From Local Similarity to Global Coding; An Application to Image Classification

Amirreza Shaban[†], Hamid R. Rabiee[†], Mehrdad Farajtabar[‡], Marjan Ghazvininejad[*]

[†]Department of Computer Engineering, Sharif University of Technology

[‡]College of Computing, Georgia Institute of Technology

[*]School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne

shaban@ce.sharif.edu, rabiee@sharif.edu, mehrdad@gatech.edu, marjan.ghazvininejad@epfl.ch

Abstract

Bag of words models for feature extraction have demonstrated top-notch performance in image classification. These representations are usually accompanied by a coding method. Recently, methods that code a descriptor giving regard to its nearby bases have proved efficacious. These methods take into account the nonlinear structure of descriptors, since local similarities are a good approximation of global similarities. However, they confine their usage of the global similarities to nearby bases. In this paper, we propose a coding scheme that brings into focus the manifold structure of descriptors, and devise a method to compute the global similarities of descriptors to the bases. Given a local similarity measure between bases, a global measure is computed. Exploiting the local similarity of a descriptor and its nearby bases, a global measure of association of a descriptor to all the bases is computed. Unlike the locality-based and sparse coding methods, the proposed coding varies smoothly with respect to the underlying manifold. Experiments on benchmark image classification datasets substantiate the superiority of the proposed method over its locality and sparsity based rivals.

1. Introduction

Image classification, i.e. the task of assigning an image to a class chosen from a predefined set of classes, has gained much attention in recent years. Most of the recent works in this area can be categorized into three groups based on the utilized model. These models include part based models [8], bag of words(BoW) models [5] and attribute-based models [16]. Among these, BoW models, which are based on the representation of affine invariant descriptors of image patches, have proved to have great performance and are widely used in many applications such as image classification [15], image retrieval [20], and human pose estimation [1].

In spite of recent advancements, image classification remains a challenging task. The complexity is caused by many factors such as background clutter and highly nonlinear variations in object appearance such as pose, illumination, and occlusions.

The BoW model is based on representing features extracted from local patches of an image. Researchers have empirically found that, assigning each feature to nearby bases leads to remarkable improvement in accuracy. Authors in [25] proved that under the manifold assumption, considering local bases in coding is essential for successful nonlinear feature learning. Developing this idea, Wang et. al. [23] in their LLC method use k-nearest neighbor bases in the coding process and set the coding coefficient for other bases to zero. Since the manifold is locally linear, a linear similarity measure is used for neighboring bases. Although this coding scheme captures the local manifold structure, it's not capable of binding this information to derive and utilize the global structure of the manifold. To be more specific, two features that have different bases in their neighborhood generate completely different codings independent of their distance on the manifold.

To overcome this drawback, we propose a novel method called Local Similarity Global Coding (LSGC), that uses the local similarities between bases to obtain a nonlinear global similarity measure between local features and bases. We first show that this coding scheme captures the global manifold structure and generates a smoother coding compared to LLC. Next, we formulate the coding as a linear transformation of any local coding (which is obtained by an arbitrary local coding scheme such as LLC). This formulation is of practical interest when the transformation from local to global coding is obtained by matrix-vector multiplication. The experimental results show that our method outperforms the state-of-the-art on several benchmark datasets.

The rest of the paper is organized as follows. In Section 2, we describe basic aspects of an image classification system. section 3, discusses related works. In Section 5, we introduce our method, Finally, in section 6, experimental results on several benchmark datasets are reported. Finally

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.360

we conclude in section 7.

2. Basic and Notations

In this section, we review aspects of a state-of-the-art image classification system. The flowchart of this process can be seen in figure 1. First, local points in the image are selected or densely sampled, and an affine invariant feature vector x_i called the descriptor vector is extracted from each local point. Among feature extraction methods, SIFT [18] and HoG are the most commonly used as in [23, 3, 15]. Each descriptor is represented according to elements in a codebook B. Each of these elements are called coding vectors u_i. Columns of B are salient features in the image. Different algorithms use different codebook learning and coding schemes.

