Locality Preserving Sparse Group Lasso based Image Categorization

Xiang Yu
Rutgers University, Dept. of Computer Science
617 Bowser Road
Piscataway, N.J., USA
xiangyu@cs.rutgers.edu

Xin Lian
Rutgers University, Dept. of Computer Science
617 Bowser Road
Piscataway, N.J., USA
xl182@cs.rutgers.edu

ABSTRACT

Image categorization is a canonical problem in computer vision field. Many algorithms have been proposed in the past decades and achieved good performance. Most of the methods start with either global features or local features. Though different features bring in different performance due to their ability to represent original images, and different features are suitable to different classifiers, we aim to investigate approaches to preserve the semantic similarity inside the features while they are mapped to other space by classifiers or they are translated into new features by feature selection or reduction strategies. Furthermore, based on the locality preserved features, we expect to design specific classifiers using the maintained local structure. Inspired by such ideas, in this paper, we propose a graph based locality preserving method for feature level ensemble and apply the sparse group lasso algorithm to find the local linear reconstruction of the test samples. The sparse group lasso method uses the neighborhood structure sufficiently and we believe the test samples are best represented by their neighborhood other than those far away training samples in the manifold, since the neighborhood preserves most of the semantic similarities. Experimental results demonstrate that our proposed method achieves far better accuracy than baseline methods and boosted-classifier approaches as good as state-of-the-art methods.

Keywords

image categorization, sparse group lasso, locality preserving

1. INTRODUCTION

Image categorization is a classical problem in computer vision. It refers to labeling images into one of those predefined image categories. Although it seems not a difficult task for humans, due to some unresolved techniques, it is still a hard topic for computer programs. For example, the image acquisition may be unstable or even uncontrolled. Objects inside images are always so complex that it is hard to represent them. Moreover, the ability of representing images by computer data structure is far from the conceptual ability of human beings.

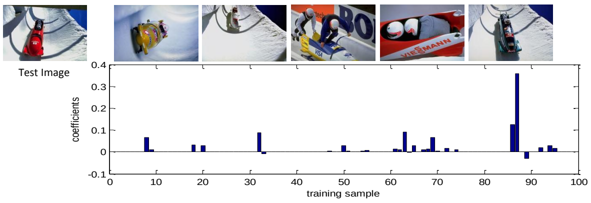


Figure 1: result of sparse group lasso representation of image classification

A number of image classification frameworks [6, 19, 17, 18, 8] have been developed recently. These methods could be mainly divided into two groups. One is global feature based approaches. In this kind of methods, image information is represented as a whole feature vector, or feature vectors are extracted based on the whole image content. As an instance, the global HSV color histograms are used to classify images in the early decades [3]. The others are local feature or region based means. For the local methods, images are segmented into regions or divided into blocks [16]. Each region or block is characterized by one feature vector describing color, texture or shapes. Consequently, an image is a collection of feature vectors. These methods assume the image components could efficiently represent the category information.

However, both of these two types of methods have their own drawbacks. Global systems cannot precisely represent the semantics of an image. Local systems may break an object into several regions or put different objects into a single region due to inaccurate image segmentation. Block based methods may also break the object or have overlaps among different objects.

Due to above considerations, we are motivated to combine the global methods' integral representation and the local methods' semantic indicating ability. In this paper, we propose a locality preserving sparse group lasso based image categorization framework. locality preserving step is based on local methods. Local methods try to explicitly segment regions or blocks out. But our locality preserving method implicitly keeps the local information by maintaining its neighborhood topology. We assume images can be linearly approximated by their neighborhood images, which shows its correctness in Local Linear Embedding (LLE) [14]. It is also worth mentioning that locally linear estimation is a canonical method not only in mathematics but also in physics, e.g. Taylor expansion. Borrow the idea from those fields, we firstly rearrange training images inside the same category according to their neighborhood structure. Then global feature GIST [13, 15] is adopted to form feature matrix A of each category. Since the columns in neighborhood of a feature matrix map to the images in neighborhood in the category manifold, the linear representation to estimate a test image is just related to the columns in some neighborhood. Thus, we expect to learn certain coefficient vector x such that the reconstruction error $\frac{1}{2}\|y-Ax\|_F^2$ achieves minimum from certain category of images, where y is the test image feature vector. The coefficient x must be sparse and group sparse because the test image is just represented by several feature vectors in A and those vectors are in neighborhood. Consequently, we formulate our objective function by regularizing group sparsity to learn the coefficient weights. The sample result is in Fig.1.

