Locality-Constrained Discriminative Learning and Coding

Shuyang Wang[1] Yun Fu[1,2]
[1]Department of Electrical & Computer Engineering,
[2]College of Computer & Information Science,
Northeastern University, Boston, MA, USA
{shuyangwang, yunfu}@ece.neu.edu

Abstract

This paper explores the enhancement by locality constraint to both learning and coding schemes, more specifically, discriminative low-rank dictionary learning and auto-encoder. Previous Fisher discriminative based dictionary learning has led to interesting results by learning more discerning sub-dictionaries. Also, the low-rank regularization term has been introduced to take advantage of the global structure of the data. However, such methods fail to consider data's intrinsic manifold structure. To this end, first, we apply locality constraint on dictionary learning to explore whether the identification capability will be enhanced or not by using the geometric structure information. Moreover, inspired by the recent advances from auto-encoders for learning compact feature spaces, we propose a locality-constrained collaborative auto-encoder (LCAE) for feature extraction. The improvement from applying locality to dictionary learning and auto-encoder is evaluated on several datasets. Experimental results have demonstrated the effectiveness of locality information compared with state-of-the-art methods.

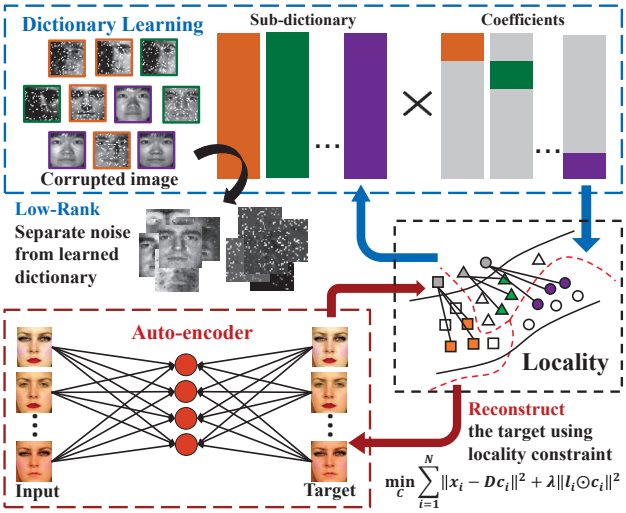

Figure 1. Illustration of our methods. The locality constraint is adopted in both dictionary learning (DL) and auto-encoder (AE) schemes. For DL, the negative effect of noise contained in training samples is narrowed by learning low-rank sub-dictionary. For AE, the target is reconstructed to be consistent with locality-constrained criterion.

1. Introduction

Recent researches have led to the rapid growth in the theory and application of sparse representation and demonstrated its promising results in face recognition and image classification etc. The key idea is to find a representation for each input signal using atoms from a given dictionary D as a linear combination. Thus, the quality of dictionary is a critical factor for sparse representation.

A problem arising with directly using the original training samples as the dictionary [26] is that, the test samples could not be faithfully represented owing to the noise and ambiguity in the dictionary. In addition, this strategy will ignore the discerning information hidden behind the training samples. Actually, the mentioned problems above can be solved by learning a proper dictionary from the original training samples. The intention of dictionary learning is to learn a set of basis from the training data where we could well represent the given signal. The recognition rate of image classification has been improved significantly with a well-adapted dictionary. A lot of research efforts have been made in order to seek a well-learned dictionary for distinctive representing the test samples. Recently, based on K-SVD [1], a discriminative constraint was added to the dictionary learning model that considers classification error in order to gain discriminability [31]. Jiang et al. enforced discerning ability by associating label information with each dictionary atom [8]. For learning a structured dictionary, the Fisher criterion was introduced to make sub-dictionaries according to different class labels [27].

The algorithms above, however, only work well for the

situation that the signals are clear or corrupted by small noise. If the training samples are corrupted with large noise, the dictionary atoms will be introduced corruptions resulting in representing the training samples.

Recently, low-rank representation [15] has been successfully applied to unsupervised subspace segmentation [14], object detection [23], and 3D visual recovery [29]. From corrupted input data, it determines a low-rank matrix. If a given matrix Y in which each atom shares the same pattern and corrupted by a sparse noisy matrix E, via rank minimization, Y could be practically recovered while sparse noisy E is removed. As for the case that using dictionary learning to deal with face recognition, the within-class samples are drawn from a low-dimensional subspace and linearly correlated. Therefore, each sub-dictionary for representing with-in class samples should reasonably be low-rank. Inspired by the previous work, low-rank regularization was integrated into sparse representation so that the sparse noises were separated from inputs while the dictionary atoms were simultaneously optimized to reconstruct the de-noised signals. The DLRD [16] algorithm achieves impressive results especially when corruption existed.

