Neurocomputing 120 (2013) 318–324

Contents lists available at ScienceDirect

Neurocomputing

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

Beyond visual features: A weak semantic image representation using exemplar classifiers for classification

Chunjie Zhang [a], Jing Liu [b,*], Qi Tian [c], Chao Liang [d], Qingming Huang [a]

[a] Graduate University of Chinese Academy of Sciences, 100049 Beijing, China
[b] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
[c] Department of Computer Sciences, University of Texas at San Antonio, TX 78249, USA
[d] National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, 430072 Wuhan, China

ARTICLE INFO

Article history:
Received 11 December 2011
Received in revised form
26 June 2012
Accepted 23 July 2012
Available online 29 March 2013

Keywords:
Image classification
Exemplar classifier
Weak semantic representation
Structured sparsity

ABSTRACT

Usually, the low-level representation of images is unsatisfied for image classification due to the well-known semantic gap, and further hinders its application for high-level visual applications. To deal with these problems, in this paper, we propose a simple but effective image representation for image classification, which is denoted as the responses to a set of exemplar image classifiers. Each exemplar classifier corresponding to a training image is learned using SVM algorithm to distinguish the image from others in different classes, and hence exhibits some discriminative information, which can also be regarded as a kind of weak semantic meaning. In such a one-vs-all manner, we can obtain the exemplar classifiers for all training images. We then train a linear classifier with structured sparsity constraints for each image category by taking advantages of the weak semantic image representation. Experiments on several public datasets demonstrate the effectiveness of the proposed method.

1. Introduction

Automatically classifying image based on its semantic content remains a very challenging problem in computer vision. Recently, the use of local features for image classification has become very popular and has been shown very effective. Typically, local features are first encoded with visual words by nearest neighbor assignment [1] or sparse coding [2]. Images are then represented by the occurrence histogram of visual words. The bag-of-visual-words (BoW) representation is inspired by the bag-of-words approach to text categorization [3]. However, due to the semantic gap, there are no explicit semantic correspondences between visual words and human perception which hinders its discriminative power for high-level visual applications.

To attack this drawback, researchers have done a lot of work which dramatically improves the image classification performance. On one hand, more discriminative and sophisticated models [4–10] are proposed. These models combine different types of visual information (*e.g.*, spatial and contextual information) for better classification. With constant updating of computation capabilities, the design of more sophisticated models will still be an important and potential solution in future, but it is beyond the scope of this paper.

On the other hand, the use of semantically meaningful space for image classification has attracted the attention of researchers [11–21], which is our focus in this paper. The semantic spaces can be generated by psychophysical experiments [11,12], latent space learning [13–15] or using the training image concepts as well as generic objects [16–21]. The use of semantic space makes image representation more interpretable than using visual features. This is often achieved by using a set of pre-learnt classifiers or object detectors. However, due to the semantic gap, except for a few objects (*e.g.*, "face"), it is still very hard to learn effective classifiers or detectors for generic image classes. Furthermore, some objects often exhibit visual polysemy (*e.g.*, a functional object like "container") or view-dependent (*e.g.*, a side-view or a frontal-view car). Thus, it is hard or even impossible to learn a single object classifier competent for various functional objects and different views.

To address above problems, we present a weak semantic image representation using the responses of some learnt exemplar classifiers. The exemplar classifier is specific to each exemplar image in a training set, and it is trained to distinguish the exemplar image from the others in different image classes. Although an exemplar image may offer a local reflection of its corresponding class, many ones in the class are mutually complementary and jointly present a comprehensive description. The response to each exemplar classifier can be deemed as an explicit representation about the image class, namely a kind of weak

* Corresponding author. Tel.: +86 1062632267.
E-mail addresses: ivazhangchunjie@gmail.com (C. Zhang), jliu@nlpr.ia.ac.cn (J. Liu), qitian@cs.utsa.edu (Q. Tian), liangchao827@gmail.com (C. Liang), qmhuang@jdl.ac.cn (Q. Huang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.07.056

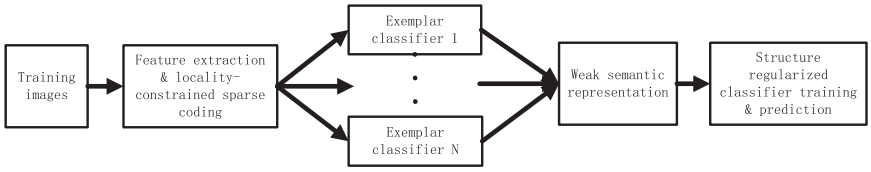


Fig. 1. Flowchart of the proposed image classification using weak semantic representation using exemplar classifiers with sparsity constraints.

