Beyond Bag of Words: Image Representation in Sub-semantic Space

Chunjie Zhang[1], Shuhui Wang[2], Chao Liang[3], Jing Liu[4], Qingming Huang[1,2], Haojie Li[5], Qi Tian[6]

[1]School of Computer and Control Engineering, University of Chinese Academy of Sciences, 100049, Beijing, China
[2]Key Lab of Intell. Info. Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[3]National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, 430072, Wuhan, China
[4]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[5]School of Software, Dalian University of Technology, Liaoning, China
[6]Department of Computer Sciences, University of Texas at San Antonio, TX, 78249, U.S.A

{cjzhang, shwang, qmhuang}@jdl.ac.cn, cliang@whu.edu.cn, jliu@nlpr.ia.ac.cn, hjli@dlut.edu.cn, qitian@cs.utsa.edu

ABSTRACT

Due to the semantic gap, the low-level features are not able to semantically represent images well. Besides, traditional semantic related image representation may not be able to cope with large inter class variations and are not very robust to noise. To solve these problems, in this paper, we propose a novel image representation method in the sub-semantic space. First, examplar classifiers are trained by separating each training image from the others and serve as the weak semantic similarity measurement. Then a graph is constructed by combining the visual similarity and weak semantic similarity of these training images. We partition this graph into visually and semantically similar sub-sets. Each sub-set of images are then used to train classifiers in order to separate this sub-set from the others. The learned sub-set classifiers are then used to construct a sub-semantic space based representation of images. This sub-semantic space is not only more semantically meaningful but also more reliable and resistant to noise. Finally, we make categorization of images using this sub-semantic space based representation on several public datasets to demonstrate the effectiveness of the proposed method.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*representation, data structures, and transforms*; I.4.10 [**Image Processing and Computer Vision**]: Image Representation—*statistical*

Keywords

Object categorization, sub-semantic space, examplar classifier, sparse coding.

MM '13, October 21 - 25 2013, Barcelona, Spain

http://dx.doi.org/10.1145/2502081.2502132.

1. INTRODUCTION

The local feature based image representation is widely used by researchers recently. Local features are first extracted either by detection or dense sampling. k-means clustering or sparse coding is then used to generate the codebook and encode these local features. Images are finally represented by bag-of-visual-words (BoW) representation. The BoW model has been proven effective by many researchers [1-5]. However, the visual words have no explicit semantic correspondence with human perception which hinders the discriminative power of the BoW model.

To alleviate this problem, a lot of works have been done. On one hand, more discriminative and powerful features are proposed [2-5]. These well designed features capture more information and are more robust to outliers than traditional BoW model. For example, the spatial pyramid matching (SPM) is proposed by Lazebnik *et al.* [2] and is widely used since its introduction. With the fast development of computational power, the explore of more powerful features will be a hot topic in the future.

On the other hand, the use of semantic related representation is also widely studied. Semantic space based image representation is more interpretable than using local features directly for humans. These semantic spaces can be generated by latent space learning [6-8], using the training images [9-10] or generic object classifiers [11]. However, the learning of effective and robust semantic space is very hard due to the well-known semantic gap. Besides, the semantic space is often learned using all the training images of the same class. However, objects may pose large inter class

variations which makes it very difficult to learn reliable classifiers for robust semantic representation. For example, the frontal view and the side view of a car are quite different. To alleviate this problem, the use of attribute is introduced [12-15] which helps to improve the image representation effectiveness. However, the attributes have to be pre-defined. Besides, choosing the proper attributes requires experience and extensive human labor which limits its scalability for large scale applications.

Recently, the use of examplar image for object detection [16] and categorization [17] become popular. The use of examplar classifier takes the advantage of semantic space based image representation and is also more efficient and easy to train than traditional methods [6-11]. Although proven effective, not all of the examplar classifiers are equally useful for image representation. It would be more effective if we can choose some discriminative examplar classifiers instead of use all of them. Besides, images of the same class often exhibit large inter class variations which means the semantic meaning of examplar classifiers may not be so semantic meaningful for efficient image representation.