Previous sparse representation based approaches assume that each sample has independent sparse linear combination, which ignores the spatial consistency of neighbor points and fails to utilize the relationship between similar samples. Recent studies have witnessed more promising results using the idea of locality on the task of classification [25]. They presented method names Local Coordinate Coding (LCC), a modification to sparse coding, which theoretically proved that locality is more essential than sparsity under certain assumptions, and the coding is encouraged specifically to rely on local structure. Since then, several locality-constrained coding method has been proposed to replace sparse constraint on scene categorization [21], human action recognition [6] and image colorization [13] problems.

Motivated by above techniques, this paper explores the enhancement of classification by adding locality constraint on both learning and coding schemes, especially for discriminative low-rank dictionary learning and auto-encoder. First, an algorithm with low-rank regularization on discriminative sub-dictionary, and locality-constrained on coefficients is introduced. Second, different from previous locality linear coding works [22, 21], we study the locality on more complicated auto-encoder method to further study the performance of locality. A locality-constrained collaborative auto-encoder (LCAE) is proposed to extract feature with local information for enhancing the classification ability. our paper's main contributions are: 1) we investigate the impact of locality constraint on dictionary learning and improve the results on several benchmark datasets, 2) a locality-constrained collaborative auto-encoder (LCAE) is proposed to provide features with intrinsic local information.

The rest of this paper is structured as follows. Our proposed locality-constrained dictionary learning method and its optimization solution are presented in Section 2. Section 3 introduces the locality-constrained collaborative auto-encoder (LCAE) model. Section 4 shows the experiments and analysis along with the drawn conclusions in Section 5.

2. Locality-constrained Discriminative Low-Rank DL (LC-LRD)

We first briefly review a discriminative dictionary learning algorithm with low-rank regularization [16], in order to improve the performance even when large noise exists in the training samples. Moreover, locality constraint is added to take place of sparse coding to exploit the manifold structure of local features in a more thorough manner.

2.1. Discriminative Low-Rank DL

Given a training dataset $Y = [Y_1, Y_2, \ldots, Y_c]$, $Y \in \mathbb{R}^{d\times N}$, where c is the number of classes, d denotes the feature dimension, N is the total training samples' number, and $Y_i \in \mathbb{R}^{d\times n_i}$ is the samples from class i which has n_i samples. From Y, we want to learn a discriminative dictionary D and the coding coefficient X, which is utilized to future classification task. Then Y is equal to $DX + E$, with E as the noises. Different from using all the training samples to learn a whole dictionary, each sub-dictionary D_i for the i-th class is learned separately. Then X and D could be represented as $[X_1, X_2, \ldots, X_c]$ and $[D_1, D_2, \ldots, D_c]$, where D_i denotes the sub-dictionary for corresponding class, and X_i is the partial coefficients over D to represent Y_i.

Sub-dictionary D_i should be endowed with the discriminability of well represent samples from i-th class. Using mathematical formula, Y_i's coding coefficients X_i on D can be written as $[X_i^1; X_i^2; \ldots; X_i^c]$, in which X_i^j is Y_i's coefficient matrix on D_j. The discerning power of D_i comes from following two aspects: first, Y_i is expected to be well represented by D_i rather than by D_j, $j \neq i$. Therefore, it is reasonable to minimize $\|Y_i - D_i X_i^i - E_i\|_F^2$. At the meanwhile, D_i is not suppose to be good at representing other classes' samples, that is each X_j^i, where $j \neq i$ should have nearly zero value so that $\|D_i X_j^i\|_F^2$ is as small as possible. Thus we denote the discriminative fidelity term for sub-dictionary D_i as follows:

$$R(D_i, X_i) = \|Y_i - D_i X_i^i - E_i\|_F^2 + \sum_{j=1, j\neq i}^{c} \|D_i X_j^i\|_F^2. \quad (1)$$

In the task of dealing with face images, the within-class samples consist in a low dimensional manifold and are

linearly dependent. Therefore, sub-dictionaries should be properly trained as low-rank to represent samples from the same class. To this end, we want to find the one with the most concise atoms from all the possible sub-dictionary D_i, that is to minimize the rank of D_i. Recent researches suggest that the rank function can be replaced by the convex surrogate [4], that is $\|D_i\|_*$, where $\|.\|_*$ is the sum of singular values of the matrix, called nuclear norm.