The contribution of this paper can be summarized as follows. (1) we propose locality preserving approach to maintain the neighborhood structure in the images spanned manifold. This strategy could implicitly maintain the semantics of local regions without additional preprocessing such as segmentation or blocking. (2) the group sparse regularization over the objective function is a consequent and seamless constraint according to the neighborhood property. By such regularization, the test image could be mapped into the neighborhood of training images.

The rest of this paper is organized as follows: Related work will be briefly discussed in Section 2. Section 3 contains basic acknowledge in sparse group lasso. We will describe the details of our proposed method in Section 4. Experiments will be conducted in Section 5. And conclusions and future work are in Section 6.

2. RELATED WORK

The research community has investigated various techniques to semantically and visually categorize images. As one of the simplest representation of images, histograms have been widely used in image categorization problems. Chapell et al. [3] applied SVM with color histogram features to classify images. Although histogram is fast and with little cost in computation, it lacks the spatial configuration in images. In Huang et al.'s method [10], a classification tree is built with color correlograms. Color correlogram captures spatial correlation of colors in an image. A number of local feature based methods are also proposed to exploit local and spatial properties. Gorkani et al. [9] divided images into non-overlapping equal-sized blocks to extract features. Wang et al. [16] proposed the method of using graphic models to tackle the problem. Using bag-of-visual-words (BOV) to describe images and combining SVM classifiers [6] is a mainstream method in image classification. Recently Multiple Instance Learning (MIL) has been applied in this field [4]. Maron and Ratan [12] use the Diverse Density (DD) learning algorithm for natural scene classification. Zhang and Goldman [19] brought in Expectation Maximization (EM) with DD to achieve a fast and scalable categorization.

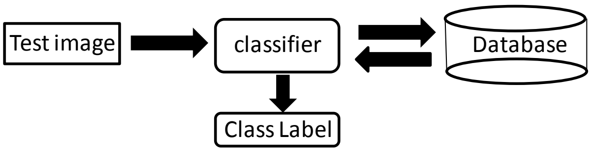


Figure 2: Image categorization framework

In addition to the above sophisticated and canonical methods, sparse coding has been successfully applied in computer vision field, such as image classification [17] and image annotation [20]. Given a test image, the non-zero linear reconstruction should also only lie in one or just a few of the classes. In this way, group sparse coding problem was proposed [1, 18]. Using sparse group lasso, instances are firstly clustered into groups. Overall coefficients should be l_1 norm based since only a few non-zero entries are needed for the representation. And for the group layer, only one or several groups are allowed for representation. Thus the l_1 norm over the l_2 norm of the coefficients of certain groups are introduced. Gao et al. [8] proposed three layer group sparse coding for image classification. They denoted images as class-layer, tag-layer from image content splitting and instance-layer from tags recombining. To solve such kind of problems, sparse group lasso algorithms [7, 11] are recently developed. Our proposed method is also based on the group lasso method. In a similar way, we group the instances according to their class labels. Unlike Gao et al.'s strategy, we do not do regional refinement to further depict the objects in the images. We allow only those images in neighborhood of the test image to reconstruct it, which is valid from the applications of Locally Linear Embedding (LLE) [14]. Nevertheless, the model is simplified from multi-layer structure and thus effectiveness is expected to improve.

3. BACKGROUND INFORMATION

Classification is a main branch of pattern recognition. Image classification is a sub-area of classification applications. Classification is such a problem, given a set of training samples, try to correctly give class label of test samples. The generic image categorization framework can be illustrated as Fig.2. Database stores the feature vectors extracted from training images. Based on training samples, we may establish generic or discriminative models to learn certain parameters of the models. These models are so-called classifiers in the framework. When one test image came in, it follows the same rule of firstly being extracted as feature vector. Then it is sent to the classifier. As the model trained according to the training samples, we expect to tell the right class label because we assume that the test image obeys the same data distribution as the training samples do.

