noisy information. However, these retrieved text contains some useful information which reflects the semantics of the corresponding image region. Note that although we use the retrieved text for implicitly semantic representation, other information extraction methods, such as topic models (probabilistic latent semantic analysis (pLSA) and Latent Dirichlet Allocation (LDA)) can also be used. The explicitly semantic representation $\boldsymbol{x}_{n,ex}$ and implicitly semantic representation $\boldsymbol{x}_{n,im}$ are concatenated for joint representation as $\boldsymbol{x}_n = [\boldsymbol{x}_{n,ex}; \boldsymbol{x}_{n,im}]$.

### 3.4. Image classification

After obtaining the final image representation $\boldsymbol{X}$, we can learn the classification model for image classification. We use subscript the indicate the index of image regions and use superscript to indicate the index of training images. Formally, let $(\boldsymbol{X}^p, y^p)$, $p = 1, \ldots, P$ be the $P$ training images with the corresponding labels, we learn a prediction model $\psi(*)$ to predict the classes of images as:

$$\widehat{y^p} = \psi(\boldsymbol{X}^p) \tag{2}$$

by minimizing the summed loss over training images as:

$$\psi(*) = argmin_{\psi(*)} \sum_{p=1}^{P} \ell(\psi(\boldsymbol{X}^p), y^p) + \lambda\Omega(\psi(*)) \tag{3}$$

where $\ell(*, *)$ is the loss function, $\Omega(*)$ is the regularization term, $\lambda$ is the parameter.

We use bi-linear classifier for image class prediction with $\psi(\boldsymbol{X}^p) = \boldsymbol{\alpha}^T\boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{\alpha} \in \mathbb{R}^{(M+\widetilde{M})\times 1}$ is the parameter which combines the influences of semantic elements while $\boldsymbol{\beta} \in \mathbb{R}^{N\times 1}$ considers the correlations of selected regions. Usually, the hinge loss is used for classification. However, hinge loss is not differentiable. Hence, we use the quadratic hinge loss [5] as:

$$\ell(\boldsymbol{\alpha}^T\boldsymbol{X}\boldsymbol{\beta}, y^p) = max^2(0, 1 - \boldsymbol{\alpha}^T\boldsymbol{X}\boldsymbol{\beta} \times y^p) \tag{4}$$

As to the regularization term, we use the $L_2$ norm which is widely used for classification as $\Omega(\psi(*)) = \|\boldsymbol{\alpha}\|^2 + \|\boldsymbol{\beta}\|^2$. After the parameters are learned, we can predict the classes of images using Eq. 2. Algorithm 1 gives the procedures of the proposed image classification by search with explicitly and implicitly semantic representation method.

**Algorithm 1** Training phrase of the proposed image classification by search with explicitly and implicitly semantic representation method.

```
Input:
    The training images and labels, $\lambda$, random selection number $N$, initial $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$;
Output:
    The learned $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\alpha \in \mathbb{R}^{(M+\widetilde{M})\times 1}$ is the parameter which combines the influences of semantic elements while $\beta \in \mathbb{R}^{N\times 1}$ considers the correlations of selected regions;
Randomly select $N$ regions for each image;
For each region, calculate the explicitly semantic representation $\boldsymbol{x}_{n,ex}$ with Eq. (1);
For each region, calculate the implicitly semantic representation $\boldsymbol{x}_{n,im}$ by search as described in Subsection 3.3;
Concatenate the explicitly and implicitly semantic representations $\boldsymbol{x}_{n,ex}$ and $\boldsymbol{x}_{n,im}$ of image regions for image representation as $\boldsymbol{x}_n = [\boldsymbol{x}_{n,ex}; \boldsymbol{x}_{n,im}]$;
Train the classifiers by optimizing over Eq. (3);
return The learned $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$.
```