To aggregate the information of different local codings into one feature vector, local codings from image patches are merged together using a predefined pooling function:

$$v = \mathbf{F}(U) \tag{1}$$

where the i^{th} column of U is the coding vector u_i, $\mathbf{F}$ is the pooling function and v is the image feature vector. Different pooling functions producing different feature vectors are used in literature. Among them are max pooling [24], sum normalization [19], sum pooling and ℓ^2 normalization. However, recent work empirically shows that the max pooling function leads to superior performance [24, 23]. The max pooling function can be defined as:

$$v_j = \max(|u_{1j}|, |u_{2j}|, \ldots, |u_{lj}|) \tag{2}$$

where u_{ij} is the j^{th} element of u_i and l is the number of local points for each image.

In the last step, image features v are used for classification. A typical choice is the SVM classifier with Mercer Kernel such as linear kernel, intersection kernel or Chi-square kernel.

The method described so far does not take into acount the spatial information of the local points. Following the procudure suggested in [15] each image is divided into $2^\ell \times 2^\ell$ subregions for $\ell = 0, 1, \ldots$ and temporal features are computed by applying the pooling function to each region. The final feature vector is represented by concatenating all the temporal features.

Recent works mainly differ from each other in their dictionary learning and coding schemes. We pay close attention to these aspects in the following section.

In the rest of the paper consider base b_i as the i^{th} column of dictionary matrix B which has a total of c columns. x_i and u_i are the local feature and corresponding coding for the i^{th} local keypoint respectively.

3. Related Work

In this section, we review commonly used methods in coding and dictionary learning for image classification. Inspired by the success of BoW in text categorization, authors in [5] used BoW for image classification task. In this method the codebook is the cluster centers that are learned using k-means. Vector quantization (VQ) is used to generate coding. Therefore, each code has only one non-zero element that indicates to which cluster the vector x_i belongs. An SVM with nonlinear kernel is used for classification.

In ScSPM [24] the VQ constraint is relaxed in such a way that each local feature can be represented by a few number of bases. The objective function is defined as:

$$\begin{aligned} &\min_{U,B} \sum_{i=1}^{n} (||x_i - Bu_i||^2 + \lambda |u_i|) \\ &\text{subject to } ||b_k|| \leq 1, k = 1, 2, \ldots, c \end{aligned} \tag{3}$$

where x_i and u_i are descriptor and coding vectors of the i^{th} local point respectively. n is the total number of local points and there are c bases in the dictionary B. The first term represents the reconstruction error and the second term controls the sparsity of coding u_i. λ balances the trade-off between reconstruction error and sparsity. The sparsity prior plays a key role in coding, because it ensures that the coding captures outstanding patterns in local features. Besides, the reconstruction error in this method is less than that of VQ coding. Classification is performed using a linear SVM that surpassed the state-of-the art performance of its time. Therefore, the complexity of $O(n^3 \sim n^2)$ in training and $O(n)$ in testing is reduced to $O(n)$ and constant time respectively.

Although ScSPM proves its performance it has one a major drawback: the coding does not change smoothly when x_i varies on the manifold. LScSPM [10] tries to overcome this problem by using manifold assumption. Its objective function is:

$$\begin{aligned} &\min_{U,B} \sum_{i=1}^{n} (||x_i - Bu_i||^2 + \lambda |u_i|) + \beta \sum_{ij} ||u_i - u_j||^2 w_{ij} \\ &\text{subject to } ||b_k|| \leq 1, k = 1, 2, \ldots, c \end{aligned} \tag{4}$$

where w_{ij} denotes the similarity between local features i and j. This objective function differs from standard sparse coding in the regularization term, which guarantees that the sparse code varies smoothly on the data manifold. The interpretation of smoothness term is that when w_{ij} for two local feature is high, their codings must be close in Euclidean space. Despite the novelty of the idea, the optimization is hard to tackle due to the large quantity of local features. Therefore, the execution time for this algorithm is