3.1 GIST Feature

In our proposed method, we adopt GIST feature in the feature extraction step. The initial idea proposing GIST[13] is to develop a low dimensional representation of the scene, which does not require any form of segmentation. In Ref.[13], the authors proposed several perceptual dimensions,e.g, naturalness, openness, roughness, expansion and rugged-

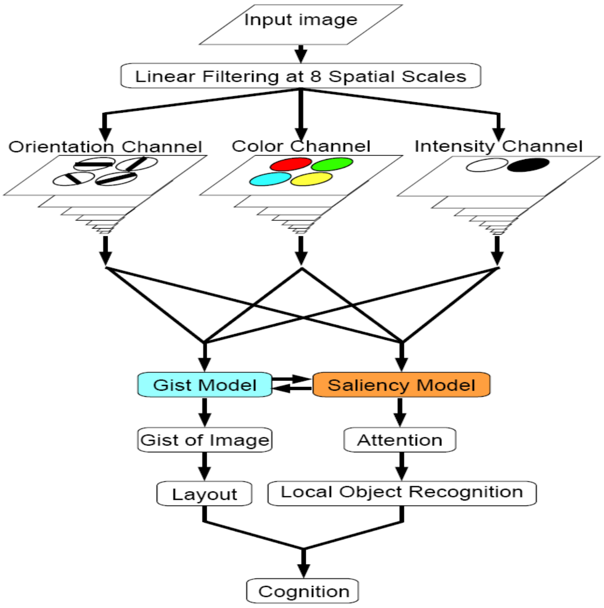


Figure 3: Model of Human Vision with Gist and Saliency[15]

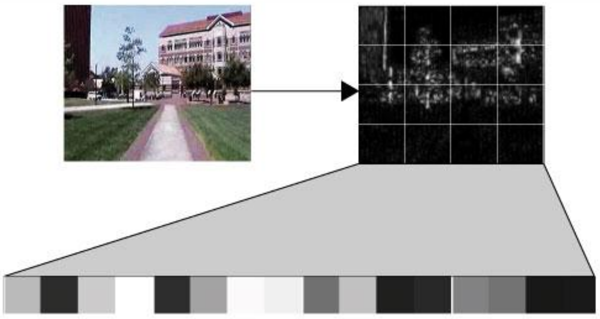


Figure 4: GIST feature extraction[15]

ness that represent the dominant spatial structure of a scene. They argued that these dimensions may be reliably estimated using spectral and coarsely localized information. From Fig.3, we know the model of human vision is constituted by mainly Gist Model and Saliency Model. Gist and saliency appear to be complementary opposites: salient objects are always significantly outstanding than their neighbors, while Gist involves accumulating image statistics over the entire scene. Just like a trade off in recognizing a scene, the two models can help each other and provide a more complete description of a scene. Fig.4 illustrates how GIST feature is extracted. First the image should be resized into regular size, usually the size could be 256×256 or 128×128. Then the resized image is blocked by a 4×4 grid for which orientation histograms are extracted. In parallel, the color channel and intensity channel are also block into the same grid. By different scales of the templates, the average value of a block are calculated to represent the grid. Then different channels and blocks are concatenated to form an integrated feature vector.

3.2 Sparsity and Sparse Group Lasso

The concept of sparsity is widely applied in signal processing, machine learning and statistics communities for model fitting and solving inverse problems. The sparse problem can be depict as: there is a signal $y \in \mathbb{R}^n$, with respect to a dictionary of basis functions $A \in \mathbb{R}^{n \times m}$, can be implemented via an l_1-penalized least-square problem referred as Lasso.

$$\arg\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1 \tag{1}$$

where λ is a regularization parameter that controls the trade-off between the quality of the reconstruction error and the sparsity. Here l_1-norm counts the absolute value of coefficients. In origin, the l_1-norm should be l_0-norm. l_0-norm counts number of non-zeros entries. The less number the non-zero entries of the coefficient x, the more sparse the coefficient is. However, since l_0-norm based regularization is not differentiable, which causes no convergent computational method to solve those problems, l_0-norm is replaced with l_1-norm. l_1-norm regularization is convex and thus has complete close-form solution to practical problems.

Moreover, if the problem is redefined as: suppose the m coefficients are divided into G groups, with p_k elements in group k. And the object is also achieving the minimal reconstruction error. Here the coefficients are not required sparse. But only one or several groups' coefficients should be non-zero. Suppose we add another weight parameter on each group. Those non-zero entry groups get non-zero weights. Thus we expect such weights to be sparse. We formulate the problem as Eq.2.

$$\arg\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda\sum_{k=1}^{M} w_k\|x_k\|_2 \tag{2}$$

Actually x_k is not sparse. $\sum_k w_k\|x_k\|_2$ is equivalent to calculate the averaged l_1-norm of $\|x_k\|_2$. We know that l_1-norm is sparse. Thus the problem is to pursue sparsity at group level, which is referred as group lasso.