semantic characteristics. We believe that the use of exemplar images offers a promising way to enhance the performance of image classification, because it presents a higher-level representation than traditional visual words. Specifically, we perform as follows. First, we explore the linear SVM algorithm to train the exemplar classifiers corresponding to each image in training set, while the spatial pyramid matching (SPM) approach for image representation [5] is used to leverage the spatial and scale information. Second, with the learnt exemplar classifiers, each image can be described as a vector of classifiers' outputs, i.e., a so-called weak semantic representation in this paper. In the following, we train a linear classifier with structured sparsity constraints to choose the most discriminative components of the weak semantic representation for efficient image classification. We perform experiments on several public datasets, and the results demonstrate the encouraging performance of the proposed method over the state-of-the-arts. Fig. 1 shows the flowchart of the proposed method.

The rest of this paper is organized as follows. The related work are given in Section 2. Section 3 presents the proposed weak semantic representation of images using exemplar classifiers. In Section 4, we classify images using the weak semantic image representation via structured regularized learning. We give the experimental results in Section 5 and finally conclude in Section 6.

2. Related work

In recent years, the bag-of-visual-words (BoW) model with local image features has become popular and been proven very effective for image classification. However, due to the well-known semantic gap, the lack of explicit correspondence between visual features and semantic concepts limits its discriminative power. Researchers have done a lot of work to alleviate this problem which can be broadly divided into two perspectives. Some tried to solve this problem by designing more sophisticated and discriminative models while others used more semantically meaningful space for image representation.

On one hand, the design of more sophisticated models [4–10] tried to combine different types of visual information (*e.g.*, spatial and contextual information) for better visual applications. Grauman and Darrell [4] proposed pyramid matching to take advantage of local feature's information in feature space. Motivated by this, Lazebnik et al. [5] proposed the spatial pyramid matching (SPM) method to combine the spatial information of local features and was widely used since its introduction. Shape context was proposed by Belongie et al. [6] which represents a binary shape as a discrete set of points sampled from its contour. These points are mapped into a log-polar coordinate system centered at a reference point and each bin of the log-polar space is then determined by the distance and angle intervals. Dalal and Triggs [7] used histograms of oriented gradients (HOG) for human detection which can be accelerated by using cell-based interpolation [8] and integral image [9] techniques. Felzenszwalb et al. [10] proposed a part based model for efficient object detection and was widely used by researchers. The more sophisticated one model is, the more computational power it needs. Although the fast development of computer helps to alleviate this problem, the careful design of efficient models is still a very challenging problem.

On the other hand, the use of semantic space for image representation has also been widely used by researchers [11–20]. Some researchers used semantic space determined through psychophysical experiments [11,12]. Oliva et al. [11] organized images with three semantic axes which are determined by psychophysical experiments while Mojsilovic et al. [12] used 20 concepts. Other researchers tried to learn the latent semantic space by text analysis methods such as probabilistic Latent Semantic Analysis [13] (pLSA) and Latent Dirichlet Allocation [14] (LDA). However, this latent space modeling is hard to interpret for humans. Pang et al. [15] extended the pLSA model by extracting both global and local topics which is then used for tourist destination summarization. To obtain more explicit semantic space, the training concepts as well as generic objects are also used [16–21]. Rasiwasia and Vasconcelos [16] proposed to learn a low dimensional semantic "theme" from casual image annotations for scene classification. Hauptmann et al. [17] studied the influence of the number of high-level concepts for reliable video retrieval. Rasiwasia et al. [18] represented each image with respect to the response of a set of visual concept classifiers and applied it for image retrieval. Vogel and Schiele [19] proposed to use the concept-occurrence vector (COV) for semantic modeling of natural scenes. Images are first divided into regions and the categories of these regions are then predicted. The normalilzed histogram of the concept occurrences in an image is then used to represent this image. Torresani et al. [20] used the whole images for generic object classifier training without considering the location and scale changes of objects. To model the object detectors more efficiently, Li et al. [21] proposed the ObjectBank which learnt generic object detectors using the images as well as the human labeled object bounding boxes using the LabelMe dataset [22] and the ImageNet dataset [23]. However, due to the semantic gap, it is often hard and time consuming to train efficient classifiers or detectors for generic image classes. To alleviate this problem, the use of exemplar training data is also used by researchers [24,25]. Malisiewicz and Efros [24] used the segmented per-exemplar image to learn distances for object recognition by association. Malisiewicz et al. [25] proposed a conceptually simple method by combining the effectiveness of a discriminative object detector with the explicit correspondence offered by nearest neighbor approach for efficient object detection with good performance.