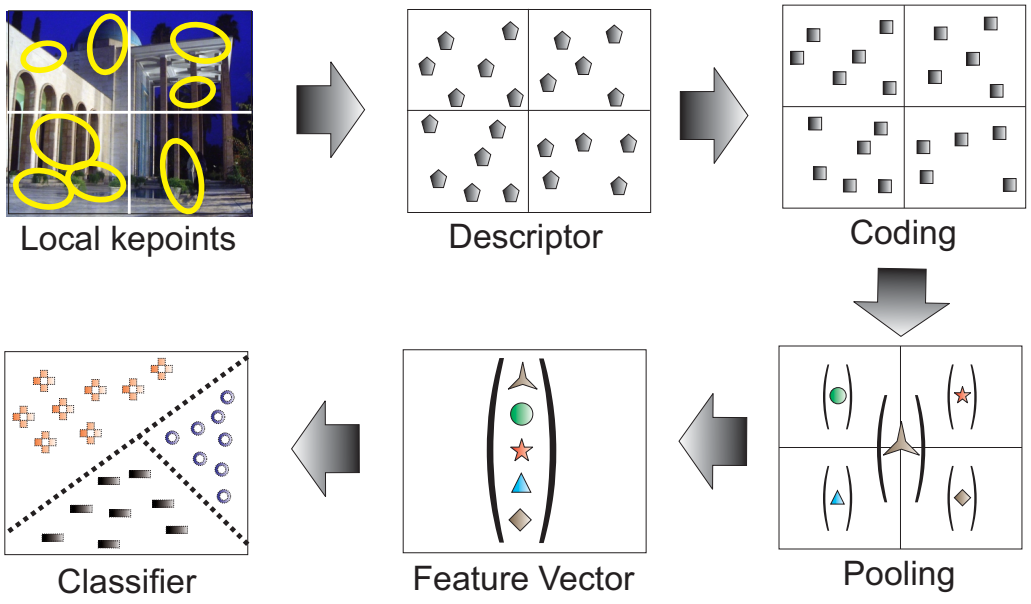


Figure 1: BoW image classification system

very great, and it is impractical for many real-world applications.

As suggested in [23], locality is more important than sparsity. The fundamental assumption in their method is that the local features lie on a nonlinear m dimensional manifold where m is less that the dimension of the ambient space. The LLC [23] is also based on this assumption. Mentioning locality, brings into consideration the nonlinear structure of the data manifold in the coding process. The coding coefficient is obtained by solving:

$$\begin{aligned} &\min_{U,B} \sum_{i=1}^{n} (||x_i - Bu_i||^2 + \lambda ||d_i \odot u_i||^2) \\ &\text{subject to } ||b_k|| \leq 1, k = 1, 2, \ldots, c \\ &\mathbf{1}^\top u_i = 1 \end{aligned} \tag{5}$$

where $\odot$ denotes pairwise multiplication. Let $dist(x_i, b_j)$ denote the Euclidean distant between local feature i and basis j. Elements of d_i are given by:

$$d_{ij} = \exp\left(\frac{dist(x_i, b_j)}{\sigma}\right) \tag{6}$$

σ is a parameter that controls locality. In practice the second term is ignored and coding for each descriptor is obtained by optimizing only the first term using only k-nearest bases. This leads to non-zero coefficients for the k-nearest bases and zero for the others. The remarkable success of LLC supports the assumption that data are laid on the manifold.

Localized soft-assignment coding [17] expresses the coding coefficient as the probability that a local feature x_i belongs to a basis b_j and surpasses the performance of LLC. Its local similarity measure is defined as:

$$\begin{aligned} p_{ij} &= \frac{\exp\left(-\beta \hat{d}(x_i, b_j)\right)}{\sum_{l=1}^{n} \exp\left(-\beta \hat{d}(x_i, b_j)\right)} \\ \hat{d}(x_i, b_j) &= \begin{cases} \hat{d}(x_i, b_j) = dist(x_i, b_l) & \text{if } b_l \in \text{k-NN}(x_i) \\ \infty & \text{otherwise} \end{cases} \end{aligned} \tag{7}$$

where $dist(x_i, b_j)$ is a distance (e.g. Euclidian) in the ambient space. Similar to LLC, coefficients for the k-nearest bases are computed and the others are set to zero.

Local coding methods like LLC and soft-assignment coding implicitly give regard to manifold structure, since local similarity a is valid approximation only for neighboring points. However, these methods disregard global similarities between data, which could be captured using nonlinear similarity estimation methods.