In real applications, we may encounter the requirement that the coefficients should be not only sparse but also group sparse. Based on the group lasso formulation, we add a third regularization term $\|x\|_1$ to the objective function, it would strike a balance between the reconstruction error, the sparsity of coefficients and the sparsity of the groups, which can be demonstrated in Eq.3.

$$\arg\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda_1\|x\|_1 + \lambda_2\sum_{k=1}^{M} w_k\|x_k\|_2 \tag{3}$$

where λ_1 and λ_2 are two independent regularization parameters, λ_1 controls the coefficients' sparsity and λ_2 indicates the group sparsity.

4. LOCALITY PRESERVING GROUP SPARSE LEARNING

In this section we will introduce the locality preserving strategy and group sparse learning method. In order to preserve the neighborhood structure of a graph, we take Euclidean norm measure in the vicinity of a node because local distance can be approximated by linear distance. And we expect to span a certain Minimal Spanning Tree (MST) out of the

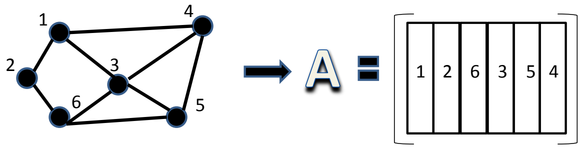


Figure 5: Locality Preserving Strategy from graph to feature matrix

graph and arrange the traning instances in group matrix A in the same order. Then sparse group lasso algorithm is adopted over each group matrix to learn sparse and group sparse coefficients. Those coefficients would reconstruct the test instance in the neighborhood of only a few training instances, where those training instances are in each other's neighborhood.

4.1 Locality Preserving

In the generic classification framework, training samples in one class are always all used to learn parameters, except a few cases, such as SVM only uses the support vectors. Suppose a test image is input, the linear reconstruction only takes the training images which lie in the neighborhood of the test image, assuming the test image and the training images are in its feature space manifold. It is because the far-away images have little local semantic similarity with the test image. Consequently the reconstruction coefficients appears non-zero only at those neighborhood images.

While formulating the training dataset, we always ensemble the feature vectors into feature matrix. Here we also ensemble those feature vectors in one group to form a group matrix A. But if they are put together in random order or certain sorted order, there is no guarantee from feature space to feature matrix the neighborhood property is preserved. Thus we seek a graph based algorithm to rearrange the feature vectors while keeping the local structure.

Given a graph $G = (V, E)$ as Fig. 5 shows, the distance from one node to every other node, $d(i, j)$ denotes the dissimilarity of the pair of nodes and each node represents one image or feature vector. The starting node could be any node. Assume we start at node 1. We would like to search all node 1's neighborhood nodes and put them following feature vector 1 in matrix A. But there may be ambiguity as node 2 is the neighbor of both node 1 and node 3. The order to put feature vector 2 and feature vector 3 is not determined. Thus we decide to take in one node each time that the node is the nearest to the current node set S. Initially $S = \{n_1\}$. Suppose n_2 is the nearest to n_1, $S = \{n_1, n_2\}$. We then consider which node is the nearest to either n_1 or n_2. Such procedure continues until all feature nodes are visited. We formulate the procedure as Algorithm. 1.

4.2 Sparse Group Lasso

As discussed above, we would like to achieve the linear reconstruction of test images using the neighborhood training images. Those reconstruction coefficients must be sparse and group sparse since the training instances must be in neighborhood and clustered together. Thus, we formulate the classification problem as choosing the smallest reconstruction error class label with our group sparse learning coefficients.

Algorithm 1 Locality preserving by rearranging feature matrix.

```
Input: $S = \{n_1\}$, $S' = \{n_2, ...n_N\}$, number of nodes $N$,
Start Node = $n_1$.
Output: rearranged feature matrix $A$.
repeat
    find $n_k$, s.t. $min(d(n_i, n_j)) = d(n_i, n_k)$, $n_i \in S$ and $n_j, n_k \in S'$
    add $n_k$ to $S$, delete $n_k$ from $S'$
    $A = [S; n_k]$
    repeat
        switch $n_k$ and $n_t$, $t \in [i + 1, k]$
    until $d(n_i, n_t) < d(n_i, n_k)$
until $S' = \emptyset$
```

Suppose there are M classes in all. All the training images of the i^{th} class form the matrix A_i $(1 \leq i \leq M)$. Each A_i is rearranged by Algorithm.1. In other words, A_i keeps the neighborhood structure of the original image manifold. Denoting A as the matrix of all the images, we could depict such grouping relationship as Eq.4.

