LARGE SCALE LINEAR CODING FOR IMAGE CLASSIFICATION

Mustafa Ibrahim[1], Mohamed Waleed Fakhr[2] and Mustafa Abdel Aziem[1]

[1]The Arab Academy for Science and Technology
[2]University of Bahrain
Cairo – Egypt

Mostafa_ebrahim87@yahoo.com,waleedfakhr@yahoo.com,melbakary@aast.edu

Abstract

Image classification, including object recognition and scene classification, remains to be a major challenge to the computer vision community. As machine can be able to extract information from an image and classify it in order to solve some tasks. Recently SVMs using Spatial Pyramid Matching (SPM) kernel have been highly successful in image classification. Despite its popularity, this technique cannot handle more than thousands of training images. In this paper we develop an extension of the SPM method, by generalizing Vector Quantization to Sparse Coding followed by multi-scale Spatial Max Pooling, and also propose a large scale linear classifier based on Scale Invariant Feature Transform (SIFT) and Sparse Codes. This new adapted algorithm remarkably can handle thousands of training images and classify them into different categories.

Keywords-Scale Invariant Feature Transform, Spatial Pyramid Matching, Support Vector Machine, Bag-of-Words and Bag-of-Features.

1. INTRODUCTION

One of the most significant developments in the last decade is the application of local features to image classification, including the introduction of "Bag-of-Words" (BoW) representation that inspires and initiates many research efforts [1]. In recent years, the Bag-of-Features (BoF) model has been extremely popular in image categorization [2]. The method treats an image as a collection of unordered appearance descriptors extracted from local patches, quantizes them into discrete “visual words”, and then computes a compact histogram representation for semantic image classification, e.g. object recognition or scene categorization [2].

One particular extension of the BoF model, called Spatial Pyramid Matching (SPM) [3], has made a remarkable success on a range of image classification benchmarks like Caltech-101 [4] and Caltech-256 [2, 5].

Linear classification has become one of the most promising learning techniques for large sparse data with a huge number of instances and features. For example, it takes only several seconds to train an image classification problem from Caltech 101 that has more than 100,000 examples. For the same task, a traditional SVM solver such as LIBSVM would take several hours. Moreover, LIBLINEAR is competitive with or even faster than state of the art linear classifiers such LIBSVM [6].

The rest of the paperorganized as follows. In section2, discuss the problem statement. Section 3 talk about some related works. Section4 and 5 presents the framework of our proposed algorithm and supported by our efficient implementation in section 6.Section 7 display of experiment results. Finally, section 8 concludes our paper.

2. PROBLEM STATEMENT

The traditional SPM approach based on Bag-of-Features (BoF) requires nonlinear classifiers to achieve good image classification performance [7]. However, these results are not effective in classifying real data so using sparse coding with SPM allow us to use linear classifier instead of nonlinear, which used before, but the linear Classifier still has limitation in huge data. There for, in this paper we proposed a technique that uses the capabilities of large-scale linear classifier in classification process with Sparse coding Spatial Pyramid Matching technique, which require a linear classifier to have good results with huge data.

3. RELATED WORK

In computer vision, the Bag-of-Words model (BoW model) can be applied to image classification as Bag-of-Features (BoF model), by treating image features as words. In document classification, a Bag of Words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, a bag of visual words is a sparse vector of occurrence counts of a vocabulary of local image features. To represent an image using BoW model, the image can be treated as a document. Similarly, "words" in images needs to be definedtoo. To achieve this, it usually passes by next three steps: Feature detection (computer vision), feature description and codebook generation [8]. A definition of the BoW model can be the "histogram representation based on independent features" [9].

The BoF approach discards the spatial order of local descriptors, which severely limits the descriptive power of the image representation. By overcoming this problem, one particular extension of the BoF model, called Spatial Pyramid Matching (SPM) [3], had made a remarkable success on a range of image classification benchmarks like Caltech-101 [4] and Caltech-256 [2, 5].