Another approach that is related to our work is the use of attributes [26–29]. Farhadi et al. [26] described object categories by a set of boolean attributes such as "has ears", "near water" and built the attribute classifiers by using the internet resources. Lampert et al. [27] used the attribute information to detect unseen object classes with transfer learning. Parikh and Grauman [28] tried to build a discriminative nameable attributes vocabulary with humans in the loop. To distinguish the discriminative power of each attribute for different images, Parikh and Grauman [29] learnt a ranking function per attribute to predict the relative strength of each property for images. The use of attributes helps to boost the performance of visual applications. However, the attributes have to be pre-defined which limits its efficiency for

large scale visual applications, besides, it also labor intensive and requires experiences to define proper attributes for different visual application tasks.

The nearest neighbor based approach and its generalized form are also widely used [30–32]. Boiman et al. [30] proposed a Naive–Bayes Nearest-Neighbor (NBNN) classifier which employs the nearest neighbor distances of the local feature space without feature quantization. This simple NBNN method required no training time and achieved the state-of-the-art performance on several public datasets. Wright et al. [31] used the sparse representation technique for robust face recognition which assigned images of the class with the lowest reconstruction distances while Yuan and Yan [32] proposed a multi-task joint sparse representation to combine the discriminative power of different types of features. The proposed method shares some similarities with these nearest neighbor based methods but is fundamentally different. Since we discriminatively train each exemplar classifier, we can have more freedom in deciding the decision boundary hence are able to generalize much better. Besides, instead of directly using the distances for classification, we use the outputs of these exemplar classifiers with sparsity constraints for better classification via fitting a regularized logistic regression model along with spatial pyramid matching.

3. Exemplar classifier based weak semantic image representation

We use the semantic space technique to represent images. Each image is represented by the response of a set of learnt classifiers. The histogram of visual word occurrences with spatial pyramid matching ($L=0,1,2$) is used as the initial image representation [5]. Instead of learning a single classifier for each class, we try to learn a set of exemplar classifiers for all the training images. Each exemplar classifier is trained with the corresponding training images and all the other images of different classes hence exhibit weak semantic meanings. Since this is much easier than classifying the full-class images, we can use simpler classifiers such as linear SVM. Instead of using these exemplar classifiers for classification directly, we use the response of these exemplar classifiers for image representation and then train classifiers for final prediction.

Formally, let $X=[x_1,\ldots,x_N]\in\mathbb{R}^{D\times N}$ be the set of D-dimensional BoW representation of N images, where $x_i\in\mathbb{R}^{D\times 1}$, $i=1,\ldots,N$. These images are of K classes and let $Y=(y_1,\ldots,y_N)\in\{1,\ldots,K\}^N$ denote the corresponding image labels. For each training image x_i, $i=1,2,\ldots,N$, we try to learn the optimal parameters (w_i,b_i) to separate x_i from all the other images of different classes by the largest possible margin, where $w_i\in\mathbb{R}^{D\times 1}$. This is achieved by solving the following optimization problem for all i as

$$\min_{w_i,b_i}\|w_i\|^2+C\times\ell(w_i^Tx_i+b_i)+\sum_{j=1}^{N}\ell(-w_i^Tx_j-b_i) \quad (1)$$

$$\forall y_j\neq y_i$$

where C is the weighting parameter which controls the relative importance of x_i. We use the hinge loss which is widely used by researchers as our loss function. The hinge loss has the form of

$$\ell(x)=\max(0,1-x) \quad (2)$$

We use libsvm [33] to train each exemplar classifier. After all the exemplar classifiers are trained, we can use it for weak semantic image representation. Since each exemplar is trained to only give high predicted values for visual similar samples of the corresponding image, it exhibits weak semantic information that an image belongs to a particular class. For a given image x, we predict its semantic meanings for each exemplar classifier and use the output of these classifiers as the final image representation $h\in\mathbb{R}^{N\times 1}$, where $h_i=w_i^Tx+b_i$, $i=1,2,\ldots,N$. The spatial pyramid matching technique (SPM) with three pyramid levels ($L=0,1,2$) is also used to combine the spatial information and scale changes of this weak semantic representation.

Although the discriminative power of each exemplar classifier is limited, together they can cope with different inter- and intra-class variations more efficiently and effectively than the full-class classifiers, hence is able to represent images better and boost the final image classification performances. The proposed method bears some similarities with the bagging technique [34] in machine learning. By combining a set of weak classifiers, the bagging technique can produce much more powerful classifier than training a single classifier.

The proposed exemplar classifier based image representation is different from nearest-neighbor based methods because each classifier is discriminatively trained. Each exemplar classifier has more freedom to define decision boundaries hence generalizes much better than nearest-neighbor based methods with less training data. Besides, we use the response of exemplar classifiers for image representation and train classifiers for final classification.