2.2. Locality constraint

In this paper, we deploy locality constraint on the coefficient matrix instead of the sparsity constraint. As indicated by LCC [28], compared to sparsity, locality is more indispensable under certain assumptions. That is because locality constraint results in sparsity but not necessary vice versa. Specifically, the locality constraint uses following criteria:

$$\min_x \|l_i \odot x_i\|^2, \text{ s.t. } \mathbf{1}^T x_i = 1, \forall i, \tag{2}$$

where $l_i \in \mathbb{R}^k$ is the locality adapter, and $\odot$ represents dot product. According to each basis vector's similarity to the input sample y_i, l_i gives each one different weight. Specifically,

$$l_i = \exp(\frac{\text{dist}(y_i, D)}{\sigma}). \tag{3}$$

where $\text{dist}(y_i, D)$=$[\text{dist}(y_i, d_1), \ldots, \text{dist}(y_i, d_k)]^T$, and $\text{dist}(y_i, d_j)$ is the Euclidiean distance between sample y_i and each dictionary atom d_j. σ controls the bandwidth of the distribution.

2.3. Our proposed model

Considering the low-rank regularization term on the discriminative sub-dictionaies and the locality-constrained on the coding coefficients all together, we have the following LC-LRD model for each sub-dictionary:

$$\begin{aligned} &\min_{D_i, X_i, E_i} R(D_i, X_i) + \alpha\|D_i\|_* + \beta\|E_i\|_1 \\ &+\lambda \textstyle\sum_{k=1}^{n_i} \|l_{i,k} \odot x_{i,k}\|^2 \quad \text{s.t. } Y_i = DX_i + E_i \end{aligned} \tag{4}$$

Basically speaking, LC-LRD is based on the following three observations:

1. The discriminative term is introduced to give the discerning ability to each sub-dictionary,

2. Each sub-dictionary should be low-rank to separate noise from samples and discover the latent structure,

3. Inspired by [25] and the above discussions, locality is more essential than sparsity. That is similar samples should have similar representations.

2.4. Optimization of LC-DLRD

We consider dividing Eq.(4) into two sub-problems to solve the proposed objective function: First updating each coefficient $X_i (i = 1, 2, \ldots, c)$ one by one by fixing all other $X_j (j \neq i)$ and dictionary D then putting together to produce the coefficient matrix X; Second, updating D_i by fixing others. The locality-constrained coefficients X_i, the discriminative low-rank sub-dictionary D_i, and the sparse error E_i are obtained by iteratively operating this two steps.

Algorithm 2.4 Updating coefficients via ALM

Input: Training data Y_i, Initial dictionary D, Parameters λ, σ, β_1

Initialize:$Z = E_i = P = 0, \mu = 10^{-6}, \mu_{max} = 10^{30}$, $\epsilon = 10^{-8}, \rho = 1.1, maxiter = 10^6, iter = 0$

while not converges and $iter \leq maxiter$ **do**

1. Fix others and update Z by:
   $Z = Y_i - E_i + \frac{P}{\mu}$
2. Fix others and update X_i by:
   $X_i = \text{LLC}(Z, D, \lambda, \sigma)$[1]
2. Fix others and update E_i by:
   $E_i = \arg\min_{E_i} (\frac{\beta_1}{\mu}\|E_i\|_1 + \frac{1}{2}\|E_i - (Y_i - DX_i + \frac{P}{\mu})\|_F^2)$
3. Update multipliers P by:
   $P = P + \mu(Y_i - DX_i - E_i)$
4. Update μ by:
   $\mu = \min(\rho\mu, \mu_{max})$
5. Check if it is converged:
   $\|Y_i - DX_i - E_i\|_\infty < \epsilon$

end while

output: X_i, E_i

Assume that the discriminated dictionary D is given in the first sub-problem, the coefficients $X_i (i = 1, 2, \ldots, c)$ is updated one after another, then the original objective function Eq.(4) reduces to locality-constrained coding problem as follow:

$$\begin{aligned} &\min_{X_i, E_i} \lambda \textstyle\sum_{k=1}^{n_i} \|l_{i,k} \odot x_{i,k}\|^2 + \beta_1\|E_i\|_1 \\ &\text{s.t. } Y_i = DX_i + E_i \end{aligned} \tag{5}$$

which can be solved by the following ALM method [3].