Researchers have found that [2], in order to obtain good performances, both BoF and SPM must be applied together with a particular type of nonlinear Mercer kernels, e.g. the intersection kernel or the Chi-square kernel. Accordingly, we can say that, the traditional SPM approach based on (BoF) requires nonlinear classifiers to achieve good image classification performance [7]. The nonlinear SVM method using Spatial Pyramid Matching (SPM) kernels [10,11] seems to be dominant among the top performers in various image classification benchmarks, the nonlinear SVM has to pay a computational complexity $O(n^3)$ and a memory complexity $O(n^2)$ in the training phase, where n is the training size. Furthermore, since the number of support vectors grows linearly with n, the computational complexity in testing is O(n).This scalability implies a severe limitation as it is nontrivial to apply them to real-world applications, whose training size is typically far beyond thousands [2].

Using Sparse coding with spatial pyramid matching can to represent each image by single image feature where the output of SIFT algorithm (local feature vectors for each image) becomes the input to coding phase ScSPM as shown in figure 1 sample. By using this technique, we overcome on the need to use nonlinear classifier in the level of classification.

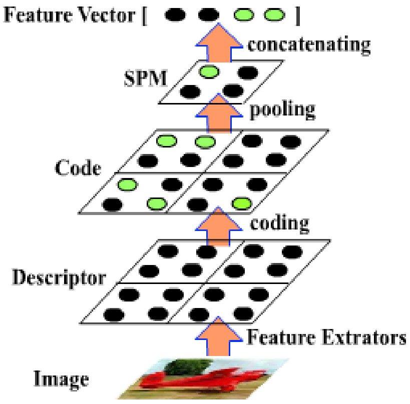


Figure1:Spatial pyramid structure for pooling features for image classification[7].

4. ENCODINGFEATURES from VECTOR QUANTIZATION to SPARSE CODING

Let X be a set of SIFT appearance descriptors in a D-dimensional feature space, i.e. X=[x_1, . . . ,x_M]^T∈ $\mathbb{R}^{M}$ x D TheVector Quantization (VQ) method applies theK-means clustering algorithm to solve the following problem[2].

$$\min_{\mathrm{V}} \sum_{m=1}^{M} \min_{\mathrm{k}=1\ldots\mathrm{k}} \| x_m - v_k \|^2 \quad (1)$$

Where V=[v_1, . . . ,v_k]^{T}are the K cluster centers to be found, called codebook, and‖ . ‖denotes the L2-norm of vectors[2]. The optimization problem can be reformulatedinto a matrix factorization problem with cluster membership indicators

$$\min_{\mathrm{U,V}} \sum_{m=1}^{M} \| x_m - u_m V \|^2 \quad (2)$$

Subject to Card (u_m)=1 , | $\mathrm{u_m}$|=1,$\mathrm{u_m} \geq 0 , \forall\, \mathrm{m}$

WhereCard(u_m) = 1 is a cardinality constraint, meaning that only one element of u_m is nonzero, u_m ≥ 0 means that all the elements of u_m are nonnegative, and | u_m| is the L1-norm of u_m, the summation of the absolute value of each element in u_m. After the optimization, the index of the only nonzero element in u_m indicates which cluster the vector x_m belongs to. In the training phase of VQ, the optimization Eq.(2) is solved with respect to both U and V. In the coding phase, a learned V which will be applied for a new set of X . Then Eq. (2)solved with respect to U only[2].

The constraint Card(u_m) = 1 may be too restrictive, giving rise to often a coarse reconstruction of X. to relax the constraint by instead putting L1-norm regularization on u_m, which enforces u_m to have a small number of nonzero elements. Then the VQ formulation turned into another problem known as sparse coding (SC):

$$\min_{\mathrm{U,V}} \sum_{m=1}^{M} \| x_m - u_m V \|^2 + \lambda\, |u_m| \quad (3)$$

Subject to $\| \mathrm{v_k} \| \leq 1 , \ \forall\, \mathrm{k} = 1,2, \ldots . \mathrm{K}$

Where a unit L2-norm constraint on $v_k$typically applied to avoid trivial equation(1)[2]. Normally, the codebook V is an over complete basis set, i.e. K > D-dimensional. Note that we drop out the no negativity constraint u_m ≥ 0 as well, because the sign of u_m is not essential, it can be easily absorbed by lettingV ^{T}⟵ = [V ^{T}, – V ^{T}] and $\mathrm{u_m^T} \longleftarrow [\mathrm{u_{m+}^T}, -\mathrm{u_{m-}^T}]$so that the constraint can be trivially satisfied, where $\mathrm{u_{m+}^T} = \min(0, \mathrm{u_m})$and $\mathrm{u_{m-}} = \max(0, \mathrm{u_m})$

Similar to VQ, SC has a training phase and a coding phase. First, a descriptor set X from a random collection of image patches is used to solve Eq. (3) with respect to U and V, where Vis retained as the codebook; In the coding phase, for each image represented as a descriptor set X, the SC codes are obtained by optimizing Eq. (3) with respect to U only[2].