4. Motivation

Recent image classification methods that look at both reconstruction error and locality in dictionary learning prove to have top-notch performance [23]. Looking at locality is a struggle to take the underlying nonlinear structure of local features into account. Locality ensures that nearby bases are preferred in coding data points, and this implicitly dis-

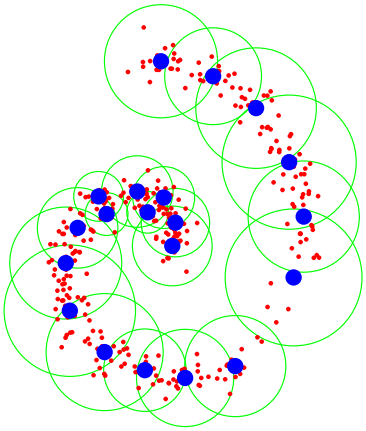


Figure 2: Kmeans dictionary learning. The blue circles are cluster centers and red circles are descriptors. The bases inherit the geometry of descriptors

criminates in favor of the bases on the underlying manifold. Since bases are usually samples of the manifold, the distance (or similarity) of data to these bases is an appropriate feature that embodies the geometry of the data.

Usual coding methods learn the bases by considering the manifold structure either explicitly or implicitly. To elaborate, we refer to a closely related trend in large-scale and online manifold learning literature that tries to find only a few bases in order to best preserve the manifold structure. These methods rely on quantization [21], sampling [11], and coarse graining [6] to reduce the number of data points and simultaneously minimize geometric information loss. Authors in [26] proposed a method with desirable theoretical properties, which quantizes the support of data into a mixture of Gaussians that cover the manifold. One can easily see that k-means is an especial case of the mixture model when the covariance matrix is an identity matrix. Thus, the bases which are learned by k-means trace the manifold structure of local features. Figure 2 illustrates how the bases learned by k-means cover the whole structure of data. The centers of the k-means can be viewed as samples of the data that inherit the underlying geometry.

The fact that these bases trace the manifold is a motivation to take the natural similarity between bases into account when coding as well. This leads to a better utilization of manifold structure in the coding process. Exploiting the non-linear dependence of bases to each other a framework is proposed in order to find a global coding scheme for a descriptor.

5. Proposed Method

Methods that take into account the manifold structure use only the k nearest bases in the coding process. This is due to the fact that the Euclidian distance in the ambient space is valid only for nearby points. In this paper we present a novel algorithm extending the methods which rely only on local similarities between data and bases. We claim that local similarities between bases are valuable in the sense that they can be used to estimate global similarities between local features and bases.

Considering bases learned by k-means, a local similarity between bases is proposed, which is then utilized to find a global similarity with regard to the manifold structure. At last a coding scheme is presented to derive global similarities between descriptors and bases.

5.1. Local Similarity

Choosing the similarity measure is arbitrary and the approach taken by any existing method (Gaussian kernel, LLC, Sparse Coding) can be adopted. We take the Gaussian kernel approach which is commonly used as a local similarity measure in the manifold learning methods [2]:

$$W(i,j) = \begin{cases} \exp(-\frac{||b_i - b_j||^2}{\sigma}) & \text{if } b_j \in \text{k-NN}(b_i) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

While W captures only local similarities, in the next subsection we propose a probabilistic framework to find a global measure of the probability that a base belongs to other bases.

5.2. From Local to Global Similarity

Given a matrix W that contains local similarities between bases, stochastic matrix P is defined by normalizing W:

$$P = D^{-1}W \tag{9}$$

where D is a diagonal matrix and $D_{ii} = \sum_j W_{ij}$. The normalization ensures $\sum_j P_{ij} = 1$. Since W contains the similarities, P may be seen as a stochastic matrix where $P_{ij} = p(b_j|b_i)$ expresses the a posteriori probability that base j belongs to i. $p(b_j|b_i)$ can be interpreted as the probability that b_j is a member of a Gaussian distribution with mean b_i and variance σ.