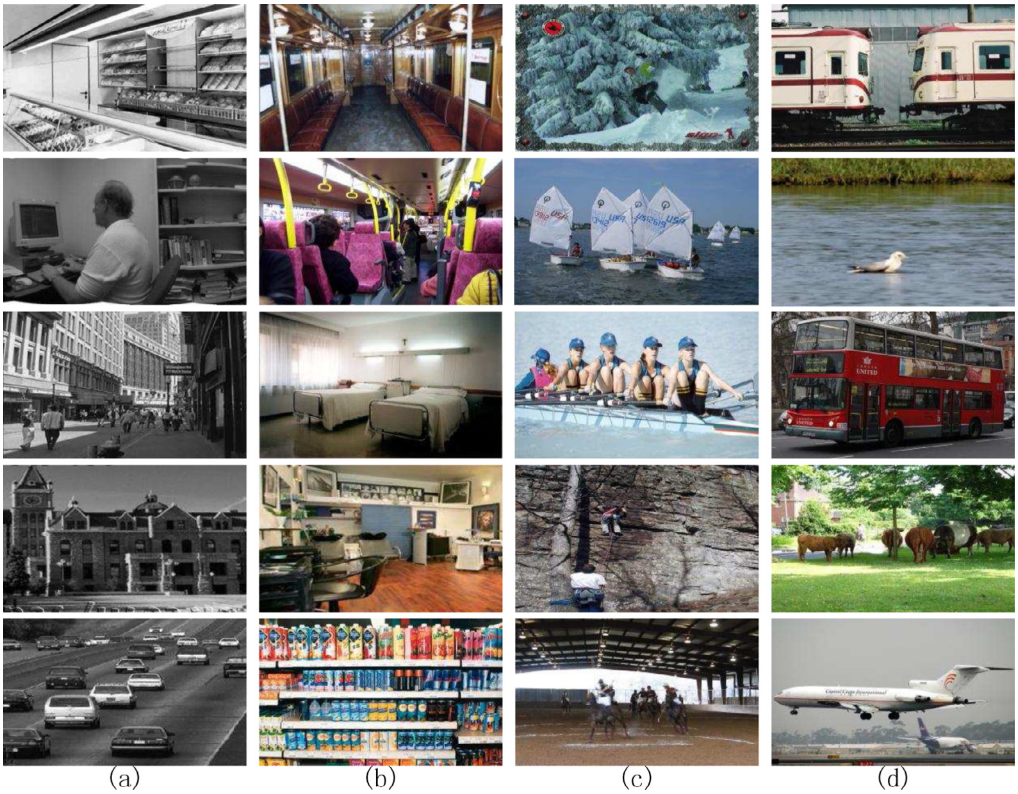

**Fig. 2.** Example images of (a) the Scene-15 dataset, (b) the MIT-Indoor dataset, (c) the UIUC-Sports dataset and (d) the PASCAL VOC 2007 dataset.

## 4. Experiments

To evaluate the effectiveness of the proposed image classification by search with explicitly and implicitly semantic representations method (EISR), we conduct experiments on the Scene-15 dataset [14], the MIT-Indoor dataset [19], the UIUC-Sports dataset [15] and the PASCAL VOC 2007 dataset [7]. Fig. 2 shows example images of the four datasets.

### 4.1. Experimental Setup

For fair comparison, we follow the experimental setup as other researchers and use the reported results for performance evaluations instead of re-implementing these algorithms. To represent images, we use two methods: (1) local feature based BoW model, (2) convolutional neural network (CNN) based deep learning model. As to the extraction of local features, we densely extract color SIFT features [27] with multiple scales. We set the minimum scale to 16 × 16 pixels with 6 pixels overlap. We use the sparse coding technique for local feature encoding [36] with a codebook of 1000 visual words. As to the CNN based image representation, we use the six layer networks as [13]. The classifiers are trained in the one-vs-all way. The final prediction is conducted by assigning the image with the class of the largest response. Classification rate is used for quantitative evaluations.

### 4.2. Scene-15 dataset

This dataset has 200–400 images for each class with the fifteen classes as: *store, office, tallbuilding, street, opencountry, mountain, insidecity, highway, forest, coast, livingroom, kitchen, industrial, suburb and bedroom*. We randomly select 100 images per class for training and view the other images as testing samples. The random selection is repeated for ten times for reliable comparison.

To quantitatively evaluate the performances of the proposed EISR method, we give the performance comparisons on the Scene-15 dataset with the baseline methods in Table 1. The use of local features and CNN based methods are denoted as EISR-SC and EISR-CNN respectively. Besides, we also give the performances without search (EISR-SC(no search) and EISR-CNN(no search)) in Table 1. The boxplots of per-class performance are also given in Fig. 3.

From Table 1 and Fig. 3, we can have four conclusions. First, EISR is able to outperform many baseline methods ranging from visually based methods [8,10,14,30,36] to semantically based methods [1,16,21,22,34,39,44]. The semantically based rep-

**Table 1**
Performance comparison of EISR with other methods on the Scene-15 dataset. Numerical values stand for mean and standard derivation respectively.

| Algorithms | Classification rate |
|---|---|
| pLSA [1] | 72.7 |
| LScSPM [8] | 89.75 ± 0.50 |
| KC [10] | 76.67 ± 0.39 |
| KSPM [14] | 81.40 ± 0.50 |
| ObjectBank [15] | 80.9 |
| Contextual Models [21] | 77.20 ± 0.39 |
| LDA [22] | 59.0 |
| Semantic Space [22] | 73.95 ± 0.74 |
| LLC [30] | 81.50 ± 0.87 |
| pLSA [34] | 63.3 |
| ScSPM [36] | 80.28 ± 0.93 |
| $S^3$R [39] | 83.72 ± 0.78 |
| WSR-EC [44] | 81.54 ± 0.59 |
| Search-Only | 46.72 ± 3.85 |
| EISR-SC | 90.15 ± 0.83 |
| EISR-CNN | 94.53 ± 0.76 |
| EISR-SC (no search) | 86.65 ± 0.80 |
| EISR-CNN (no search) | 92.18 ± 0.86 |