Sparse Coding hasbeen chosento derive image representations because it has a number of attractive properties. First, compared with the VQ coding, SC coding can achieve a much lower reconstruction error due to the less restrictive constraint; second, sparsely allows the representation to be specialized, and to capture salient properties of images; third, research in image statistics clearly reveals that image patches are sparse signals[2].

5. LINEAR SPATIAL PYRAMID MATCHING

For any image represented by a set of descriptors, a single feature vector based on some statistics of the descriptors' codes can be computed. For example, if U is obtained via Eq. (2), a popular choice is to compute the histogram

$$\mathrm{z} = \frac{1}{\mathrm{M}} \sum_{m=1}^{M} u_m \qquad (4)$$

The Bag-of-Words approach to image classification computes such a histogram z for each image I represented by an unordered set of local descriptors. In the more sophisticated SPM approach, the image's Spatial Pyramid histogram representation z is a concatenation of local histograms in various partitions of different scales. After normalization, z can be seen as again a histogram. Let z_i denote the histogram representation for image I_i.[2]For a binary image classification problem, an SVM aims to learn a decision function

$$f(z) = \sum_{i=1}^{n} \propto_i k(z, z_i) + b(5)$$

Wheref$\{(z_i;\ y_i)\}^n_{i=1}$ is the training set, and $y_i \in \{-1;\ +1\}$ indicates labels. For a test image represented by z, if f(z) > 0 then the image is classified as positive, otherwise as negative[2].

6. LARGE SCALE LINEAR CODING for IMAGECLASSIFICATION

This paper follows another line of research on building discriminative models for classification. The previous work includes nonlinear SVMs using pyramid matching kernels [10] and K-Nearest Neighbor (KNN) methods [11, 12, 13], or Linear ScSPM technique as shown in figure 2, Over the past years.

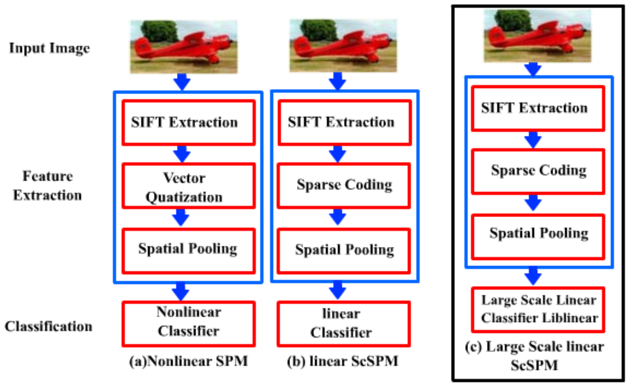


Figure 2:(a)Traditional way using nonlinear classifier (b)linear classifier with sparse coding spatial pyramid matching (c)large scale linear classifier [2].

Experimental results, have shown that Linear Sparse coding Spatial Pyramid Matching "Linear ScSPM" not enough for real-world application contain thousands of images. Where this technique was triedto use it to classify 2000 image(100 category each one has 10 images for training and 10 images for testing) but it cannot do that as shown in Table (1) using Intel Core2due 2.33 GHz and 4 GB Ram Dell machine.

Table 1:Limitation of linear SCSPM in classifying data “previous technique ”

Experiments	# categories	#Training data	#Testing data	Mean accuracy
1	2	20	20	100%
2	2	25	25	100%
3	5	25	25	93.28%
4	10	25	25	91.13%
5	15	25	25	82.53%
6	100	5	5	58.08%
7	100	10	10	Error out of Memory

Therefore, this technique triedto employ the powerful of the previous technique "linear ScSPM" and in the same time,it handles the limitation of classifying thousands of real word data. Large-scale linear classifier "liblinear-1.91" could to handle these limitations perfectly figure 2.