4. Weak semantic image representation for classification via structure regularized learning

After representing each image with the weak semantic representation, we can predict the categories of images by training classifiers. Let $H=[h_1,h_2,\ldots,h_N]\in\mathbb{R}^{D'\times N}$ be the weak semantic representation of N training images and $Y=(y_1,\ldots,y_N)\in\{1,\ldots,K\}^N$ denote the corresponding image labels, where $h_i\in\mathbb{R}^{D'\times 1}$, $i=1,\ldots,N$. Our aim is to learn K linear functions α_k^Th, $\alpha_k\in\mathbb{R}^{D'\times 1}$, $k=1,\ldots,K$, such that the label of image h is decided by

$$y=\underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}}\ \alpha_k^Th \quad (3)$$

We follow the one-versus-all strategy to learn K binary linear classifiers by solving the following optimization problem as

$$\min_{\alpha_k}\sum_{i=1}^{N}L(\alpha_k^Th_i,y_i^k)+\lambda R(\alpha_k),\quad \forall k \quad (4)$$

where $y_i^k=1$, if $y_i=k$, otherwise $y_i^k=-1$. $L(\cdot)$ is the loss function and $R(\alpha_k)$ is the regularization term. λ is the balancing parameter of the two terms whose value can be determined by cross validation. In this paper, we use *Log* loss which has the form of

$$L(\alpha_k^Th_i,y_i^k)=\log(Z/\exp(0.5y_i^k\times\alpha_k^Th_i)) \quad (5)$$

The log loss is widely used by researchers both for its good performance and differentiability.

A proper regularization term is very important for robust image classification. Since we train exemplar classifiers for each training image and as the number of training images increases, it would be more effective to choose the most discriminative exemplar classifiers instead of using all of them equally. A popular choice is to use the sparsity constraints as $R(\alpha_k)=\|\alpha_k\|_1$ for exemplar classifier selection.

Besides, although each exemplar classifier is trained separately, images of the same class are often correlated which means the corresponding exemplar classifiers should also have some prediction consistency. To take this information into consideration, we set the regularization term in a structured form by joint ℓ_1/ℓ_2 regularization as $R(\alpha_k)=\|\alpha_k\|_{1,2}=\sum_{j=1}^{K}\|\alpha_k^j\|_2$, where α_k^j is the parameters corresponding to exemplar classifiers of the jth class. This regularization term encourages parameters corresponding to

the same class to be jointly zero. Such structured sparsity is more robust and semantically meaningful than using individual sparsity constraints whose effectiveness has been proven by many researchers [21,35]. Moreover, since we use the SPM technique to alleviate the scale and location changes of objects, it would be more effective to select the optimal scale and location of objects instead of treating them equally. We add a ℓ_1 to the structured regularization function and the final regularizer used in this paper has the form of

$$R(\alpha_k) = \|\alpha_k\|_{1,2} + \lambda_1\|\alpha_k\|_1 \tag{6}$$

where λ_1 is a balancing parameter which can be determined by cross validation. This makes the final optimization problem still convex. To solve this problem, we use the coordinate descent algorithm proposed by Li et al. [21] to learn the optimal parameters α_k, $\forall k$. After all the parameters are learnt, we can predict the classes of images using Eq. (3).

5. Experiments

We evaluate the proposed method for image classification on several public datasets: the Scene-15 dataset [5], the Caltech-256 dataset [36] and the MIT Indoor dataset [37]. We densely extract SIFT descriptors [38] on overlapping 16×16 pixels with an overlap of 6 pixels. Sparse coding [2] with locality constraints [39] is used to encode local features as it has been proven more effective than k-means clustering method for image classification. Max pooling is then used to extract image representation which is used for training exemplar classifiers. The codebook size is set to 1024 for the three datasets. We found our method is robust to a wide range of Cs. This is because we did not use the outputs of exemplar classifiers for final prediction but use them as weak semantic image representation. The final performance is relatively stable as long as the outputs of exemplar classifiers are consistent.

5.1. Scene-15 dataset

The Scene-15 dataset has 15 categories with a total of 4485 images and ranges from natural scenes like mountains and forest to man-made environments like store and living room. Each class of the Scene-15 dataset has 200–400 images. The average image size is 300×250 pixels. For fair comparison, we randomly choose 100 training images per category and use the rest images for test, as did in [2,5] and repeat this process for six times. We report our final results by the mean and standard deviation of the average of per-class classification rates.