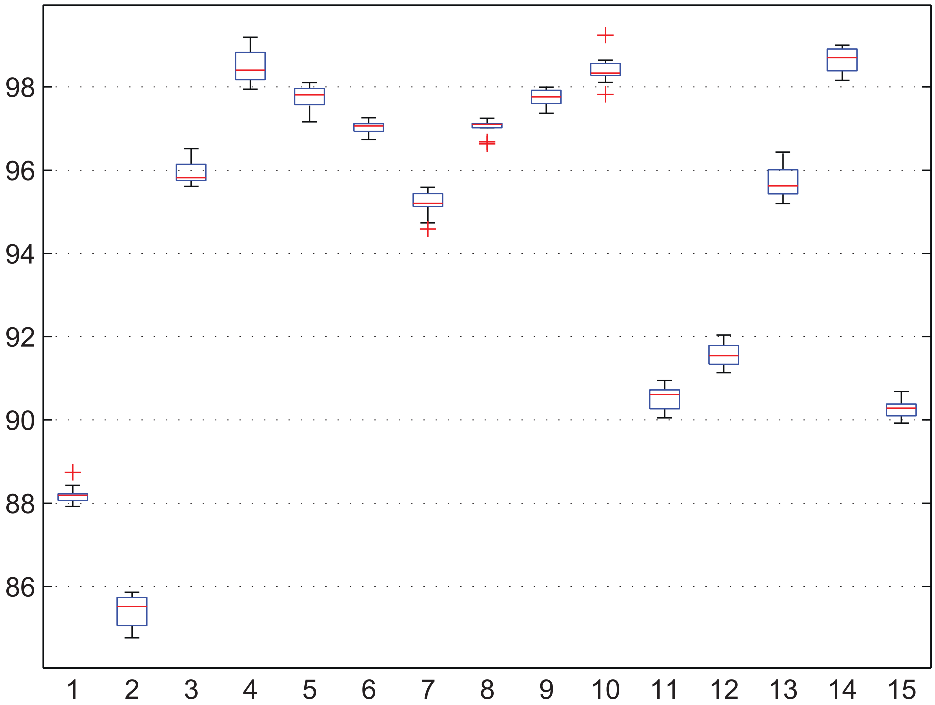

**Fig. 3.** Boxplot of the per-class performance of EISR-CNN on the Scene-15 dataset. From left to right: store, office, tallbuilding, street, opencountry, mountain, insidecity, highway, forest, coast, livingroom, kitchen, industrial, suburb and bedroom.

resentations help to cope with the discrepancy between visual information and human perception. Second, without search for implicitly semantic representation, EISR still be able to improve over other semantically based methods. The use of overcomplete regions for image representation helps to encode more spatial correlations. Third, text based semantic analysis methods cannot alleviate the semantic gap of visual features. Hence, directly using pLSA and LDA cannot achieve comparable performances as visually based methods. However, by generating the semantic representations using visual information, we are able to improve the performances over visually based methods. Fourth, EISR also improves over ObjectBank which also uses Internet images. This is because we also try to generate the semantic representations with carefully collected images. The search results are contaminated with various noise which degenerates the performances. The large variation also indicates using search results only is not very robust. However, by combing the explicitly semantic representation, we can alleviate this problem. Overall, these results prove the effectiveness of the proposed EISR method for improving the classification performances.

**Table 2**
Mean classification rate comparisons of EISR with other methods on the MIT-Indoor dataset.

| Methods | Classification rate |
|---|---|
| Doersch [6] | 66.87 |
| Gong [11] | 68.90 |
| Lin [17] | 68.50 |
| KSPM [14] | 34.40 |
| Quattoni [19] | 26.50 |
| Razavian [23] | 69.00 |
| LPR-LIN [25] | 44.84 |
| CENTRIST [33] | 36.90 |
| Zhou [49] | 70.08 |
| EISR-SC | 50.37 |
| EISR-CNN | 71.62 |
| EISR-SC (no search) | 43.58 |
| EISR-CNN (no search) | 66.25 |

**Table 3**
Performance comparisons of EISR with other methods on the UIUC-Sports dataset.

| Algorithms | Performance |
|---|---|
| LScSPM [8] | 85.31 ± 0.51 |
| CSDL [9] | 86.54 ± 0.56 |
| LLC [30] | 83.09 ± 1.30 |
| ScSPM [36] | 82.74 ± 1.46 |
| LRSC [44] | 88.17 ± 0.85 |
| EISR-SC | 87.46 ± 0.62 |
| EISR-CNN | 92.73 ± 0.55 |
| EISR-SC (no search) | 85.39 ± 0.58 |
| EISR-CNN (no search) | 89.65 ± 0.73 |

### *4.3. MIT-Indoor dataset*

There are 67 classes of indoor images in the MIT-Indoor dataset with a total of 15,620 images. We follow the data split setup as Quattoni and Torralba [19] and use 80 images per-class for training. Table 2 gives the mean classification rate comparisons of EISR with other baseline methods on the MIT-Indoor dataset.