Using Scale Invariant Feature Transform (SIFT) algorithm local feature descriptors for each image can be extracted then Sparse coding with Spatial Pyramid Matching (ScSPM) encode the extracted features into single features vector which represent the salient properties of images figure 3, whereupon, Large scale linear classifier can be used to classify large scale of data speedily and accurately.

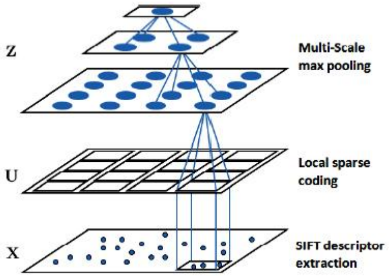


Figure3:The architecture of new algorithm based on sparse coding[2].

Sparse coding measures the responses of each local descriptor to the dictionary’s visual elements. These responses are pooled across different spatial locations over different spatial scales.

Linear classification has become one of the most promising learning techniques for large sparse data with a huge number of instances and features. Using linear- SVM "LIBLINEAR" instead of traditional SVM "LIBSVM" thousand of images can be classified perfectly, where, LIBLINEAR is optimized to deal with linear classification (i.e. no kernels necessary), whereas linear classification is only one of the many capabilities of libsvm, So logically it may not match up to LIBLINEAR in terms of classification accuracy [6].

LIBLINEAR supports two popular binary linear classifiers: LR and linear SVM. Given a set of instance-label pairs (x_i; y_i); i = 1, . . ,l , x_iR^n , y_i {+1,-1}, both methods solve the Following unconstrained optimization problem with different loss functions ε(w; x_i; y_i):

$$\min_{\mathrm{w}} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi(w; x_i;\ y_i)$$

Where, C > 0 is a penalty parameter. For SVM, the two common loss functions are max(1-$y_i w^T x_i$; 0) and max(1-$y_i w^T x_i$, 0)2. The former is referred to as L1-SVM, while the latter isL2-SVM. For LR, the loss function is log (1+$e^{-\text{yiwTxi}}$), which derived from a probabilistic model. The software is available at http://www.csie.ntu.edu.tw/~cjlin/liblinear.

7. EXPERIMENTS and RESULTS

By evaluating ourtechnique on real data using IMAGENET and Caltech-101 dataset, hundreds of classes have been chosen to test our algorithm comparing with IMAGENET competition results September 16 2010, and libsvm experimental results 2011 [2].We represent results of Locality-constrained Linear Coding (LLC) algorithm that only used single descriptor (HoG) and simple linear SVM as the classifier.Using Caltech-256 dataset figure 4[7].In our experimentsweonly used one patch size to extract SIFT descriptors, namely, 16 x 16 pixels as in SPM [3],maximum image size was 300 x 300 pixels for width and height. In addition, we used LIBLINEAR -1.91 libraries as large-scale linear classifier.