Fig. 2 shows some example images with the top five images whose corresponding exemplar classifiers output the largest responses (in descending order). Each exemplar classifier only needs to classy visually similar images and we use the joint representation of the outputs of these exemplar classifiers for better image classification. We give the performance comparison of the proposed method with [2,5,16,21,40] in Table 1. We give the results of the proposed method with no regularization on the classifier parameters, using k-means clustering and sparse coding for local feature encoding respectively. Note that the WSR-EC(no regularization) result is achieved by using k-means clustering and nearest neighbor based local feature quantization. We also give the re-implemented results of KSPM by Yang et al. [2]. We can see from Table 1 that the proposed method achieves good performance which clearly demonstrates the effectiveness of the proposed method. The use of exemplar classifier for image representation makes it robust to inter and intra class variation than LSS [16] which used all the training samples, besides, we

Fig. 2. Example images with the top five images whose corresponding exemplar classifiers output the largest responses (in descending order for each row).

jointly choose the most discriminative parameters for final classification using structured sparsity constraints which helps to further improve the final classification performance.

Table 1
Performance comparison on the Scene-15 dataset. (ScSPM, sparse coding along with spatial pyramid matching; KSPM, spatial pyramid matching and kernel SVM classifier; LSS, low-dimensional semantic spaces with weak supervision; OB, object bank; KCSPM, kernel codebook and spatial pyramid matching; WSR-EC (no regularization/*k*-means/sparse coding): the proposed weak semantic image representation using exemplar classifiers with no regularization on classifier parameters/*k*-means clustering/sparse coding).

Algorithm	Performance
KSPM [2]	76.73 ± 0.65
ScSPM [2]	80.28 ± 0.93
KSPM [5]	81.40 ± 0.50
LSS [16]	72.20 ± 0.20
OB [21]	80.9
KCSPM [40]	76.70 ± 0.40
WSR-EC(no regularization)	74.19 ± 0.47
WSR-EC(*k*-means)	77.82 ± 0.63
WSR-EC(sparse coding)	$\mathbf{81.54 \pm 0.59}$

Table 2
Performance comparison on the Caltech-256 dataset. (Classemes: classification with weakly trained object classifiers based descriptor; NBNN, Naive–Bayes nearest-neighbor; LLC, locality-constrained linear coding.)

Algorithm	15 training	30 training	45 training
KSPM [2]	23.34 ± 0.42	29.51 ± 0.52	–
ScSPM [2]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55
Classemes [20]	–	36.00	–
OB [21]	–	39.00	–
NBNN(1 Desc) [30]	30.45	38.18	–
KSPM [36]	–	34.10	–
LLC [39]	34.36	41.19	45.31
KCSPM [40]	–	27.17 ± 0.46	–
WSR-EC	$\mathbf{35.28 \pm 0.65}$	$\mathbf{42.01 \pm 0.47}$	$\mathbf{45.82 \pm 0.54}$

We can see from Table 1 that the proposed WSR-EC (with sparse coding) has comparable performance with KSPM [5]. We believe this is because images of the Scene-15 dataset are relatively easy to be separated. However, with the increase of image classes, it becomes more difficult for KSPM to separate images correctly. The proposed WSR-EC method uses the discriminative power of each exemplar classifier and organized them more efficiently, hence improves image classification performance. We can see from Table 2 that the KSPM does not work as well as WSR-EC on the Caltech-256 dataset which has 256 classes.

Our method performs not as good as OB [21] with *k*-means clustering and nearest-neighbor assignment based feature quantization. This is because the OB algorithm also used human labeled images from other sources for efficient object detector training while we only make use of the training samples. However, by using sparse coding instead of *k*-means for local feature encoding, we are able to improve the performance. This also demonstrates the effectiveness of using sparse coding for image classification [2].

5.2. Caltech-256 dataset

The Caltech-256 dataset contains 256 categories of 29,780 images with high intra-class variability and object location variability. Each class of the Caltech-256 dataset has at least 80 images. We follow the experimental setting as [2,39] did and randomly choose 15, 30 and 45 images per class for training and use the rest of images for testing.

We give the performance comparison of the proposed method with other methods [2,20,21,30,36,38,40] in Table 2. The combinations of spatial location of objects and exemplar classifier based image representation make our algorithm perform better than the Classemes [20] by 6% and OB [21] by 3% for 30 training images per class. Besides, the use of locality-constrained sparse coding helps to increase the discriminative power of exemplar classifiers which also helps for the final performance improvements. The use of more discriminative exemplar classifiers helps to improve the final performance than less discriminative exemplar classifiers.