As matrix P measures the similarities between neighboring bases, the similarity between non-neighboring bases can be computed indirectly by random walks on the graph which has the adjacency matrix W [13]. Suppose $p^{(2)}(b_k|b_i)$ represents indirect belonging of b_i to b_k which is not among b_i's neighbors. Superscript 2 means an indirect dependence via 2 steps:

$$p^{(2)}(b_k|b_i) = \sum_{l=1}^{c} p(b_k, b_l|b_i) = \sum_{l=1}^{c} p(b_k|b_l)p(b_l|b_i) \tag{10}$$

Therefore, elements of the matrix P^2 are indirect similarities of order 2. Similarities of higher orders are defined in the same manner, i.e. for the similarity of order t we use similarity of order $t-1$:

$$p^{(t)}(b_k|b_i) = \sum_{l=1}^{c} p(b_k|b_l)p^{(t-1)}(b_l|b_i) \tag{11}$$

One can easily see that matrix P^t captures the similarities of order t.

This representation is crucially affected by the locality-scale parameter t. Although for every t, P^t can be regarded as a measure of non-local similarity, a better measure of dependence of basis j on basis i can be defined as:

$$S = \frac{1}{t}\sum_{m=0}^{t-1} P^m, \tag{12}$$

which considers a multi-resolution non-local dependence from very local to more global ones. S can be regarded as a new probability measure of dependence for the bases. In fact $S_{ij} = p^{(<t)}(b_j|b_i)$, which is the probability that b_i belongs to b_j considering locality of order 0 to $t-1$.

5.3. LSGC coding

Suppose a new descriptor, x_i, has to be encoded. First its local coding u_i is computed via its k nearest bases:

$$u_{ij} = p(b_j|x_i) = \begin{cases} \frac{1}{Z}\exp(-\frac{||x_i-b_j||^2}{\sigma}) & \text{if } b_j \in \text{k-NN}(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where Z is a normalizing constant. It is the same as used by soft-assignment coding in [17]. Any other local coding scheme may be used alternatively.

Belonging of x_i to a non-neighboring basis b_k can be computed indirectly using the global similarity between bases and the local coding of x_i:

$$p(b_k|x_i) = \sum_{l=1}^{c} p^{(<t)}(b_k|b_l)p(b_l|x_i) = \sum_{l=1}^{c} S_{lk}u_{il} \tag{14}$$

which may be equivalently written as:

$$g_i = S^\top u_i, \tag{15}$$

where g_i is the global coding we acquire for the descriptor x_i.

The overall algorithm is shown in Algorithm 1.

Now we mention some nice aspects of our proposed coding:

Remark 1. The especial case $t = 1$ in equation (12) leads to the same coding as that of the previous works which consider local similarities. Larger ts lead to more global measures. In the extreme case, when $t \to \infty$ all the bases become indistinguishable in the term of similarity. Sufficiently large t should be selected based on the resolution at which we look at the locality.

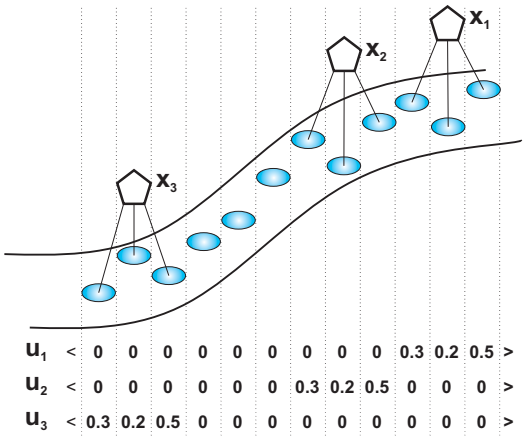
Algorithm 1 *LSGC Coding*

Input: bases b_j, locality parameter t, descriptors x_i, bandwidth parameter σ
Output: global coding for each descriptor g_i