$$\begin{aligned} A_{ig} &= [a_{ig}^1, a_{ig}^2, ..., a_{ig}^k, ..., a_{ig}^{M_{ig}}] \in \mathbb{R}^{d\times M_{ig}} \\ A_i &= [A_{i1}, A_{i2}, ..., A_{ig}, ..., A_{iG_i}] \in \mathbb{R}^{d\times \sum_g M_{ig}} \\ A &= [A_1, A_2, ..., A_i, ..., A_M] \in \mathbb{R}^{d\times \sum_i \sum_g M_{ig}} \end{aligned} \tag{4}$$

Inside a class feature matrix A_i, it can be further grouped into G_i groups $(1 \leq g \leq G_i)$. Each group represents the neighborhood structure. Inside one such group of class i, A_{ig} consists of M_{ig} feature elements a_{ig}^k $(1 \leq k \leq M_{ig})$. Then we denote the reconstruction coefficients corresponding to A, A_i and A_{ig} as X, X_i and X_{ig} respectively. It is denoted as Eq.5.

$$\begin{aligned} X_{ig} &= [x_{ig}^1, x_{ig}^2, ..., x_{ig}^k, ..., x_{ig}^{M_{ig}}] \in \mathbb{R}^{M_{ig}\times 1} \\ X_i &= [X_{i1}, X_{i2}, ..., X_{ig}, ..., X_{iG_i}] \in \mathbb{R}^{\sum_g M_{ig}\times 1} \\ X &= [X_1^T, X_2^T, ..., X_i^T, ..., X_M^T] \in \mathbb{R}^{\sum_i \sum_g M_{ig}\times 1} \end{aligned} \tag{5}$$

Combing the reconstruction error and the regularization of group coefficients, we derive the objective function as Eq.6.

$$\arg\min_{X_i} \frac{1}{2}\|y - A_i X_i\|_F^2 + \lambda_1 \|X_i\|_1 + \lambda_2 \sum_{j=1}^{G_i} w_j \|X_{ij}\|_2 \tag{6}$$

Here λ_1 and λ_2 are two regularization parameters to control the coefficients' sparsity and group sparsity. Parameter w_j depicts the weights of different groups inside one class. If the distribution approximates uniform Gaussian, we could set w_j to be uniform distribution.

For each class i, we train coefficients X_i. Then the classification rule is pick the class label i such that it approaches the minimal reconstruction error, which is as Eq.7 shows.

$$\begin{aligned} l(y) &\Leftarrow \arg\min_i \{e(1), e(2), ..., e(i), ..., e(M)\} \qquad (7) \\ e(i) &= \frac{1}{2}\|y - A_i X_i\|_F^2 \end{aligned}$$

here $l(y)$ is the class label of test feature vector y.

We know that Eq.6 is convex. Thus we could use convex optimizers to solve it. Ref. [11] discussed the Moreau-Yosida regularization term group lasso algorithm in detail. The Moreau-Yosida regularization divides the coefficients into tree structure. The same layer's coefficients have no overlap. The children layer's coefficients are the subset of the parent layer. Here we just has one layer which contains the coefficients of different groups inside one class. The rest formulation is exactly the same as Ref. [11]. Thus we took it as the workhorse to optimize our objective function Eq.6.

5. EXPERIMENTS

5.1 Experimental Setup

COREL is a general purpose image database [5] consisting of 145000 images. In our experiment, we choose 5000 images from the collection which are of 50 semantic categories with each class having 100 images. These images are stored in JPEG format with size 384 × 256 or 256 × 384 and each image is represented in RGB color space. In Ref. [4], the experiment uses COREL2k which is the closest setup of our experiments. In COREL, each category contains 100 images. So COREL2k select 20 categories out of 1450 categories, while ours choose 50 out of 1450.

In order to achieve exactly the same experimental environment, we will use the same COREL2k for comparison with state-of-the-art method (Multiple Instance learning Classification MILc). In our proposed method, we randomly pick 10 images as test images. And the validation is leave-one-out, which is once we find a test image, the rest of the samples in the category are all used to train the reconstruction coefficients. The same validation method is taken by our comparison methods, e.g. k-NN and in-class sparse classification (in-CSc). The in-class sparse lasso method is similar as our proposed method. But its regularization term is just sparse, no group sparse constraint, which allows training samples not in the neighborhood to linearly represent the test image. For the baseline comparison method SVM, we randomly choose 90 images out of 100 images as training images for each class, the rest 10 images as test images. We then run multi-class SVM on our training dataset and obtain the classifier for 50 different categories.

In feature extraction aspect, we adopt GIST feature. GIST feature depicts texture information of images, which is also known as Gabor multi-scale multi-directional wavelets. In Gabor model with Euclidean distance measurement, noise is assumed Gaussian and thus feature difference is $l2$-norm based.