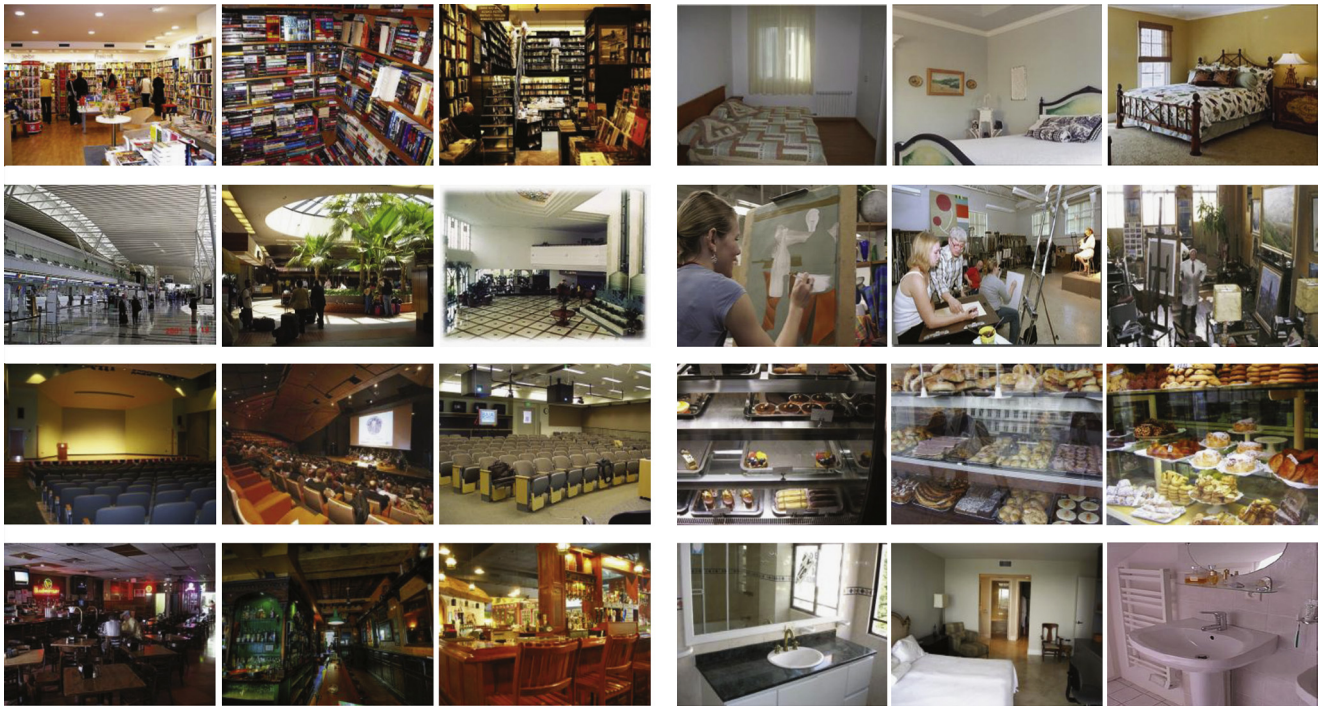


Fig. 3. Some example images of the MIT Indoor dataset.

Table 3
Sorted classification performance on the MIT Indoor dataset.

church inside 78.3	elevator 73.7	auditorium 63.7
concert hall 63.6	classroom 62.3	bowling 61.9
computer room 61.4	buffet 61.4	inside bus 59.5
greenhouse 58.8	corridor 58.0	dentaloffice 57.5
cloister 56.6	studiomusic 52.4	trainstation 51.8
locker room 49.7	closet 48.9	library 48.7
laundromat 47.7	tv studio 47.5	grocerystore 47.5
videostore 45.3	bathroom 44.5	florist 44.1
hospitalroom 43.7	garage 43.2	pantry 43.2
gameroom 42.2	stairscase 41.8	nursery 41.7
bookstore 40.8	deli 39.2	inside subway 39.1
kitchen 38.4	winecellar 37.8	poolinside 36.7
movietheater 35.8	clothingstore 35.0	gym 34.1
toystore 33.7	fastfood restaurant 33.6	livingroom 33.0
bar 31.9	dining room 30.7	casino 30.5
prison cell 29.3	hairsalon 29.0	artstudio 27.8
airport inside 26.3	waitingroom 25.4	subway 24.9
museum 24.4	lobby 24.0	bakery 23.1
bedroom 21.3	laboratorywet 20.8	restaurant 20.2
mall 20.2	meeting room 18.4	warehouse 17.6
operating room 16.8	office 16.1	restaurant kitchen 15.6
children room 14.3	kindergarden 12.6	jewelleryshop 12.2
shoeshop 11.9		

Moreover, the proposed method also outperforms the NBNN method which used the nearest-neighbor measurement with no quantization of local features.