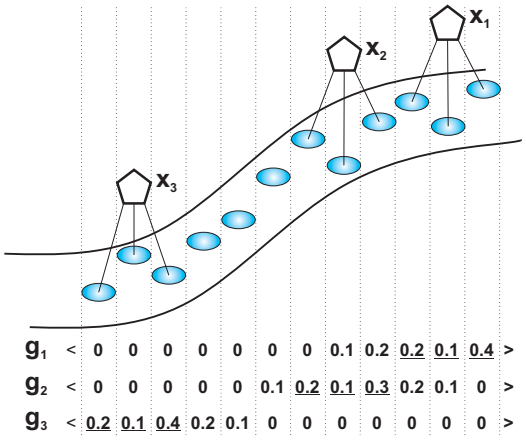
$W(i,j) \leftarrow \begin{cases} \exp(-\frac{||b_i-b_j||^2}{\sigma}) & \text{if } b_j \in \text{k-NN}(b_i) \\ 0 & \text{otherwise} \end{cases}$
{Normalize W}
$D_{ii} \leftarrow \sum_j W_{ij}$
$P \leftarrow D^{-1}W$
{compute the global similarity measure between bases}
$S \leftarrow \frac{1}{t}\sum_{m=0}^{t-1} P^m$
for all x_i **do**
  { Compute its local coding $p(b_j|x_i)$}
  $u_{ij} \leftarrow \begin{cases} \frac{1}{Z}\exp(-\frac{||x_i-b_j||^2}{\sigma}) & \text{if } b_j \in \text{k-NN}(x_i) \\ 0 & \text{otherwise} \end{cases}$
  {Compute its global coding}
  $g_i \leftarrow S^T u_i$
end for

Remark 2. Equation (15) is a linear transformation on the previously computed local coding. It is surprising how a linear transformation can encode the descriptors considering the nonlinear geometry of the data. In fact, S itself is built based on a non-linear transformation of the matrix P and stores our prior belief about the geometry or distribution of bases along the manifold. Moreover it is of practical interest, because global coding of every descriptor can be calculated efficiently via a transformation by a precomputed matrix S.

Remark 3. Our method is superior to conventional methods which try to consider manifold structure of data by solely using k-NN to construct the local coding. Our method leads to a coding which varies smoothly with respect to the manifold. For illustration consider figure 3. Three descriptors x_1, x_2, and x_3 are going to be coded. For the methods like LLC and soft-assignment coding which only consider k nearest bases (figure 3a) $d(u_1, u_2) = d(u_1, u_3)$; or equivalently similarity of coding u_1 to u_2 is the equal to similarity of u_1 to u_3. However, by considering the manifold structure x_1 and x_2 are closer compared with the pair x_1 and x_3. The proposed method in figure 3b overcomes this shortcoming by propagating similarities along bases, so g_1 and g_2 share more nonzero elements i.e., $d(g_1, g_2) < d(g_1, g_2)$.



(a) Local coding



(b) Global coding

Figure 3: Comparing local and global coding scheme.

6. Experiments

6.1. Hand-written Image Classification

In this study we aim to compare the classification performance of a Linear SVM with different coding schemes. We compare LSGC to LLC [23], soft assignment coding (SAC) [22], and localized soft assignment coding (LSAC) [17] on different benchmark image and hand-written letter datasets. Table 1 summarizes the characteristic of each dataset. Each feature vector is normalized to have length of 1. In SAC method for each data point x, despite localized soft assignment coding, all coding coefficients are computed. In this experiment, LSAC is equivalent to LSGC with $t = 1$. The parameters are set by 5-fold cross-validation. In LSGC, we set k to one the best result is achieved for $t = 1$ in cross-validation. We then fix k and increment t until the best accuracy is achieved in cross-validation. In all the methods, at most 4000 points are used for learning a dictionary containing 1000 atoms. Table 2 reports the average accuracy and standard deviation in 20 independent runs with 20 labeled points per a class. The value of t obtained by cross-validation are reported for each dataset. In Table 2, with the exception of Letter dataset, LSGC with t larger than one improves the accuracy by propagating the coefficients. However both LLC and LSAC use the locality constraint, LLC does better in most cases. It can be concluded that on empirical data, representing coefficient by the reconstruction term is more efficient compared to the Gaussian kernel. In general, LSGC outperforms other coding methods with small numbers of labeled points.

6.2. Natural Image Classification

To evaluate our method, we compare LSCG to other methods in the literature using two benchmark datasets: CIFAR-10, Caltech-101.

In the preprocessing step, each image is converted to a gray-scale and its size is reduced to be less that 300 pixels in both width and height. We use SIFT descriptors extracted from one level 8×8 pixel patches, where the center of each patch lies on a grid with step size of 4 pixels. After obtaining the descriptors, the codebook is learned by using k-means clustering . Codebook sizes are fixed to 1000 in each dataset. Employing the max pooling method, we obtain the temporal features. We use SPM with $l = 0, 1, 2$ to calculate the final feature vectors. The parameter t is fixed to 3 for all experiments. We consider five nearest neighbors in coding process and the bandwidth size parameter σ is set to the mean of standard deviation of the bases.