To solve the problems mentioned above, in this paper, we go one step further byond the bag of words based image representation and propose to represent images in sub-semantic space. First, we train exampler classifier for each training image which serves as the weak semantic similarity measurement. A visual and semantic similarity graph is constructed by combing the visual similarity and weak semantic similarity of training images. We then partition this graph to get sub-sets of images which are visually and semantically similar. Each sub-set of images are used to construct a sub-semantic space representation of images. This is achieved by learning SVM classifiers which separate one sub-set of images from the others. Since we use a sub-set of images for representation, this semantic space is named as sub-semantic space. This sub-semantic space based image representation is not only more semantically meaningful than examplar based representation but also more resistant to noise than traditional semantic space based image representation. Finally, to demonstrate the effectiveness of the proposed image representation method, we conduct object categorization experiments on two public datasets.

2. SUB-SEMANTIC SPACE BASED IMAGE REPRESENTATION

In this section, we give the details of the proposed sub-semantic space based image representation method. We use the sparse coding with locality constraints [18] and max pooling technique as the raw image representation. For each training image, visual similarity along with examplar classifier based weak semantic representation are used to construct a visual-semantic similarity graph. This graph is then partitioned to get sub-class partitions of training images. These sub-class partitioned images are trained to get the final sub-semantic representation of images by separating each sub-class from the rest sub-classes.

2.1 Examplar classifier based weak semantic similarity

We use the semantic space technique to represent images. We try to learn a set of exemplar classifiers for each of the training images. Each exemplar classifier is trained with the corresponding training image and all the other images of different classes which exhibits weak semantic meanings. Since this is much easier than classifying the full-class images, we can use simpler classifiers such as linear SVM. This weak semantic information is then used to measure the weak semantic similarity of images.

Formally, let $X = [x_1, ..., x_N] \in \mathbb{R}^{D \times N}$ be the set of D-dimensional BoW representation of N images, where $x_i \in \mathbb{R}^{D \times 1}, i = 1, ..., N$. These images are of K classes and let $Y = (y_1, ..., y_N) \in \{1, ..., K\}^N$ denote the corresponding image labels. For each training image $x_i, i = 1, 2, ..., N$, 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:

$$min_{w_i, b_i} \parallel w_i \parallel^2 + C \times \ell(w_i^T x_i + b_i) + \sum_{j=1}^{N} \ell(-w_i^T x_j - b_i) \quad (1)$$

$$\forall y_j \neq y_i$$

Where C is the weighting parameter which controls the relative importance of x_i. We use the hinge loss as our loss function which has the form of:

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

Libsvm is used to train each exemplar classifier and we can use it for weak semantic image representation. For a given image x, we predict its semantic meanings for each exemplar classifier and use the output of these classifiers as the weak semantic representation $h \in \mathbb{R}^{N \times 1}$, where $h_i = w_i^T x + b_i, i = 1, 2, ..., N$. The spatial pyramid matching (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.

The weak semantic similarity $s_{wss}(i, j)$ between image i and j is then defined as:

$$s_{wss}(i, j) = exp^{\|h_i - h_j\|^2 / \sigma_1} \quad (3)$$

where σ_1 is the parameter which controls the relative influence of weak semantic similarity.

2.2 Visual-semantic graph construction and partition

Similarly, we can define the visual similarity $s_{vs}(i, j)$ between image i and j as:

$$s_{vs}(i, j) = exp^{\|x_i - x_j\|^2 / \sigma_2} \quad (4)$$

where σ_2 is the parameter which controls the relative influence of visual similarity.

We propose to use a visual-semantic graph to model the visual similarity as well as the weak semantic similarity of images. The nodes correspond to the training images while the edges correspond to the visual-semantic similarity which is defined as the sum of visual similarity and weak semantic similarity

$$w(i, j) = \alpha s_{wss}(i, j) + (1 - \alpha) s_{vs} \qquad y_i == y_j \quad (5)$$

α is a balancing parameter which controls the relative importance of visual similarity and weak semantic similarity. Not that only images of the same class are considered. Let W be the graph matrix with $W = (w_{i,j})_{i,j=1,...,n}$ and D

be the corresponding diagonal degree matrix with its diagonal elements as $d_i = \sum_{j=1}^{n} w_{i,j}$. The corresponding graph Laplacian matrix is then defined as:

$$L = D - W \tag{6}$$

We adopt the spectral clustering technique to group the visual-semantic similarity graph into M sub-sets for each image class. This is achieved by finding the first M eigenvalues (ordered increasingly) and corresponding eigenvectors of L [22]. After these sub-sets of images are obtained, we can use them to construct the sub-semantic representation of images.