5.3. MIT Indoor dataset

The MIT Indoor dataset has 67 indoor scenes of 15,620 images of different sources. All images have minimum 200 pixel resolution in the smaller axis. We follow the same experimental setup as did in [36] and use 80 images per class for classifier training and 20 images for testing. Fig. 3 shows some example images of the MIT Indoor dataset.

We achieved 38.6% on the MIT Indoor dataset which outperforms the OB (37.6%) algorithm and Classemes (26%) algorithm. Table 3 gives the detailed performance of WSR-EC. This again demonstrates the effectiveness of the proposed method. The use of locality-constrained sparse coding helps to encode local features more effectively than nearest neighbor assignment hence improves the discriminative power of exemplar classifiers. Besides, we also impose structured sparsity constraints on the weak semantic image representation for efficient classifier training.

6. Conclusions

This paper proposed a novel image classification model using weak semantic image representation with exemplar classifiers. We train exemplar classifiers for all the training images and use the outputs of these learnt exemplar classifiers for image representation. Since each exemplar classifier only concentrates on similar objects with this exemplar, the proposed image representation bears weak semantic information. To take advantages of this weak semantic representation, we train a logistic regression model with structured sparsity constraints to jointly choose the most discriminative components for efficient image classification. The spatial pyramid matching technique is also used to combine different image locations and scale changes. Experimental results on several public datasets demonstrate the effectiveness of the proposed method. Since exemplar classifiers are trained separately, it can be easily extended in a parallel way for large scale visual applications.

Our future work concentrates on the following two aspects. First, the use of internet resources for efficient exemplar classifier training will be studied. Second, how to combine different types of features efficiently to boost the performance will also be investigated.

Acknowledgement

This work is supported by National Basic Research Program of China (973 Program):2010CB327905, National Natural Science Foundation of China: 61025011, 61272329, the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), China Postdoctoral Science Foundation: 2012M520434.

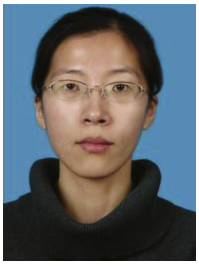
References

[1] J. Sivic, A. Zisserman, Video Google: a text retrieval approach to object matching in videos, in: Proceedings of the Ninth International Conference on Computer Vision, Nice, France, 2003, pp. 1470–1477.
[2] 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, Miami, USA, 2009.
[3] G. Salton, M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
[4] K. Grauman, T. Darrell, The pyramid match kernel: discriminative classification with sets of image features, in: Tenth Proceedings of International Conference on Computer Vision, Beijing, China, 2005.
[5] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the Computer Vision and Pattern Recognition, New York, USA, 2006, pp. 2169–2178.
[6] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522.
[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the Computer Vision and Pattern Recognition, Beijing, China, 2005.
[8] Y. Pang, Y. Yuan, X. Li, J. Pan, Efficient HOG human detection, Signal Process. 91 (4) (2011) 773–781.
[9] Y. Pang, H. Yan, Y. Yuan, K. Wang, Robust CoHOG feature extraction in human-centered image/video management system, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 42 (2) (2012) 458–468.
[10] P.F. Felzenszwalb, R.B. Girschick, D. McCallester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[11] A. Oliva, A. Torralba, A. Guerin-Dugue, J. Herault, Global semantic classification of scenes using power spectrum templates, in: Challenge of Image Retrieval, Newcastle, UK, 1999.
[12] A. Mojsilovic, J. Gomes, B. Rogowitz, Semantic-friendly indexing and querying of images based on the extraction of the objective semantic cues, Int. J. Comput. Vis. 56 (1) (2004) 79–107.
[13] Thomas Hofmann, Probabilistic latent semantic analysis, in: Proceedings of Uncertainity in Artificial Intelligence, Stockholm, Sweden, 1999.
[14] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (1) (2003) 993–1022.
[15] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, L. Zhang, Summarizing tourist destinations by mining user-generated travelogues and photos, Comput. Vis. Image Understand. 115 (3) (2011) 352–363.
[16] N. Rasiwasia, N. Vasconcelos, Scene classification with low-dimensional semantic spaces and weak supervision, in: Proceedings of the Computer Vision and Pattern Recognition, Alaska, USA, 2008.
[17] A. Hauptmann, Rong Yan, W. Lin, M. Christel, H. Wactlar, Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news, IEEE Trans. Multimedia 9 (5) (2007) 958–966.
[18] N. Rasiwasia, P. Moreno, N. Vasconcelos, Bridging the gap: query by semantic example, IEEE Trans. Multimedia 9 (5) (2007) 923–938.
[19] J. Vogel, B. Schiele, Semantic modeling of natural scenes for content-based image retrieval, Int. J. Comput. Vis. 72 (2) (2007) 133–157.
[20] L. Torresani, M. Szummer, A. Fitzgibbon, Efficient object category recognition using classesmes, in: Proceedings of the European Conference of Computer Vision, Crete, Greece, 2010.
[21] L. Li, H. Su, E. Xing, Li 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.
[22] B. Russell, A. Torralba, K. Murphy, W. Freeman, Labelme: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (3) (2008) 157–173.
[23] J. Deng, W. Dong, R. Socher, L. Li, K. Li, Li Fei-Fei, ImageNet: a large-scale hierarchical image database, in: Proceedings of the Computer Vision and Pattern Recognition, Florida, USA, 2009.
[24] T. Malisiewicz, A. Efros, Recognition by association via learning per-exemplar distances, in: Proceedings of the Computer Vision and Pattern Recognition, Alaska, USA, 2008.