[1]We set Z, D, λ and σ as the input of LLC [25] and the code can be downloaded from http://www.ifp.illinois.edu/ jyang29/LLC.htm.

$$\begin{aligned}&\min_{X_i,E_i} \lambda\textstyle\sum_{k=1}^{n_i}\|l_{i,k}\odot x_{i,k}\|^2+\beta_1\|E_i\|_1\\&+<P,(Y_i-DX_i-E_i)>+\tfrac{\mu}{2}\|Y_i-DX_i-E_i\|_F^2\\&=\min_{X_i,E_i} \lambda\textstyle\sum_{k=1}^{n_i}\|l_{i,k}\odot x_{i,k}\|^2+\beta_1\|E_i\|_1+\tfrac{\mu}{2}\|Z-DX_i\|_F^2\end{aligned} \tag{6}$$

where Z denotes to $Y_i - E_i + P/\mu$, μ is a positive penalty parameter, P denotes the Lagrange multiplier and $l_i = \exp(\text{dist}(z_i, D)/\sigma)$. Different from traditional locality-constrained linear coding (LLC) [25], we add an error term which could handle large noise in samples.

The detail of the coefficient updating can be referred to Algorithm (2.4). For the procedure of updating sub-dictionary, we have the same method with [16].

2.5. Classification based on our model

A linear classifier is used for final classification. In previous training process, the dictionary is learned, the locality-constrained coefficients X of training data Y and X_{test} of test data Y_{test} are calculated. The test sample i's representation x_i is X_{test}'s i-th column vector. The linear classifier $\hat{V}$ is obtained by a multivariate ridge regression model [30]:

$$\hat{V} = \arg\min_{V} \|L - VX\|_2^2 + \gamma\|V\|_2^2 \tag{7}$$

where L is the class label matrix for Y. This produces $\hat{V}$ = $LX^T(XX^T + \gamma I)^{-1}$. When testing points Y_{test} comes in, we first compute $\hat{V}X_{\text{test}}$. Then label for sample i is assigned by the position corresponding to the largest value in the label vector, that is: $label$ = $\arg\max_{label}(v = \hat{V}x_i)$.

3. Locality-constrained Collaborative Auto-Encoder (LCAE)

Suppose we have input image $x \in \mathbb{R}^D$, and hidden unit $z \in \mathbb{R}^d$ in which D is the visual descriptor's dimension. There are two important non-linear transformation in the auto-encoder's feed-forward process: "input→hidden units", and "hidden units→output" as:

$$z_i = \sigma(W_1 x_i + b_1); \quad h(x_i) = \sigma(W_2 z_i + b_2) \tag{8}$$

where $W_1 \in \mathbb{R}^{d\times D}$, $b_1 \in \mathbb{R}^d$, $W_2 \in \mathbb{R}^{D\times d}$, $b_2 \in \mathbb{R}^D$, and σ is the sigmoid function in the form of $\sigma(x) = (1+e^{-x})^{-1}$. Auto-encoder is basically a single hidden layer neural network, in which the input and target have same identity. Consequently, the output of auto-encoder is encouraged to be as similar to the target as possible. That is,

$$\min_{W_1,b_1,W_2,b_2} L(x) = \min_{W_1,b_1,W_2,b_2} \frac{1}{2n}\sum_i \|\hat{x}_i - h(x_i)\|_2^2, \tag{9}$$

where n is the number of samples, $\hat{x}$ is the target and $h(x_i)$ is the reconstructed input. By this means, the neurons in the hidden layer of auto-encoder are able to reconstruct the data and can be seen as a good representation for the input.

In order to introduce locality into the coding procedure, the input data is first reconstructed by LLC coding criteria then to work as the target of the auto-encoder. That is $\hat{x}$ in Eq. (9) is replaced by a locality reconstruction which followed as:

$$\begin{aligned}&\min_{C} \textstyle\sum_{i=1}^{N}\|x_i - Dc_i\|^2 + \lambda\|l_i \odot c_i\|^2\\&\text{s.t. } \mathbf{1}^T c_i = 1, \forall i\end{aligned} \tag{10}$$

where dictionary D will be initialized by PCA on the input training matrix X. The proposed LCAE can be trained using the backprop algorithm, which updates W and b by back propagation the reconstruction error gradient from the output layer to the locality coded target layer. After the iteration of forward and backward propagation, the locality coefficients will be updated using new output layer $h(x_i)$.

4. Experiments

We verify the performance of our LC-LRD and LCAE on various visual classification applications to demonstrate the efficiency and generality of the proposed methods. First, the LC-LRD is evaluated on four datasets including two face datasets: Extend YaleB [11], AR [17], one object categorization dataset COIL-100 [19], and one handwritten digits recognition dataset MNIST [10]. Second, Extend YaleB, AR, CMU PIE [24] and a newly built Virtual MakeUp (VMU) dataset [5] (samples shown in Fig. 3) are used to evaluate our LCAE method. Experimental results will be presented with some analysis in this section.