To evaluate the sensitivity of our method to the training size, we calculate the accuracy of our algorithm with different training sizes. After feature extraction the linear SVM is used to classify the test data points. Results are reported under 10 independent runs on each dataset. The LLC implementation is provided by the authors.

6.2.1 CIFAR-10

CIFAR-10 [14] contains 60000 natural images in 10 categories. We use test batch that consists of 10000 images. To evaluate the effect of parameter t, the results are reported for locality steps $t = 1, 2, 3, 4$. Note that for $t = 1$, our method is reduced to soft-assignment coding [17]. For training, we randomly sample 25, 50, 75, 100, 200 data from each class. To have a fair comparison results of LLC [23] are reported in the same setting in Table 3.

The results show the superiority of LSGC. This is due to our method takes the geometry of descriptors into account. Increasing t from 1 to 3 results into an increase in accuracy, however, saturates for $t > 3$. It is emprically seen that for $t > 3$ all the coefficients are nonzero, i.e. the local similarity propagate sufficiently on the whole manifold. This

Table 1: Characteristics of datasets.

Dataset Name	#Instances	#Attributes	#Classes	Feature Set
COIL20	1440	1024	20	32×32 raw image
COIL [4]	1500	241	6	-
Digit [9]	5620	64	10	-
pendigits [9]	10992	16	10	points regularly spaced in time
USPS	11000	256	10	-
Letter [9]	20000	16	26	statistical moments and edge counts
MNIST	60000	784	10	28×28 raw image

Table 2: (Prediction Accuracy rate $\pm$ standard deviation) with 20 labels per class.

Dataset	Linear SVM	LLC	SAC	LSAC	LSGC	LSGC t
COIL20	0.953 ± 0.009	0.983 ± 0.007	0.965 ± 0.007	0.965 ± 0.008	$\mathbf{0.990 \pm 0.005}$	5
COIL	0.721 ± 0.028	0.818 ± 0.017	0.803 ± 0.020	0.804 ± 0.017	$\mathbf{0.850 \pm 0.028}$	35
Digit	0.924 ± 0.007	0.960 ± 0.006	0.952 ± 0.008	0.954 ± 0.009	$\mathbf{0.975 \pm 0.003}$	35
pendigits	0.866 ± 0.010	0.943 ± 0.007	0.945 ± 0.009	0.946 ± 0.010	$\mathbf{0.951 \pm 0.011}$	3
USPS	0.809 ± 0.010	0.891 ± 0.012	0.827 ± 0.017	0.834 ± 0.016	$\mathbf{0.902 \pm 0.007}$	32
Letter	0.629 ± 0.012	$\mathbf{0.718 \pm 0.006}$	0.707 ± 0.009	0.707 ± 0.009	0.708 ± 0.012	1
MNIST	0.789 ± 0.009	0.872 ± 0.007	0.824 ± 0.019	0.834 ± 0.016	$\mathbf{0.891 \pm 0.009}$	16

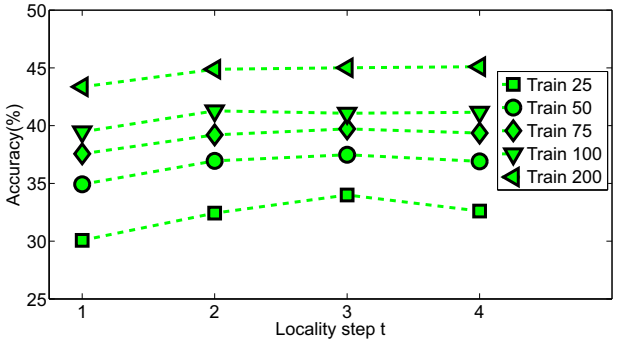


Figure 4: Effect of locality step parameter in CIFAR-10 accuracy. By increasing t over a threshold there will be no increase in the accuracy

phenomena is illustrated in Figure 4.

6.2.2 Caltech-101 Dataset

Caltech 101 [7] contains 9144 images of 101 classes including animals, vehicles, flowers, etc which is highly diverse. The number of images per category varies from 31 to 800. We examined the proposed algorithm on 5, 10, 15, 20, 25 and 30 training images per class. For comparison the results for several image classification method which is reported in the literature is illustrated in Table 4. We claim that the superiority of the results is due to considering global similarities in the coding process.