[25] T. Malisiewicz, A. Gupta, A. Efros, Ensemble of exemplar-SVMs for object detection and beyond, in: Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 2011.
[26] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Proceedings of the Computer Vision and Pattern Recognition, Florida, USA, 2009.
[27] C. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: Proceedings of Computer Vision and Pattern Recognition, Florida, USA, 2009.
[28] D. Parikh, K. Grauman, Interactively building a discriminative vocabulary of nameable attributes, in: Proceedings of the Computer Vision and Pattern Recognition, Colorado, USA, 2011.
[29] D. Parikh, K. Grauman, Relative attributes. in: Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 2011.
[30] O. Boiman, E. Shechtman, M. Irani, In defense of nearest-neighbor based image classification, in: Proceedings of the Computer Vision and Pattern Recognition, Alaska, USA, 2008.
[31] J. Wright, A. Yang, A. Ganesh, S. Satry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[32] X. Yuan, S. Yan, Visual classification with multi-task joint sparse representation, in: Proceedings of Computer Vision and Pattern Recognition, San Francisco, USA, 2010.
[33] C. Chang, C. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011).
[34] Leo Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[35] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 68 (1) (2006) 49–67.
[36] G. Griffin, A. Holub, P. Perona, Caltech-256 Object Category Dataset, Technical Report, CalTech, 2007.
[37] A. Quattoni, A. Torralba, Recognizing Indoor Scenes, in: Proceedings of the Computer Vision and Pattern Recognition, Florida, USA, 2009.
[38] 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.
[39] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Proceedings of the Computer Vision and Pattern Recognition, San Francisco, USA, 2010.
[40] J.C. Gemert, C.J. Veenman, A. Smeulders, J. Geusebroek, Visual word ambiguity, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1271–1283.

Chunjie Zhang received his PhD degree in Pattern Recognition and Intelligent Systems from Institute of Automation, Chinese Academy of Sciences, China in 2011. He received his BE degree from Nanjing University of Posts and Telecommunications, China, 2006. He worked as an engineer in the Henan Electric Power Research Institute during 2011–2012. He is currently working as postdoc at Graduate University of Chinese Academy of Sciences, Beijing, China.

Zhang's current research interests include Image Processing, Machine Learning, Cross Media Content Analysis, Pattern Recognition and Computer Vision.



Jing Liu received the BE and ME degrees from Shandong University, Shandong, in 2001 and 2004, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2008.

She is an associate professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her current research interests include multimedia analysis, understanding, and retrieval.

Qi Tian received his PhD degree in Electrical and Computer Engineering from University of Illinois at Urbana-Champaign, Illinois in 2002. He received his MS degree from Drexel University, Philadelphia, Pennsylvania, 1996, and BE degree from Tsinghua University, China, 1992, respectively.

He is currently an associate professor in the Department of Computer Science at the University of Texas at San Antonio, and Adjunct Professor in Zhejiang University and Xidian University. Tian's current research interests include Multimedia Information Retrieval, Computational Systems Biology, Biometrics, and Computer Vision. He is a Senior Member of IEEE , and a Member of ACM.

Chao Liang received the BS degree in Automation from Huazhong University of Science and Technology, Wuhan, China, in 2006, and the PhD degree in Pattern Recognition and Intelligent System from Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently working as postdoc teacher at National Engineering Research Center for Multimedia Software, Wuhan, China.

His research interests include multimedia content analysis, machine learning, computer vision, and pattern recognition.

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

He was a postdoctoral fellow with the National University of Singapore from 1995 to 1996 and was with the Institute for Infocomm Research, Singapore, as a Member of Research Staff from 1996 to 2002. He joined the Chinese Academy of Sciences, Beijing, China, under Science100 Talent Plan in 2003, and is currently a professor with the Graduate University, Chinese Academy of Sciences. His current research areas are image and video analysis, video coding, pattern recognition, and computer vision.