4.1. Experiments on LC-LRD

Several state-of-the-art algorithms were compared on each dataset, to show our advantage, including LDA [2], linear regression classification (LRC) [18] and several latest DL based classification methods, i.e. FDDL [27], DLRD [16], $D^2L^2R^2$ [12] and DPL [7]. In each experiment, we keep all the steps the same as that of the baselines except for the learning stage for fair comparison.

Parameter selection One of the most important parameters in majority of dictionary learning methods is the number of atoms in every sub-dictionary which denoted by m_i. In this paper's experiments, we set all the m_i equal, $i = 1, 2, \ldots, c$. We analyze the effect of m_i on the performance of LC-LRD, $D^2L^2R^2$, DLRD, FDDL and DPL. We take Extended YaleB as an example (20 training samples per class and the other setting is given in next subsection). Fig. 2 shows the accuracy of five methods versus different number of dictionary atoms. We can see that all methods have an increasing performance along with more atoms, and in

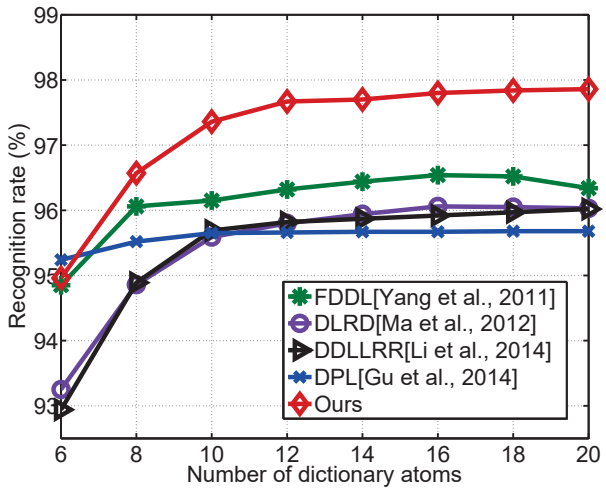

Figure 2. The recognition rates of five DL based methods versus the number of dictionary atoms with 20 training samples per class on Extend YaleB dataset.

all cases our LC-LRD method has nearly 2% improvement over other methods. Since each method's performance has a similar trend with the atoms' increasing, we fix the number of the dictionary columns of each class as training size for all of following experiments except for MNIST dataset, which is set to 30 each class. We will study the influence of neighbors' number k used for approximated LLC in experiments on LCAE, and in this section k is set equal to 10 as suggested in [25].

Figure 3. Sample images in the (a) Extended YaleB with 10%, 20%, 30% random pixel corruption; (b)AR dataset; (c) COIL-100 and (d) VMU datasets

There are five parameters in our approach: α, λ, σ in Eq. (4) and β_1, β_2 for error term of dictionary updating and coefficients updating separately. In the experiments, we found that β_1 and β_2 make more difference in recognition, therefore, other parameters α and λ are set as 1 in this paper. If no specific, the parameters β_1, β_2 and the parameters of other compared methods are chosen by 5-fold cross validation. For Extended YaleB, $\beta_1 = 15$, $\beta_2 = 100$; for AR, $\beta_1 = 5$, $\beta_2 = 100$; for COIL-100 , $\beta_1 = 3$, $\beta_2 = 150$; for MNIST, $\beta_1 = 2.5$, $\beta_2 = 2.5$.

The two face recognition datasets and splits subsets are downloaded from CAD website[2]. Through these datasets, the robustness of our algorithm to illumination changes, pose variations will be tested. Furthermore, we will evaluate LC-LRD's performance to noise by adding pixel corruptions.

Extended YaleB Dataset. The Extended YaleB dataset contains 2414 frontal-face images of 38 subjects captured under various lighting conditions. For each class, there are between 59 and 64 images for each person normalized to size 32×32. This dataset is diverse due to different illumination conditions, therefore we denote two experiments on this dataset. First, we choose random subsets with $p(= 5, 10, ..., 40)$ images per subject as the training set, and the rest of the dataset formed the testing set. There are 10 randomly splits for each given p; Second, a certain percentage of randomly selected pixels from the images are replaced by setting the pixel value as 255 (show in Fig. 3 (a)). Then randomly take 30 images as training samples, and the rest as testing samples and also repeat the experiment ten times. These two experiments results are given in Table. 1 and Table. 2 respectively.

Table. 1 shows the recognition rates with different training size. It can be observed that under all situations our method archives the best accuracy. Our method's robustness to noise is demonstrated in Table. 2, that along with the percentage of corruption increases our algorithm performs the best constantly. The performance of FDDL as well as DPL, LRC and LDA drops rapidly, by contrast, our method, $D^2L^2R^2$ and DLRD can still get much better recognition accuracies under different levels of corruption. This demonstrates the effectiveness of low-rank regularization and the error term when noise exists. Comparing with $D^2L^2R^2$ and DLRD, our method still performs better due to the locality constraint part, especially in cases that the occlusion is small.