We can have similar conclusions as on the Scene-15 dataset. First, EISR is again able to improve over both visually and semantically based methods. Second, the searching strategy helps to improve the performances better than that on the Scene-15 dataset. We believe this is because images of the MIT-Indoor dataset are more complicated and concentrate on the indoor scenes. Third, CNN based representations are more discriminative than local feature based methods. Hence, the EISR-CNN can improve over EISR-SC dramatically. The experimental results on the MIT-Indoor dataset again prove the usefulness of the proposed EISR method.

### *4.4. UIUC-Sports dataset*

The UIUC-Sports dataset has sport images of eight classes as: *badminton, bocce, croquet, polo, rock climbing, rowing, sailing* and *snow boarding*. There are 1792 images with the number of each class ranges from 137 to 250. We follow the experimental setup of [15] and randomly select 70 images per class for training and use the other images for testing. Table 3 shows the performance comparisons of EISR with other methods on the UIUC-Sports dataset. To show the per-class performance, we also give the boxplot of the performance in Fig. 4.

Since images of this dataset are relatively easier to classify than the MIT-Indoor dataset, the relative improvement of EISR over other methods is small. However, EISR is still able to improve the performances over these baseline methods. Besides, by combining with the search strategy, we are able to further improve the classification performances. Moreover, the per-class performances of EISR are stable for different classes which shows the robustness of the proposed method.

### *4.5. The PASCAL VOC 2007 dataset*

This dataset has more than 10,000 images which consist of twenty classes (*aeroplane, bicycle, boat, bottle, bus, bird, car, cat, cow, chair, dining table, dog, horse, person, sheep, motorbike, train, potted plant, soft* and *tv/monitor*). The images are provided with train/validate/test splits. Classifiers are firstly trained using the train split and the optimal parameters are selected with the validate split. The train and validate splits are then merged to re-train the classifiers with the learned parameters. The final performance is evaluated on the test split. Average precision is used for performance evaluation.

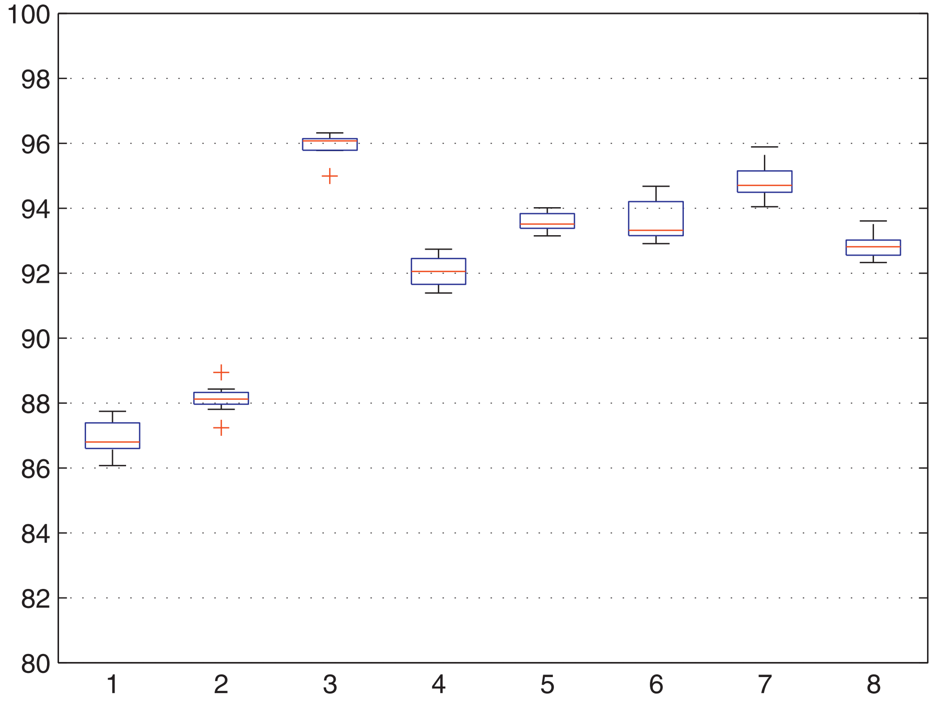

**Fig. 4.** Boxplot of the per-class performance of EISR-CNN on the UIUC-Sports dataset. From left to right: badminton, bocce, croquet, polo, rock climbing, rowing, sailing and snow boarding.