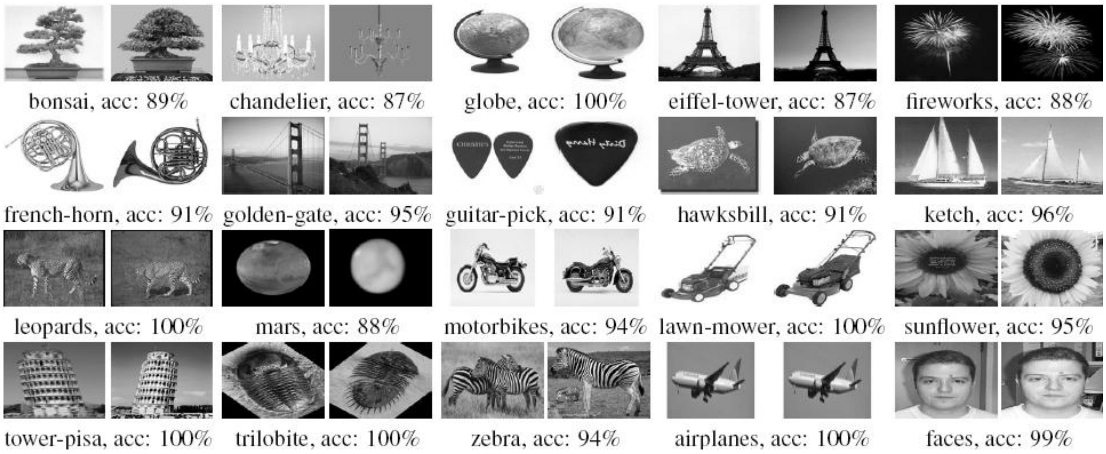
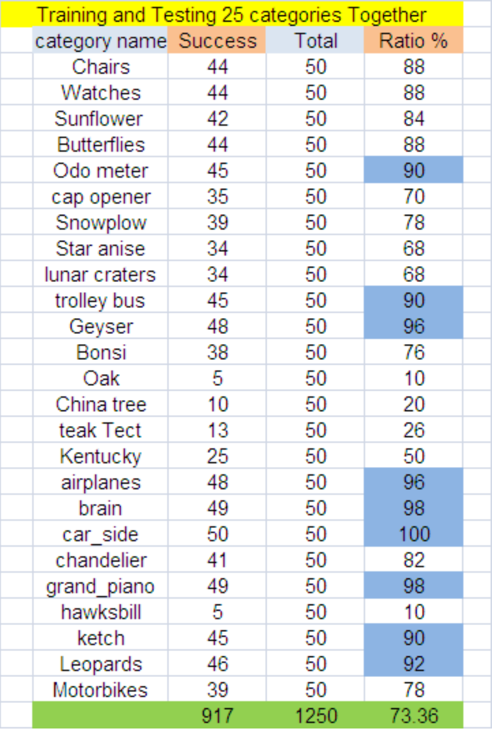


Figure4:Example images from classes with highest classification accuracy from the Caltech-256 datasetClassifying 97 categories using "LLC algorithm"[7]

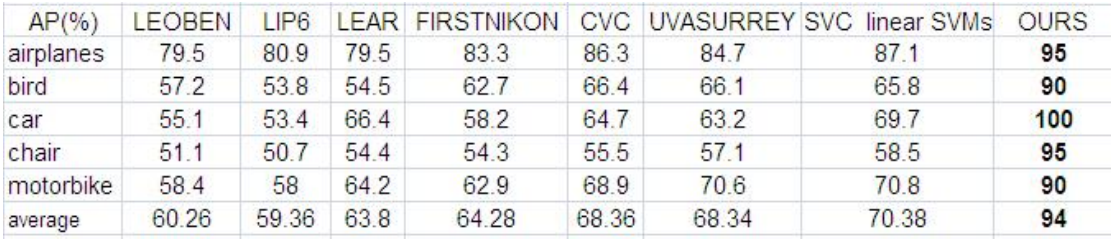
Table 2:Classifying 25 categories using Liblinearclassifier. "Our algorithm"



Training and Testing 25 categories Together

category name	Success	Total	Ratio %
Chairs	44	50	88
Watches	44	50	88
Sunflower	42	50	84
Butterflies	44	50	88
Odo meter	45	50	90
cap opener	35	50	70
Snowplow	39	50	78
Star anise	34	50	68
lunar craters	34	50	68
trolley bus	45	50	90
Geyser	48	50	96
Bonsi	38	50	76
Oak	5	50	10
China tree	10	50	20
teak Tect	13	50	26
Kentucky	25	50	50
airplanes	48	50	96
brain	49	50	98
car_side	50	50	100
chandelier	41	50	82
grand_piano	49	50	98
hawksbill	5	50	10
ketch	45	50	90
Leopards	46	50	92
Motorbikes	39	50	78
	917	1250	73.36

Table 3:Comparison of our method with top performers of some categories inECCV 10.[1]



AP(%)	LEOBEN	LIP6	LEAR	FIRSTNIKON	CVC	UVASURREY	SVC linear SVMs	OURS
airplanes	79.5	80.9	79.5	83.3	86.3	84.7	87.1	**95**
bird	57.2	53.8	54.5	62.7	66.4	66.1	65.8	**90**
car	55.1	53.4	66.4	58.2	64.7	63.2	69.7	**100**
chair	51.1	50.7	54.4	54.3	55.5	57.1	58.5	**95**
motorbike	58.4	58	64.2	62.9	68.9	70.6	70.8	**90**
average	60.26	59.36	63.8	64.28	68.36	68.34	70.38	**94**