AR Dataset. The AR dataset consists of more than 4,000 frontal-face images of 126 subjects, that is there are 26 pictures for each subject taken in two separated sessions. We follow the experimental setting in [27], for fair comparison, to choose 50 male subjects and 50 female subjects as a subset. For each subject, the 7 images from session 1 with illumination and expression changes were used for training, and the other 7 images from session 2 under the same condition served as testing. We do experiments on different features: original 60×43 images, resized 27×20 images and the feature provided by [9].

We illustrate the recognition rates under different feature in Fig. 4. From the figure, we can observe that our method achieves the best results on all the features and the improvement is larger compared with which on YaleB dataset. This could result in the variation of AR dataset and locality is proved to be better on dealing with this kind of data.

[2]http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

Table 1. Recognition rate(%) of different algorithms on Extended YaleB dataset with different number of training samples per class.

Training images	DPL [7]	$D^2L^2R^2$ [12]	DLRD [16]	FDDL [27]	LRC [18]	LDA [2]	Ours
5	75.17±1.86	75.96±1.20	76.17±1.16	77.75±1.34	60.24±2.02	74.12±1.52	**78.62±1.20**
10	89.31±0.62	89.60±0.89	89.94±0.89	91.16±0.85	82.98±0.82	86.67±0.90	**92.07±0.89**
20	95.69±0.90	96.02±0.91	96.03±0.85	96.15±0.66	91.80±0.97	90.64±1.07	**97.86±0.91**
30	97.80±0.36	97.87±0.42	97.90±0.47	97.86±0.35	94.60±0.60	86.84±0.92	**99.23±0.47**
40	98.67±0.43	98.09±0.39	98.80±0.37	98.84±0.46	96.10±0.58	95.27±0.79	**99.54±0.44**

Table 2. Recognition rate(%) of different algorithms on Extended YaleB dataset with various corruption percentage(%).

Occlusions	DPL [7]	$D^2L^2R^2$ [12]	DLRD [16]	FDDL [27]	LRC [18]	LDA [2]	Ours
0	97.80±0.36	97.87±0.42	97.90±0.47	97.86±0.35	94.60±0.60	86.84±0.92	**99.23±0.47**
5	78.27±1.22	91.90±1.14	91.84±1.07	63.55±0.87	80.49±1.10	29.03±0.82	**93.31±0.69**
10	64.58±1.09	85.71±1.51	85.82±1.54	44.65±1.22	67.61±1.33	18.53±1.15	**86.97±0.86**
15	53.77±0.86	80.46±1.64	80.89±1.37	32.76±1.03	56.81±1.24	13.63±0.53	**81.71±0.81**
20	44.95±1.38	73.59±1.54	73.56±1.63	25.26±0.42	47.23±1.59	11.30±0.46	**74.14±1.01**
25	35.87±1.01	65.93±1.15	65.88±1.50	18.45±0.82	38.85±1.18	9.23±0.81	**66.45±1.06**

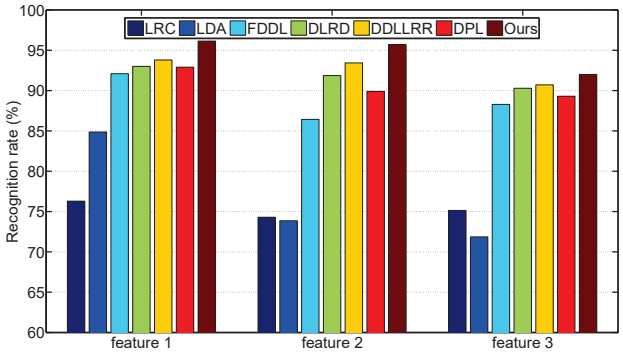

Figure 4. Average recognition rate(%) of different algorithms on AR dataset with three different features. Feature 1: row pixel 60×43; feature 2: row pixel 27×20; feature 3: feature provided by [9] .