**Table 4**
Performance comparisons on the PASCAL VOC 07 dataset.

| object class | LLC [30] | Best07 [7] | FV [26] | DECAF [2] | CNN [2] | SPM [24] | OCP [24] | EISR-SC | EISR-CNN |
|---|---|---|---|---|---|---|---|---|---|
| airplane | 74.8 | 77.5 | 80.0 | 87.4 | 95.3 | 72.5 | 74.2 | 76.5 | 96.8 |
| bicycle | 65.2 | 63.6 | 67.4 | 79.3 | 90.4 | 56.3 | 63.1 | 67.2 | 92.2 |
| bird | 50.7 | 56.1 | 51.9 | 84.1 | 92.5 | 49.5 | 45.1 | 54.8 | 93.1 |
| boat | 70.9 | 71.9 | 70.9 | 78.4 | 89.6 | 63.5 | 65.9 | 73.3 | 91.7 |
| bottle | 28.7 | 33.1 | 30.8 | 42.3 | 54.4 | 22.4 | 29.5 | 33.5 | 57.3 |
| bus | 68.8 | 60.6 | 72.2 | 73.7 | 81.9 | 60.1 | 64.7 | 71.4 | 83.5 |
| car | 78.5 | 78.0 | 79.9 | 83.7 | 91.5 | 76.4 | 79.2 | 80.7 | 92.8 |
| cat | 61.7 | 58.8 | 61.4 | 83.7 | 91.9 | 57.5 | 61.4 | 62.5 | 93.3 |
| chair | 54.3 | 53.5 | 56.0 | 54.3 | 64.1 | 51.9 | 51.0 | 57.8 | 68.5 |
| cow | 48.6 | 42.6 | 49.6 | 61.9 | 76.3 | 42.2 | 45.0 | 52.2 | 79.4 |
| table | 51.8 | 54.9 | 58.4 | 70.2 | 74.9 | 48.9 | 54.8 | 53.1 | 77.6 |
| dog | 44.1 | 45.8 | 44.8 | 79.5 | 89.7 | 38.1 | 45.4 | 46.8 | 90.2 |
| horse | 76.6 | 77.5 | 78.8 | 85.3 | 92.2 | 75.1 | 76.3 | 78.8 | 92.7 |
| motorbike | 66.9 | 64.0 | 70.8 | 77.2 | 86.9 | 62.8 | 67.1 | 69.2 | 88.1 |
| person | 83.5 | 85.9 | 85.0 | 90.5 | 95.2 | 82.9 | 84.4 | 85.4 | 95.8 |
| plant | 30.8 | 36.3 | 31.7 | 51.1 | 60.7 | 20.5 | 21.8 | 33.5 | 63.9 |
| sheep | 44.6 | 44.7 | 51.0 | 73.8 | 82.9 | 38.1 | 44.3 | 46.4 | 84.2 |
| sofa | 53.4 | 50.9 | 56.4 | 57.0 | 68.0 | 46.0 | 48.8 | 55.8 | 71.4 |
| train | 78.2 | 79.2 | 80.2 | 86.4 | 95.5 | 71.7 | 70.7 | 82.3 | 96.3 |
| tv | 53.5 | 53.2 | 57.5 | 68.0 | 74.4 | 50.5 | 51.7 | 57.4 | 76.8 |
| mAP | 59.3 | 59.4 | 61.7 | 73.4 | 82.4 | 54.3 | 57.2 | 61.9 | 84.3 |

We give the performance comparisons on the PASCAL VOC 2007 dataset in Table 4. Since the search base strategy has been proven useful on the other datasets, we only give the performances of EISR-SC and EISR-CNN in Table 4. We can see from Table 4 that EISR is again able to improve the classification performances over many methods. For example, when local features are used, EISR-SC improves over other sparse coding based methods [7,26,30]. When the CNN based strategy is used, the performance can be further improved. This is because the resulting explicitly semantic representations are more discriminative than sparse coding based methods. The final representations are more separable which helps to improve the average precision. Besides, the improvements of EISR over non-rigid objects are larger than rigid objects. We believe this is for two reasons. First, non-rigid objects are more difficult to classify using visual features but can be better represented with