For parameter settings, our proposed method chooses $\lambda_1 = 10^{-2}$ and $\lambda_2 = 10^{-2}$. We pick the group weights as equal weight $w_j = \frac{1}{100/G_i}$. The group number G_i is uniformly decided as 5. For those comparison methods, k-NN we choose k as 5 in order to be consistent with our proposed group number; In-CSL we take $\lambda = 10^{-2}$ as well; while SVM we set it linear kernel with regularizer C decided by libsvm [2].

Table 1: Accuracy comparison

Algorithms	Precision(%)	Recall(%)
k-NN	49.2	43.8
SVM	48.7	48.4
in-CSc	41.7	40.5
LPGSc	62.4	58.9
Combined	67.6	69.8
$MILc^{[4]}$	68.7	-

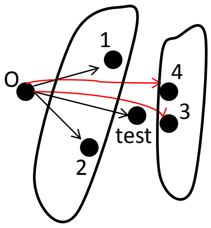


Figure 6: in-CSc shortcoming reasoning

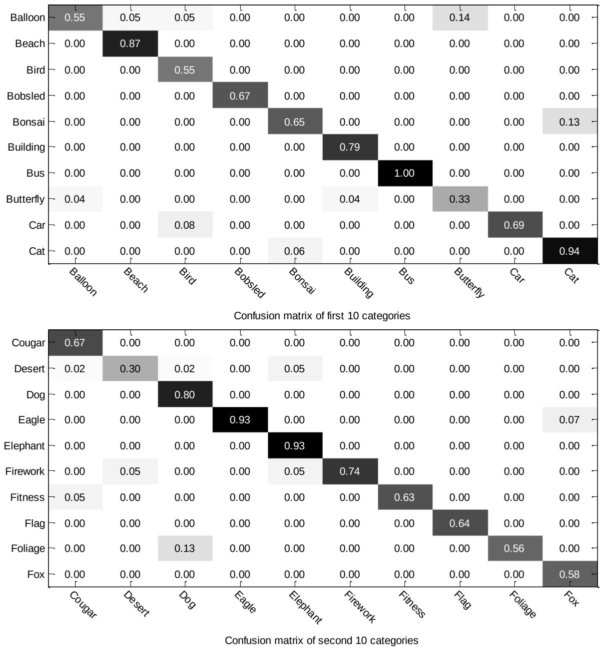
As most of the image categorization works do, we evaluate the performance by its precision of each category, recall, average precision (AP) and the confusion matrix across different categories. We denote precision as P, recall as R and AP as AP in Eq.8.

$$\begin{aligned} P_i &= \frac{\#\ true\ classified\ samples\ to\ Class\ i}{\#\ samples\ classified\ to\ class\ i} \\ R_i &= \frac{\#\ true\ classified\ samples\ to\ Class\ i}{\#\ samples\ class\ i\ has} \qquad (8) \\ AP &= \frac{1}{M}\sum_i P_i \end{aligned}$$

5.2 Experimental results

We compare our Locality Preserving Group Sparse Classification (LPGSc) with the following related work: (i) k-NN. (ii) SVM. (iii)in-class sparse Classification (in-CSc). As a test of boosting single classifier's accuracy, we tried to combine our LPGSc with k-NN. Firstly the proposed method gives the first 3 classes that have least reconstruction error. Then based on the 3 classes, k-NN is adopted to further give out the final class label. Such classifier combination method is also compared in our experiments. (iv) classifier combined method. (v)MILc in Ref [4].

Table.1 shows the average precision and recall over six different methods. Our proposed method is LPGSc and the combined method is our trial expecting to boost the performance. The rest are either baseline comparison methods or state-of-the-art method. It reveals that our proposed method outperforms about 12% to 15% comparing to k-NN and SVM baseline methods. The in-CSc method is below the baseline methods. Usually the training sample space is not Gaussian distributed and it is not homogeneous in different dimensions. Euclidean distance evaluating the k nearest neighbors has bias from the true nearest neighbors. Nevertheless, such training sample space is not linearly separable.