2.3 Sub-semantic space based image representation

Let $X = [X_1, ..., X_{MK}]$ be the partitioned training images with X_k is the k-th subset. We propose to train linear SVM classifiers to separate each subset of images with the others and use the output of these learned classifiers for sub-semantic space representation. This image representation not only takes advantage of the semantic based representation, but also is more reliable and robust to noise than examplar based weak semantic representation [17]. This is because we can get ride of some confusing training samples during the graph partition process. This makes the final image representation more semantically meaningful. In fact, the traditional semantic space based image representation [9-10] and examplar based method [17] can be viewed as special cases of the proposed sub-semantic space method. If we set α to 0 and M to 1, the sub-semantic space based image representation will degenerate to the traditional semantic space method. If we set α to 1 and M to the number of training images per class, the proposed sub-semantic space method will degenerate to the examplar based method.

We choose to conduct object categorization experiments to demonstrate the effectiveness of the proposed sub-semantic space based image representation method. We follow the one-versus-all strategy as [17] did and learn K binary SVM classifiers to predict the categories of images.

3. EXPERIMENTS

We evaluate the proposed sub-semantic space based image representation method for categorization on two public datasets: the Scene-15 dataset [2] and the Caltech-256 dataset [19] as [17] did for fair comparison. We follow the same experimental setup and densely extract SIFT descriptors on overlapping 16×16 pixels with an overlap of 6 pixels. Sparse coding [3] with locality constraint [18] is used to encode local features as it has been proven more effective than k-means clustering method. Max pooling is then used to extract image representation which is used for training exemplar classifiers. The codebook size is set to 1,024 for the two datasets.

3.1 Scene-15 dataset

The Scene-15 dataset has 15 categories (bedroom, CALsuburb, coast, forest, highway, industrial, insidecity, kitchen, livingroom, mountain, opencountry, PARoffice, store, street, tallbuilding) with a total of 4,485 images and ranges from natural scenes to man-made environments. Each class of the Scene-15 dataset has 200 to 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 testing, as did in [2, 3, 17]. We repeat this process for 6 times. We report our final results by the mean and standard deviation of the average of per-class classification rates.

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: Weak semantic representation with examplar classifier; S^3R: the proposed sub-semantic space representation.

Algorithm	Performance
KSPM [3]	76.73 ± 0.65
ScSPM [3]	80.28 ± 0.93
KSPM [2]	81.40 ± 0.50
LSS [9]	72.20 ± 0.20
OB [11]	80.9
KCSPM [20]	76.70 ± 0.40
WSR-EC(k-means) [17]	77.82 ± 0.63
WSR-EC(sparse coding) [17]	81.54 ± 0.59
S^3R (k-means)	78.35 ± 0.64
S^3R (sparse coding)	$\mathbf{82.86 \pm 0.75}$

We give the performance comparison of the proposed method with [2, 3, 9, 11, 17, 20] in Table 1. KCSPM [20] tried to alleviate the information loss during hard assignment of visual words and used soft assignment instead. We give the results of the proposed method using k-means clustering and sparse coding for local feature encoding respectively. 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 sub-semantic based image representation makes the S^3R not only more semantically meaningful than examplar classifier based method but also helps to get ride of some noisy examplar image which may hinder the final performance. The use of sparse coding helps improve the performance over k-means clustering based method.

On analysis of the detailed categorization performance, we found that the relative improvement of S^3R over WSR-EC is on the indoor classes. We believe this is because the outdoor classes are relatively easy and has smaller inter class variations compared with indoor classes. It is sufficient to use examplar classifier based method while still obtaining good performance for the outdoor class.