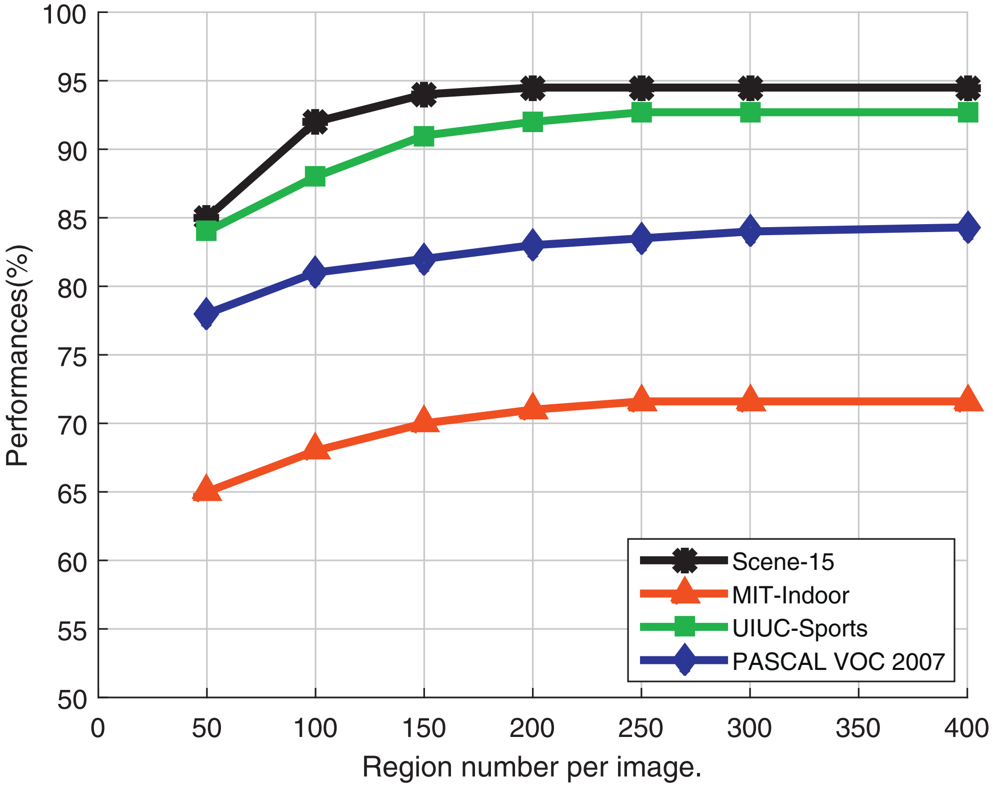

**Fig. 5.** Influences of region number on the Scene-15 dataset, the MIT-Indoor dataset, the UIUC-Sports dataset and the PASCAL VOC 2007 dataset.

semantics. Second, by randomly selecting image regions, we can model the correlations of different parts of images better and help to model the inter-class variation.

### 4.6. Influences of region selection

We give the influences of region number and the size of regions in Figs. 5 and 6 respectively. We can see from Fig. 5 that different region numbers have varied influences on the performances. If the region number is too small, we may not be able to fully combine the discriminative information. This problem can be alleviated with the increment of region number. Besides, different datasets require different region numbers. The PASCAL VOC 2007 dataset and MIT-Indoor dataset are more difficult to classify than the Scene-15 dataset and the UIUC-Sports dataset. From Fig. 6, we can see that the size of image region is not as important as the region number as long as the region is not too small. This is because we can compensate the size problem by selecting more regions. However, if we do not select enough regions, the increment of region size cannot achieve satisfiable performances. We can have the conclusion that we should select more image regions to improve the classification performances.

## 5. Conclusion

In this paper, we proposed an image classification method by search with explicitly and implicitly semantic representations. We first randomly selected image regions as the basic element for representation to make use of the spatial layouts of images. The explicitly semantic representation was then obtained using the learned semantic models which combined the discriminative information of training images. Besides, the implicitly semantic representation was collected by measuring the similarities between each image region and the Internet images. The corresponding text was used for implicitly semantic representations. Finally, we combined both the explicitly and implicitly semantic representations for joint representation. The bi-linear classifier was trained for prediction. We evaluated the proposed method on the Scene-15 dataset, the MIT-Indoor dataset, the UIUC-Sports dataset and the PASCAL VOC 2007 dataset with the results proved the effectiveness of the proposed method.