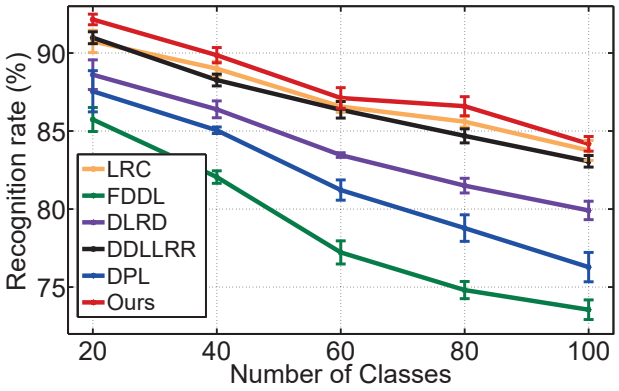

Figure 5. Recognition rate(%) with standard deviations of different algorithms on COIL-100 dataset.

COIL-100 Dataset. In this section, we assess our approach on object categorization by using the COIL-100 dataset. The training set is constructed by randomly selected 10 images per object, and the rest of the images consist the testing set. We repeat this random selection ten times, and the average results with standard deviations are reported. To evaluate the scalability of our method and competing methods, we separately utilize samples of 20, 40, 60, 80 and 100 objects in this dataset. Fig. 5 shows the average recognition rates with standard deviations of all compared methods. The results show our algorithm's generality that the locality not only works on face recognition but also on object categorization.

MNIST Dataset. We evaluate our algorithm on the subset of MNIST handwritten digit dataset downloaded from CAD website, which includes first 2000 training images and first 2000 test images with the size of each digit image is 28×28. This experimental setting follows [12], and we get consistency results. The recognition rates and traing/testing time by different algorithms on MNIST dataset are summarized in Table 3. Our algorithm achieves the highest accuracy than its competitors. Compared within the top 4 highest accuracy methods, ours' training time is the shortest because locality-constrained method only updates parts of dictionary atoms each time.

4.2. Experiments on LCAE

We report experimental results based on four datasets: three widely used face recognition datasets Extended YaleB, AR, CMU PIE and one newly built Virtual MakeUp (VMU) database. For all these datasets, we train both tradi-

Table 3. Average recognition rate(%) & running time(s) on MNIST dataset.

Methods	Accuracy	Training time	Testing time
$D^2L^2R^2$ [12]	84.23	233.429	84.863
DLRD [16]	86.15	243.271	99.787
FDDL [27]	84.85	263.137	388.219
LDA [2]	72.30	0.261	0.576
LRC [18]	82.70	365.605	–
DPL [7]	83.60	4.328	0.125
Ours	**88.75**	140.793	88.180

tional auto-encoder and LCAE separately, then use SVM as classier to provide the recognition rates. For Extended YaleB and AR datasets, we follow the setting in above section and specifically set training images in Extended YaleB as 10 per class. For CMU PIE, we also randomly choose 10 images per class as training and the rest images as testing. The VMU dataset is built to simulate the application of makeup by artificially adding makeup to 51 female Caucasian subjects (show in Fig. 3 (d)). There are 4 makeup statues (a) no makeup; (b) lipstick only; (c) eye makeup only; and (d) a full makeup including lipstick, foundation, blush and eye makeup. Hence, the assembled dataset contains total 204 images and four images per subject. We randomly select half as training and half as testing. The improvement of recognition rate on four dataset is shown in Table 4. We verify the effectiveness of the locality components of our approach by comparing it with baseline method in [20] which only differ in this aspects. We can see the LCAE algorithm gets higher recognition rate by introducing local information into the built auto-encoder, which enables it to provide similar inputs similar features.

The most important parameter in LCAE is k which used to determine how many local atoms of a dictionary are used to reconstruct the target in each iteration. As show in Fig. 6, the effect of k is explored on AR dataset. $k = 0$, means no locality reconstruct applied, is considered as baseline, and $k = 5, 25, 45, 65, 85$ are tested respectively. We can see the highest accuracy occurs when $k = 5$, and the accuracy decreases along with the increase value of k. When $k = 85$, means all the dictionary atoms are used, the accuracy falls back to the baseline, which is desirable since the local information will disappear with all the atoms used.

4.3. Discussion

From above experiments on our two proposed algorithms, we could find that locality constraint has the ability to improve both the dictionary learning method and auto-encoder by introducing local information. For LC-LRD, our method not only performs good on clear dataset but also gets better

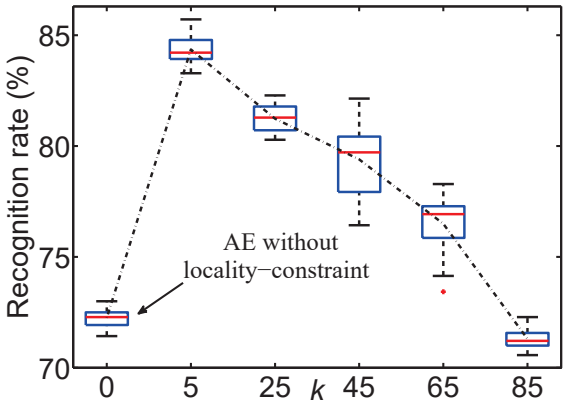

Figure 6. Accuracy on AR dataset with varying k. 0 means no locality applied, and $k = 85$ means all the dictionary atoms used.