3.2 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 [3, 17] 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 [3, 17-22] in Table 2. Classemes [21] used weakly trained object classifiers for object categorization while the NBNN [22] method worked directly on local features without local feature quantization. We can see

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 [3]	23.34 ± 0.42	29.51 ± 0.52	–
ScSPM [3]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55
Classemes [21]	–	36.00	–
OB [11]	–	39.00	–
NBNN(1 Desc)[22]	30.45	38.18	–
KSPM [19]	–	34.10	–
LLC [18]	34.36	41.19	45.31
KCSPM [20]	–	27.17 ± 0.46	–
WSR-EC [17]	35.28 ± 0.65	42.01 ± 0.47	45.82 ± 0.54
S^3R	**37.18 ± 0.52**	**42.87 ± 0.54**	**46.06 ± 0.48**

from table 2 that S^3R performs better than WSR-EC which demonstrates the effectiveness of using sub-semantic space for image representation. Besides, S^3R also outperforms the OB method which leverages the internet resources for semantic image representation. Since images of the Caltech-256 dataset have larger inter class variations than the Scene 15 dataset, more sub-semantic classes are needed in order to get better object categorization performance.

In our future work, we will study how to select the proper sub-semantic space with sparsity constraints [23, 24].

4. ACKNOWLEDGEMENT

This work is supported in part by National Basic Research Program of China (973 Program): 2012CB316400; the President Fund of UCAS; the Open Project Program of the National Laboratory of Pattern Recognition (NLPR); China Postdoctoral Science Foundation: 2012M520434, 2013T60156, 2013M530350, 2013M530739; National Natural Science Foundation of China: 61025011, 61272329; This work is supported in part to Dr. Qi Tian by ARO grant W911BF-12-1-0057, NSF IIS 1052851, Faculty Research Awards by Google, FXPAL, and NEC Laboratories of America and 2012 UTSA START-R Research Award respectively. This work is supported in part by NSFC 61128007.

5. REFERENCES

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

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

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

[4] 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, *CVPR*, 2011, pages 1673-1680.

[5] P. F. Felzenszwalb, R. B. Girschick, D. McCallester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 32(9) (2010) 1627-1645.

[6] Thomas Hofmann, Probabilistic latent semantic analysis, *UAI*, Sweden, 1999, pages 289–296.

[7] D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, *Joruanl of Machine Learning Research*, 3(1) (2003) 993-1022.

[8] A. Oliva, A. Torralba, A. Guerin-Dugue, J. Herault, Global semantic classification of scenes using power spectrum templates, *CIR*, UK, 1999.

[9] N. Rasiwasia, N. Vasconcelos, Scene classification with low-dimensional semantic spaces and weak supervision, *CVPR*, USA, 2008, pages 1-6.

[10] 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 *Transactions on Multimedia*, 9(5) (2007) 958-966.

[11] L. Li, H. Su, E. Xing, F. Li, ObjectBank: A high-level image representation for scene classification & semantic feature sparsification, *NIPS*, 2010, pages 1-9.

[12] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, *CVPR*, USA, 2009, pages 1778-1785.

[13] C. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, *CVPR*, USA, 2009, pages 951-958.

[14] D. Parikh, K. Grauman, Interactively building a discriminative vocabulary of nameable attributes, *CVPR*, USA, 2011, pages 1681-1688.

[15] D. Parikh, K. Grauman, Relative attributes. *ICCV*, Spain, 2011, pages 1-8.

[16] T. Malisiewicz, A. Gupta, A. Efros, Ensemble of exemplar-SVMs for object detection and beyond, *ICCV*, Spain, 2011, pages 89-96.

[17] C. Zhang, J. Liu, Q. Tian, C. Liang, and Q. Huang, Beyond visual features: A weak semantic image representation using exemplar classifier for classification, *Neurocomputing*, 2012.

[18] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, *CVPR*, USA, 2010, pages 3360-3367.

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

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

[21] L. Torresani, M. Szummer, A. Fitzgibbon, Efficient object category recognition using classesmes. *ECCV*, Greece, 2010, pages 776-789.

[22] O. Boiman, E. Shechtman, M. Irani, In defense of nearest-neighbor based image classification, *CVPR*, USA, 2008, pages 1-8.

[23] C. Zhang, J. Liu, Q. Tian, Y. Han, H. Lu, S. Ma, A Boosting Sparsity-Constrained Bilinear Model for Object Recognition, *IEEE Multimedia*, 2012, 19(2):58-68.

[24] C. Zhang, J. Liu, Q. Tian, C. Liang, Q. Huang, Image Classification Using Harr-like Transformation of Local Features with Coding Residuals, *Signal Processing*, 2012, 93(8):2111-2118.