	Balloon	Beach	Bird	Bobsled	Bonsai	Building	Bus	Butterfly	Car	Cat
Balloon	0.55	0.05	0.05	0.00	0.00	0.00	0.00	0.14	0.00	0.00
Beach	0.00	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Bird	0.00	0.00	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Bobsled	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00
Bonsai	0.00	0.00	0.00	0.00	0.65	0.00	0.00	0.00	0.00	0.13
Building	0.00	0.00	0.00	0.00	0.00	0.79	0.00	0.00	0.00	0.00
Bus	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Butterfly	0.04	0.00	0.00	0.00	0.00	0.04	0.00	0.33	0.00	0.00
Car	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.69	0.00
Cat	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.94

Confusion matrix of first 10 categories

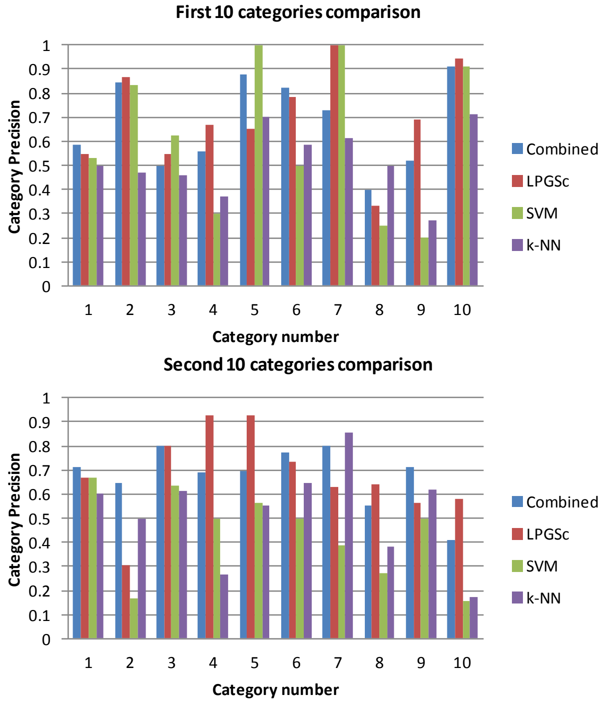
	Cougar	Desert	Dog	Eagle	Elephant	Firework	Fitness	Flag	Foliage	Fox
Cougar	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Desert	0.02	0.30	0.02	0.00	0.05	0.00	0.00	0.00	0.00	0.00
Dog	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Eagle	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.00	0.07
Elephant	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.00
Firework	0.00	0.05	0.00	0.00	0.05	0.74	0.00	0.00	0.00	0.00
Fitness	0.05	0.00	0.00	0.00	0.00	0.00	0.63	0.00	0.00	0.00
Flag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.64	0.00	0.00
Foliage	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.56	0.00
Fox	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.58

Confusion matrix of second 10 categories

Figure 7: Confusion matrix of the first 10 and second 10 categories by LPGSc

It is hard to find suitable kernel to translate the sample space into designed kernel space. Thus SVM performs poorly over such training space. The reason in-CSc performs even more poorly than baseline methods is as Fig.6 shows. Suppose node O is origin point. Node 1 and node 2 belong to wrong category, which are far away from each other. Node 3 and 4 belong to the true category, which are in neighborhood of each other. To linearly reconstruct the test node, from the picture we know that the reconstruction error node 1 and 2 is less than node 3 and 4. Thus according to the minimum reconstruction error rule, the in-CSc predicts wrong class label. Either the criteria rule is not suitable here or the model is not suitable here. We notice that if node 1 and 2 clusters together, the linear reconstruction would be accurate as node 3 and 4 do. That's one of the promotion of our proposed algorithm, which is to constrain the training samples in the neighborhood. The combined method is about 5% increase from our proposed method, which is at the same level of state-of-the-art[4] result. Since the MILc uses COREL2k, here we prepare exactly the same environment for combined method. Fig.7 reveals first 10 and second 10 categories' confusion matrices. Actually we calculated the whole 50 × 50 confusion matrix of the whole dataset. Then we truncate the first and second 10 categories submatrices out of the original confusion matrix. Diagonal of the two matrices are dominant in each row. Though there may be as low as 0.30 accuracy, 70% of the diagonal elements are above 0.6, which indicates that the proposed classifier behaves above the average precision (62.4%) for most of the categories.

Besides the comparison over the average precision as Table.1 shows, we wonder the performance on each category by different methods. Fig.7 demonstrates the performance on our proposed method. Fig.8 compares the combined method, proposed method, SVM and k-NN with the same experimental environment. From the cylinder plot, 14 out of 20 categories, our proposed method outperforms k-NN and SVM. The combined method achieves 15 out of 20 to dominate the performance. SVM only takes 2 categories and k-NN takes 4 categories in better performance. Hence, the experimental result reflect that our proposed method performs over 10% better than baseline methods in precision and in most categories LPGSc takes advantage than the baseline approaches. Moreover, our classifier-combined strategy has already achieves as good as the state-of-the-art methods.