Table 4. Average recognition rate(%) of AE and LCAE on different datasets.

Methods	YaleB	AR	PIE	VMU
AE	73.82	72.25	68.05	81.37
LCAE[Ours]	81.43	84.36	72.33	86.27

results on corrupted data. For the LCAE, we show its effectiveness on four face datasets and also explore its properties by varying k's value.

5. Conclusion

This paper investigated the efficiency of locality-constrained both on dictionary learning and auto-encoder. We first presented an algorithm which iterative learns a discriminative sub-dictionaries with low-rank regularization and locality constraint on coefficients. By applying locality constraint, we exploited the underlying manifold of data space and dictionary space in a more thorough manner than sparse representation. Second, we proposed a LCAE algorithm which introduce locality constraint to the target layer of auto-encoder. Extensive experiments have shown that our LC-LRD method outperforms the state-of-the-art methods on four benchmarks both in clean and corrupted cases and the LCAE also has the ability to give learned feature local information.

Acknowledgments

This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. -svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[2] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *TPAMI*, 19(7):711–720, 1997.

[3] D. P. Bertsekas. Constrained optimization and lagrange multiplier methods. *Computer Science and Applied Mathematics, Boston: Academic Press, 1982*, 1, 1982.

[4] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[5] C. Chen, A. Dantcheva, and A. Ross. Automatic facial makeup detection with application in face recognition. In *ICB*, pages 1–8. IEEE, 2013.

[6] Y. Chen and X. Guo. Learning non-negative locality-constrained linear coding for human action recognition. In *VCIP*, pages 1–6. IEEE, 2013.

[7] S. Gu, L. Zhang, W. Zuo, and X. Feng. Projective dictionary pair learning for pattern classification. In *NIPS*, pages 793–801, 2014.

[8] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *CVPR*, pages 1697–1704. IEEE, 2011.

[9] Z. Jiang, Z. Lin, and L. S. Davis. Label consistent k-svd: learning a discriminative dictionary for recognition. *TPAMI*, 35(11):2651–2664, 2013.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[11] K.-C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *TPAMI*, 27(5):684–698, 2005.

[12] L. Li, S. Li, and Y. Fu. Learning low-rank and discriminative dictionary for image classification. *Image and Vision Computing*, 2014.

[13] Y. Liang, M. Song, J. Bu, and C. Chen. Colorization for gray scale facial image by locality-constrained linear coding. *Journal of Signal Processing Systems*, 74(1):59–67, 2014.

[14] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *TPAMI*, 35(1):171–184, 2013.

[15] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[16] L. Ma, C. Wang, B. Xiao, and W. Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In *CVPR*, pages 2586–2593. IEEE, 2012.

[17] A. M. Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

[18] I. Naseem, R. Togneri, and M. Bennamoun. Linear regression for face recognition. *TPAMI*, 32(11):2106–2112, 2010.

[19] S. Nayar, S. A. Nene, and H. Murase. Columbia object image library (coil 100). *Department of Comp. Science, Columbia University, Tech. Rep. CUCS-006-96*, 1996.

[20] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. V. Le, and A. Y. Ng. On optimization methods for deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 265–272, 2011.

[21] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. Action classification with locality-constrained linear coding. In *ICPR*, pages 3511–3516. IEEE, 2014.

[22] A. Shabou and H. LeBorgne. Locality-constrained and spatially regularized coding for scene categorization. In *CVPR*, pages 3618–3625. IEEE, 2012.

[23] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, pages 853–860. IEEE, 2012.

[24] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *TPAMI*, 25(12):1615–1618, 2003.

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

[26] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *TPAMI*, 31(2):210–227, 2009.

[27] M. Yang, D. Zhang, and X. Feng. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, pages 543–550. IEEE, 2011.

[28] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *NIPS*, pages 2223–2231, 2009.

[29] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma. Image classification by non-negative sparse coding, low-rank and sparse decomposition. In *CVPR*, pages 1673–1680. IEEE, 2011.

[30] G. Zhang, Z. Jiang, and L. S. Davis. Online semi-supervised discriminative dictionary learning for sparse representation. In *ACCV*, pages 259–273. Springer, 2013.

[31] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *CVPR*, pages 2691–2698. IEEE, 2010.