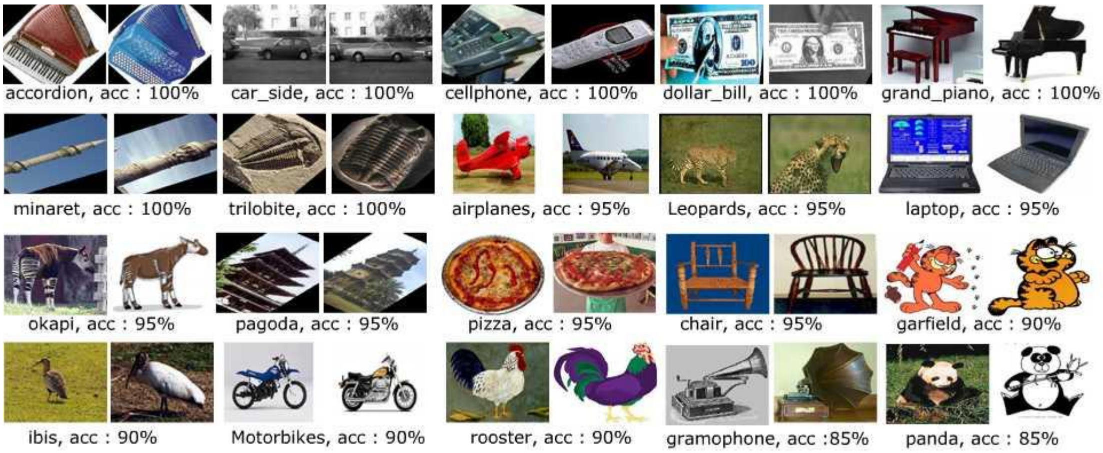


Figure 5: Example images from classes with highest classification accuracy from the Caltech-101 datasetClassifying 97 categories using "Our Algorithm"

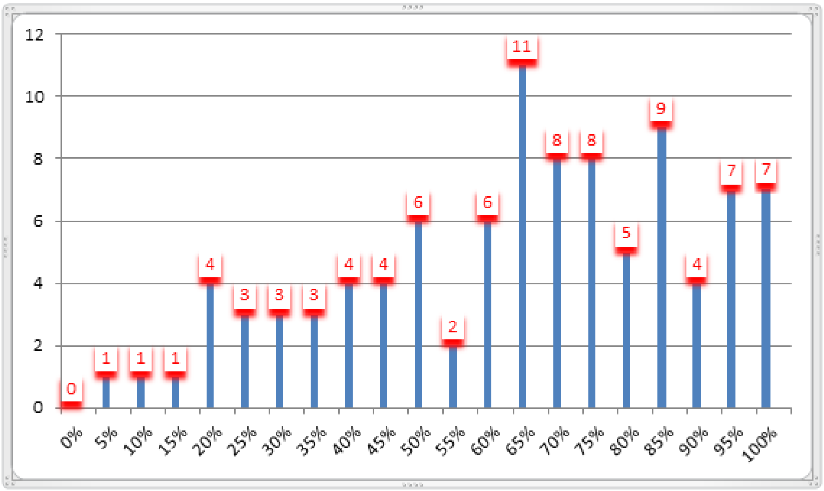


Figure 6:Chart of Classifying 97 categories using Liblinear classifier

Using our algorithm, we conducted several experiments and here we represent some results of our experiments. In table (2) appear the results of classifying 25 categories, while in table (3) representsa comparison among our results and other techniques in ECCV2010 and PASCAL VOC 2009. Figure 5 represent some results of classifying 97categories using Caltech 101 dataset, Finally figure 6 illustrate chart that represent the number of categories whose achieve specific ratio in classifying 97 categories using 2910 image “15 image in training and 20 image in testing ” where x-axis represent ratio of success while y-axis represent number of categories achieved this ratio. We executed many experiments and compared their results with ECCV2010 [1, 7],libsvm experimental results 2011 and

IMAGENET results. Results demonstrated that our technique quickly reaches the testing accuracy corresponding to the optimal solution of ScSPM figure2 and handle the scalability limitation problem.