Table 4: Image classification accuracy(%) on Caltech-101

#Training	5	10	15	20	25	30
NN [3]	-	-	65.00	-	-	70.40
Griffin [12]	44.2	54.5	59.0	63.3	65.8	67.60
ScSPM [24]	-	-	67.0	-	-	73.2
LLC [23]	51.15	59.77	65.43	67.74	70.16	73.44
LSGC	**54.01**	**63.86**	**68.70**	**71.58**	**73.73**	**75.07**

7. Conclusion and Future Work

In this paper, we presented a method that uses the information about manifold structure of descriptors, to infer a global similarity measure between bases and descriptors. We showed that by using a linear transformation that embodies the manifold information, we can obtain global similarities from the local ones. In addition, by using global similarities between bases and descriptors in the coding process, a smoother coding is obtained compared to previous methods.

Our methods relies on the fact that the bases are sampling the data manifold which is done by k-means. Incorporating dictionary learning methods which take the manifold structure into account is remains as future work. Utilizing coarse graining algorithms which are sensitive to the geometry of the data is another open issue in our work.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *Pattern Analysis and Machine Intelli-*

Table 3: Image classification accuracy(%) on CIFAR-10

#Training	25	50	75	100	200
LLC [23]	31.16 ± 1.13	35.01 ± 0.78	37.50 ± 0.68	38.80 ± 0.77	42.17 ± 0.49
LSAC [17]	30.07 ± 0.96	34.92 ± 0.91	37.56 ± 0.78	39.46 ± 0.47	43.36 ± 0.42
LSGC(t = 2)	32.43 ± 1.19	36.95 ± 0.59	39.20 ± 0.52	41.28 ± 0.79	44.88 ± 0.41
LSGC(t = 3)	$\mathbf{34.00 \pm 1.15}$	$\mathbf{37.48 \pm 0.90}$	$\mathbf{39.72 \pm 0.71}$	41.07 ± 0.78	45.01 ± 0.22
LSGC(t = 4)	32.61 ± 0.75	36.91 ± 0.88	39.36 ± 0.68	$\mathbf{41.16 \pm 0.77}$	$\mathbf{45.10 \pm 0.46}$

gence, IEEE Transactions on, 28(1):44–58, 2006.

[2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006.

[3] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. Ieee, 2008.

[4] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22, 2004.

[6] M. Farajtabar, A. Shaban, H. Rabiee, and M. Rohban. Manifold coarse graining for online semi-supervised learning. *Machine Learning and Knowledge Discovery in Databases*, pages 391–406, 2011.

[7] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[8] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 380–387. IEEE, 2005.

[9] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[10] S. Gao, I. Tsang, L. Chia, and P. Zhao. Local features are not lonely–laplacian sparse coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3555–3561. IEEE, 2010.

[11] A. Goldberg, M. Li, and X. Zhu. Online manifold regularization: A new learning setting and empirical study. *Machine Learning and Knowledge Discovery in Databases*, pages 393–407, 2008.

[12] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[13] M. S. T. Jaakkola. Partially labeled classification with markov random walks. In *Advances in Neural Information Processing Systems 14: Proceedings of the 2002 Conference*, volume 2, page 945. MIT Press, 2002.

[14] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. Ieee, 2006.

[16] L. Li, H. Su, Y. Lim, and L. Fei-Fei. Objects as attributes for scene classification. In *ECCV First International Workshop on Parts and Attributes*, 2010.

[17] L. Liu, L. Wang, and X. Liu. In Defense of Soft-assignment Coding. 2011.

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[19] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. *Computer Vision–ECCV 2010*, pages 143–156, 2010.

[20] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. Ieee, 2007.

[21] M. Valko, B. Kveton, L. Huang, and D. Ting. Online semi-supervised learning on quantized graphs. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI*, 2010.

[22] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

[23] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.

[24] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801. Ieee, 2009.

[25] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. *Advances in Neural Information Processing Systems*, 22:2223–2231, 2009.

[26] X. Zhu and J. Lafferty. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 1052–1059. ACM, 2005.