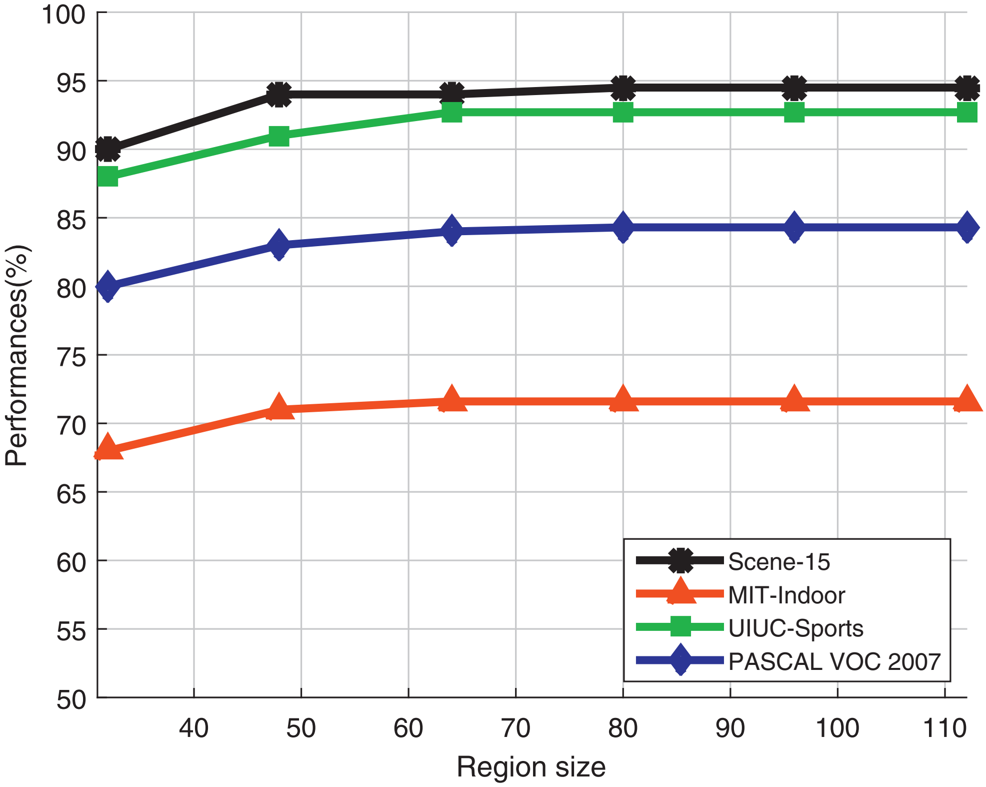

**Fig. 6.** Influences of region size on the Scene-15 dataset, the MIT-Indoor dataset, the UIUC-Sports dataset and the PASCAL VOC 2007 dataset.

## Acknowledgments

This work is supported by National Natural Science Foundation of China: 61303154.

## References

[1] A. Bosch, A. Zisserman, X. Munoz, Scene classification using a hybrid generative/discriminative approach, IEEE Trans. Pattern Anal. Mach. Intell. 30 (4) (2008) 712–727.
[2] K. Chatfield, K. Simnyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: delving deep into convolutional nets, in: British Machine Vision Conference, 2014, pp. 1–12.
[3] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[4] T. Darrell, I. Essa, A. Pentland, Task-specific gesture analysis in real-time using interpolated views, IEEE Trans. Pattern Anal. Mach. Intell. 18 (2) (1996) 1236–1242.
[5] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, N. Vasconcelos, Scene classification with semantic fisher vectors, in: Proceedings of Computer Vision and Pattern Recognition, 2015, pp. 2974–2983.
[6] C. Doersch, A. Gupta, A. Efros, Mid-level visual element discovery as discriminative mode seeking, in: Proceedings of Advances in Neural Information Processing Systems, 2013, pp. 494–502.
[7] M. Everingham, A. Zisserman, C. Williams, L.V. Gool, The PASCAL visual object classes challenge 2007 (VOC 2007) results, Pascal Challenge, 2007 Technical report.
[8] S. Gao, I. Tsang, L. Chia, Laplacian sparse coding, hypergraph laplacian sparse coding, and applications, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 35, 2013, pp. 92–104.
[9] S. Gao, I. Tsang, Y. Ma, Learning category-specific dictionary and shared dictionary for fine-grained image classification, IEEE Trans. Image Process. 23 (2) (2014) 623–634.
[10] J. Gemert, C. Veenman, A. Smeulders, J. Geusebroek, Visual word ambiguity, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1271–1283.
[11] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, in: Proceedings of European Conference on Computer Vision, 2014, pp. 392–407.
[12] C. Hong, J. Yu, J. You, X. Chen, D. Tao, Multi-view ensemble manifold regularization for 3d object recognition, Inf. Sci. 320 (2015) 395–405.
[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[14] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Proceedings of Computer Vision and Pattern Recognition, 2006, pp. 2169–2178.
[15] L. Li, L. Fei-Fei, What, where and who? classifying event by scene and object recognition, in: Proceedings of International Conference on Computer Vision, 2007, pp. 1–8.
[16] L. Li, H. Su, E. Xing, L. Fei-Fei, Objectbank: A high-level image representation for scene classification & semantic feature sparsification, in: Proceedings of the Neural Information Processing Systems, Vancouver, Canada, 2010, pp. 20–39.