First 10 categories comparison

Second 10 categories comparison

Figure 8: Accuracy comparison of methods for the first 10 and second 10 categories

Finally we give out some visual classification results in Fig.9. The most left column are test images. In the same row, we chose 5 training images with the largest corresponding group sparse coefficients. We can use the group sparse coefficients and those listed training samples to reconstruct the test image with least square error. Since our extracted feature is GIST, which is texture based feature, the chosen training images are not color related but texture related. The images with highest coefficients below should be most relevant to the test images. From the visual results, we could see that proposed classifier correctly pick the most relevant training images as the neighborhood images of test instances, where the salient coefficients are clustered in group.

6. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a novel locality preserving and sparse group lasso based framework for image categorization. Locality preserving strategy maintains the neighborhood property from original training sample space. Group sparse learning further searches the sparse and group sparse coefficients to represent the test sample by the neighborhood structure preserved from the first step. Experimental results

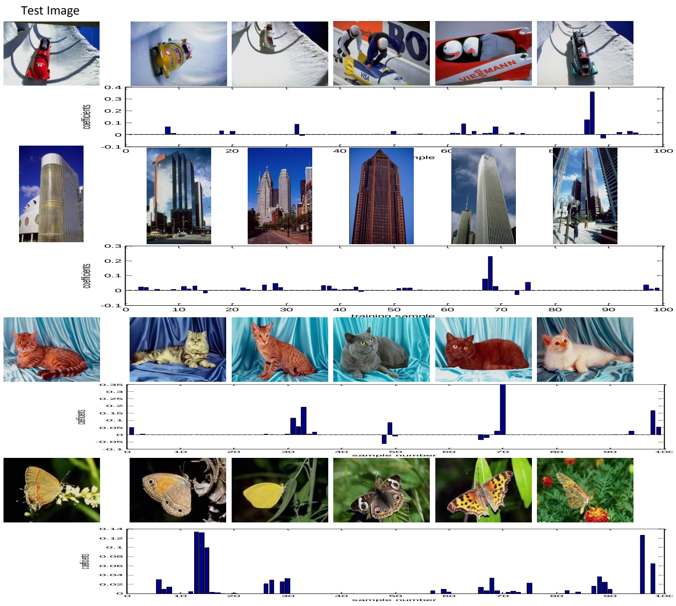


Figure 9: Visual results of group sparse representation

demonstrate our proposed method outperforms the baseline methods and our classifier-boosted method could approach approximately as good as state-of-the-art method. We need to investigate more efficient features not only global but also local features. Since feature level plays an important role in depicting images. Efficient feature boosts performance significantly. Furthermore, classifier combination should be refined aiming to capture as much information from different classifiers as possible. And as our workhorse solving convex functions is not real-time, more effective algorithms to learn the coefficients should also be considered in the future work.

7. REFERENCES

[1] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. In *NIPS*, 2009.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 2011.

[3] O. Chapell, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *IEEE Trans. on Neural Networks*, 10:1055–1064, 1999.

[4] Y. Chen, B. J, and W. J.Z. Miles: Multiple instance learning via embedded instance selection. *TPAMI*, 2006.

[5] COREL. http://www.corel.com/products/clipartandphotos.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV SLCV Workshop*, 2004.

[7] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. 2010.

[8] S. Gao, L.-T. Chia, and I.-H. Tsang. Multi-layer group sparse coding-for concurrent image classification and annotation. In *CVPR*, 2011.

[9] M. Gorkani and R. Picard. Texture orientation for sorting photos 'at a glance'. In *ICPR*, 1994.

[10] J. Huang, S. Kumar, and R. Zabih. An automatic hierarchical image classification scheme. In *ACM Multimedia*, 1998.

[11] J. Liu and J. Ye. Moreau-yosida regularization for grouped tree structure learning. *NIPS*, 2010.

[12] O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. In *ICML*, 1998.

[13] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3), 2001.

[14] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[15] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *TPAMI*, 2006.

[16] J. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. on PAMI*, 23:947–963, 2001.

[17] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[18] X. Yuan and S. Yan. Visual classification with multi-task joint sparse representation. In *CVPR*, 2010.

[19] Q. Zhang, S. A. Goldman, W. Yu, and J. Fritts. Content-based image retrieval using multiple instance learning. In *ICML*, 2002.

[20] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and M. Dimitris. Automatic image annotation using group sparsity. In *CVPR*, 2010.