8. CONCLUSION

In this paper we proposed a large-scale linear classifier liblinear using spatial pyramid matching approach based on SIFT sparse codes for image classification. The method uses selective Sparse Coding instead of traditional Vector Quantization to extract salient properties of appearance descriptors of local image patches. Furthermore, instead of averaging pooling in the histogram, sparse coding enables us to operate local max pooling on multiple spatial scales to incorporate translation and scale invariance. Where, each image can be encoded into single meaningful feature vector. The most encouraging result of this paper is the obtained image representation works surprisingly well with simple large-scale linear classifier liblinear, which dramatically improves the scalability of training and the speed of testing, and even improves the classification accuracy. Using 2910 image (97 categories), as a simple sample of large-scale data. Our experiments on a variety of image classification tasks demonstrated the effectiveness of this approach. As an indication from our work, the sparse codes of SIFT features might serve as a better local appearance descriptor for general image processing tasks.

9. FUTURE WORK

A recent work shows that sparse coding approach can be accelerated by using a feed-forward network [14]. It will be interesting to try such methods to make our approach faster. Moreover, the accuracy could be improved by learning the codebook in a supervised fashion, as suggested by another recent work [15].

References

[1] Xi Zhouy, Kai Yuz, Tong Zhang, and Thomas S. Huangy, "Image Classification using Super-Vector Coding of Local Image Descriptors", Dept. of ECE, University of Illinois at Urbana-Champaign, Illinois,NEC Laboratories America, California ,Department of Statistics, Rutgers University, New Jersey ECCV10,2010.

[2]JianchaoYangy, Kai Yuz, YihongGongz, Thomas Huangy, "Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification", Beckman Institute, University of Illinois at Urbana-Champaign ,NEC Laboratories America, Cupertino, CA 95014, USA CVPR09,2009.

[3]S. Lazebnik, C. Schmid, and J. Ponce,"Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", CVPR, 2006.

[4]F.F.Li, R. Fergus, and P. Perona, "Learning generative visual models from few training" examples: an incremental Bayesian approach tested on 101 object categories. In CVPRWorkshop on GenerativeModel Based Vision, 2004.

[5]C.C.Chang and C.J. Lin. "LIBSVM: a library for support vector machines", 2001, Software available at: http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

[6]RongEn Fan, Kai-Wei Chang, ChoJui Hsieh, Xiang Rui Wang and Chih-Jen Lin,"LIBLINEAR: A Library for Large Linear Classification", Department of Computer Science, National Taiwan University, Taipei 106, Taiwan July 14, 2012.

[7]Jinjun Wang†, Jianchao Yang‡, Kai Yu§, FengjunLv§, Thomas Huang, and Yihong Gong ,"Locality constrained Linear Coding for Image Classification",Akiira Media System, Palo Alto, California ,Beckman Institute, University of Illinois at Urbana-Champaign ,NEC Laboratories America, Inc., Cupertino, California,CVPR'10,2010.

[8] L. FeiFei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories", Procof IEEE Computer Vision and Pattern Recognition, pp. 524–531,2005.

[9]L. FeiFei, R. Fergus, and A. Torralba,"Recognizing and Learning Object Categories", CVPR 07 short course,2007.

[10]Lazebnik, S.Schmid, C. Ponce,"bags of features, Spatial pyramid matching for recognizing natural scene categories", Citeseer,2006.

[11] Bosch, A.Zisserman, and A. Munoz, "Representing shape with a spatial pyramid kernel", Proceedings of the 6th ACM international conference on Image and video retrieval, ACM 2007.

[12]Makadia, A.Pavlovic, and V. Kumar, "New baseline for image annotation", Proc. ECCV08,2008.

[13]Torralba, A.Fergus, and R.Weiss, "Small codes and large image databases for recognition", IEEE Conference on Computer Vision and Pattern Recognition. CVPR08,2008.

[14] K. Kavukcuoglu, M. Ranzato, and Y. LeCun.Fast, "sparse coding algorithms with applications to object recognition", Technical report, Computational and Biological Learning Lab, NYU,2008.

[15] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning" , NIPS, 2009.