[17] D. Lin, C. Lu, R. Liao, J. Jia, Learning important spatial pooling regions for scene classification, in: Proceedings of Computer Vision and Pattern Recognition, 2014, pp. 3726–3733.
[18] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2004) 91–110.
[19] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: Proceedings of Computer Vision and Pattern Recognition, 2009, pp. 413–420.
[20] C. Rao, A. Yilmaz, M. Shah, View-invariant representation and recognition of actions, Int. J. Comput. Vision 50 (2) (2002) 203–226.
[21] N. Rasiwasia, N. Vasconcelos, Holistic context models for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (5) (2012) 902–917.
[22] N. Rasiwasia, N. Vasconcelos, Scene classification with low-dimensional semantic spaces and weak supervision, in: Proceedings of Computer Vision and Pattern Recognition, Alaska, USA, 2008, pp. 1–6.
[23] A. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, in: Proceedings of Computer Vision and Pattern Recognition, 2014, pp. 512–519.
[24] O. Russakovsky, Y. Lin, K. Yu, L. Fei-Fei, Object-centric spatial pooling for image classification, in: Proceedings of European Conference on Computer Vision, 2012, pp. 1–15.
[25] F. Sadeghi, M. Tappen, Latent pyramidal regions for recognizing scenes, in: Proceedings of European Conference on Computer Vision, 2012. Pp.228 C241
[26] J. Sanchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: theory and practice, Int. J. Comput. Vision 105 (3) (2013) 222–245.
[27] K. Sande, T. Gevers, C. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1582–1596.
[28] L. Torresani, M. Szummer, A. Fitzgibbon, Efficient object category recognition using classesmes, in: Proceedings of European Conference of Computer Vision, Crete, Greece, 2010, pp. 776–789.
[29] D. Wang, S. Hoi, Y. He, J. Zhu, Mining weakly labeled web facial images for search-based face annotation, IEEE Trans. Knowl. Data Eng. 26 (1) (2014) 166–179.
[30] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Proceedings of Computer Vision and Pattern Recognition, 2010, pp. 3360–3367.
[31] X. Wang, L. Zhang, F. Jing, W. Ma, Annosearch: Image auto-annotation by search, in: Proceedings of Computer Vision and Pattern Recognition, 2006, pp. 1483–1490.
[32] X. Wang, L. Zhang, W. Ma, Duplicate-search-based image annotation using web-scale data, Proc. IEEE 100 (9) (2012) 2705–2721.
[33] J. Wu, J. Rehg, Centrist: a visual descriptor for scene categorization, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1489–1501.
[34] L. Wu, S. Hoi, N. Yu, Semantics-preserving bag-of-words, models and applications, IEEE Tran. Multimedia 19 (7) (2010) 1908–1920.
[35] Y. Xiao, Z. Zhu, Y. Zhao, Y. Wei, S. Wei, Kernel reconstruction ICA for sparse representation, IEEE Trans. Neural Netw. Learn. Systems 26 (6) (2015) 1222–1232.
[36] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of Computer Vision and Pattern Recognition, USA,, 2009, pp. 1794–1801.
[37] J. Yu, Y. Rui, Y. Tang, D. Tao, High-order distance based multiview stochastic learning in image classification, IEEE Trans. Cybern. 44 (12) (2014) 2431–2442.
[38] J. Yu, D. Tao, J. Li, J. Cheng, Semantic preserving distance metric learning and applications, Inf. Sci. 281 (2014) 674–686.
[39] C. Zhang, J. Cheng, J. Liu, J. Pang, C. Liang, Q. Huang, Q. Tian, Object categorization in sub-semantic space, Neurocomputing 142 (2014) 248–255.
[40] C. Zhang, J. Cheng, Y. Zhang, J. Liu, C. Liang, J. Pang, Q. Huang, Q. Tian, Image classification using boosted local features with random orientation and location selection, Inf. Sci. 310 (2015) 118–129.
[41] C. Zhang, J. Liu, C. Liang, Q. Huang, Q. Tian, Image classification using harr-like transformation of local features with coding residuals, Signal Process. 93 (8) (2013) 2111–2118.
[42] C. Zhang, J. Liu, C. Liang, Z. Xue, J. Pang, Q. Huang, Image classification by non-negative sparse coding, correlation constrained low-rank and sparse decomposition, Comput. Vision Image Understanding 123 (2014) 14–22.
[43] C. Zhang, J. Liu, Q. Tian, Y. Han, H. Lu, S. Ma, A boosting, sparsity-constrained bilinear model for object recognition, IEEE Multimedia 19 (2) (2012) 58–68.
[44] C. Zhang, J. Liu, Q. Tian, C. Liang, Q. Huang, Beyond visual features: a weak semantic image representation using exemplar classifiers for classification, Neurocomputing 120 (2013) 318–324.
[45] C. Zhang, Z. Xue, X. Zhu, H. Wang, Q. Huang, Q. Tian, Boosted random contextual semantic space based representation for visual recognition, Inf. Sci. (2016), doi:10.1016/j.ins.2016.06.029.
[46] T. Zhang, B. Ghanem, S. Liu, C. Xu, N. Ahuja, Low-rank sparse coding for image classification, in: Proceedings of International Conference on Computer Vision, 2013, pp. 281–288.
[47] W. Zhao, X. Wu, C. Ngo, On the annotation of web videos by efficient near-duplicate search, IEEE Trans. Multimedia 12 (5) (2010) 448–461.
[48] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, D. Cai, Graph regularized sparse coding for image representation, IEEE Trans. Image Process. 20 (5) (2011) 1327–1336.
[49] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Proceedings of Advances in Neural Information Processing Systems, 2014, pp. 487–495.
[50] X. Zhou, K. Yu, T. Zhang, T. Huang, 2010, Image classification using super-vector coding of local image descriptors. Proceedings of European Conference on Computer Vision, 141–154