Locality-constrained Low Rank Coding for Face Recognition

Devansh Arpit†  Gaurav Srivastava†  Yun Fu§ ∗
†SUNY at Buffalo, USA  §Northeastern University, Boston
E-mail: {devansha, gsrivast}@buffalo.edu

Abstract

This paper presents Locality-constrained Low Rank Coding (LLRC) as a novel approach for image classification. The widely used Sparse representation based algorithms reconstruct a test sample using a sparse linear combination of training samples. But they do not consider the underlying structure of the data in the feature space. On the other hand, Low Rank representation has been recently used for clustering face images into their respective classes by taking advantage of the low rank structure of the data. LLRC first imposes a locality constraint to choose the training samples that are in the vicinity of the test sample. Then it applies the low rank constraint on these training samples to further choose a subset that belongs to a subspace corresponding to one face class. In this manner, the training samples used to reconstruct a given test sample can be chosen from just one class rather than a mixture of classes, thus enhancing the classification accuracy. We evaluate our algorithm on face image datasets. Our algorithm outperforms sparse representation based algorithms, thus showing that exploiting the structure of data is important. We further demonstrate that both locality constraint and low rank constraint are imperative to obtain superior performance.

1. Introduction

The accuracy of an algorithm used for the task of pattern recognition depends significantly on the underlying structure of the data at hand, i.e., how the data is distributed in the feature space. Thus an algorithm that exploits this information about the data can achieve better accuracy than an algorithm that does not. For applications such as face recognition, it is generally assumed that the face images with varying appearances, illuminations, or expressions lie along linear subspaces [6] and hence can be linearly represented using a weighted sum of known face samples. This linearity assumption leads to easy computation and often produces good results in real applications. Therefore it has been used by recent algorithms, e.g. Sparse Representation for Classification (SRC) [8], Locality-constrained Linear Coding (LLC)[7] and Low Rank Representation (LRR) [3].

Sparse representation based approaches [8, 7, 9] rely on the assumption that representing an unknown sample using a sparse linear combination of the training samples leads to better separability between the classes. On the other hand, lately the idea of locality has been found to give more promising results [10, 7] for the task of classification. It emphasizes that reconstructing a given sample using training data located close to it in the feature space leads to more reliable classification.

Although the above approaches consider linear subspaces, they do not explicitly model an important intuition that the samples from different classes may lie in independent linear subspaces. Recently, low-rank representation has been used for unsupervised subspace segmentation and clustering of the face images into different classes [3]. In this work, the authors assume that face images from different subjects belong to independent linear subspaces. Thus the authors exploit the structure of the data for the task of finding clusters in the collection of images.

In this paper, we propose a novel approach called Locality-constrained Low Rank Coding (LLRC) to address the issue inherent to LLC that all training samples
close to a given test sample may not necessarily belong to the same class and hence may degrade the classification performance when used for reconstructing the test sample. By imposing a low rank constraint on the training samples that lie in the vicinity of the test sample, we can segment them into nearly independent linear subspaces. Then the test sample can be reconstructed using the samples belonging to one of the subspaces that lie closest to the test sample. Since LRR assumes that the different subspaces correspond to different face classes, the test sample will be reconstructed from training samples belonging to just one class rather than a mixture of classes, and this will boost the classification performance when used for reconstructing the test sample.

Different subspaces correspond to different face classes, the test sample is reconstructed using samples that lie in the vicinity of the test sample, we can segment them into nearly independent linear subspaces. Then the test sample can be reconstructed using very few samples from the dictionary. Let \( D \) be a low rank estimate of \( X \) using the columns of the dictionary \( D \). For subspace segmentation problem, the authors put \( D = X \). Thus a low rank estimate of \( D \) will enforce a low rank approximation of the matrix \( X \) while minimizing the error between \( X \) and its low rank approximation.

2. Related Work
2.1. Low Rank Representation

The task of subspace segmentation essentially means segmenting a collection of data such that the data vectors in each segmented subset belong to an independent low dimensional space and that ideally the distribution of data of each particular segment is degenerate along the other dimensions not described by that subspace. More specifically, in [3], Liu et al. proposed that the task of clustering face images into their respective classes can be converted into a subspace segmentation problem if we assume that face images from different individuals lie near independent subspaces.

Consider a matrix \( X = [x_1, x_2, x_3, \ldots, x_n] \in \mathbb{R}^{m \times n} \) where the columns represent the feature vectors. Then LRR tries to segment these \( n \) samples into \( k \) classes by minimizing the following objective function:

\[
\text{arg min}_{Z} \|Z\|_* \quad \text{s.t.} \quad X = DZ,
\]

where \( \| \cdot \|_* \) denotes the nuclear norm, \( Z = [z_1, z_2, z_3, \ldots, z_n] \) is the coefficient matrix such that each \( z_i \) provides us with a representation of the column vector \( x_i \) using the columns of the dictionary \( D \). For subspace segmentation problem, the authors put \( D = X \). Thus a low rank estimate of \( Z \) will enforce a low rank approximation of the matrix \( X \) while minimizing the error between \( X \) and its low rank approximation.

2.2. Sparse Representation based Classification

The seminal work by Wright et al. [8] introduced the idea of sparse coding for face recognition. The basic idea was to represent a new sample \( y \) using a very sparse linear combination of the training samples. Let \( D = [d_1, d_2, d_3, \ldots, d_k] \in \mathbb{R}^{m \times n} \) denote the dictionary matrix containing a total training samples of \( n \) face images from \( k \) classes where each \( d_i \) is a sub-matrix containing the images from the \( i \)-th class as its columns. Further, let \( w \in \mathbb{R}^n \) denote the weights over these \( n \) vectors. Then the objective of SRC is to minimize the following:

\[
\text{min}_{w} \|w\|_1 \quad \text{s.t.} \quad y = Dw,
\]

Since \( \ell_1 \) norm is a sparsity inducing norm, solving the above objective function yields a vector \( w \) which has very few non-zero elements leading to signal reconstruction using very few samples from the dictionary.

Classification is then done by computing the reconstruction error from each class individually and then assigning the sample to the class with the least error.

2.3. Locality-constrained Linear Coding

The motivation of LLC [7] was based on overcoming the limitation of SRC. LLC lays stress on the locality of the code rather than sparsity directly. The objective function of LLC can be expressed as:

\[
\text{min}_{w} \|(y - Dw)\|_2^2 + \lambda \|l \odot w\|_2^2
\]

where \( \odot \) denotes element wise product, \( D \) is the dictionary, \( w \) is the optimal code for signal reconstruction and \( l \) is a vector such that the elements of \( l \) indicate the distance of the signal \( y \) from every training sample (column vectors of \( D \)). Thus LLC is able to achieve a locally smooth sparse code vector where the sparsity is a result of the locality constraint since the training samples far away from \( y \) do not contribute to the reconstruction of \( y \). In this way, a signal with slight variations gets represented using a very similar code vector which is not guaranteed by sparse coding.

**Algorithm 1 ADMM algorithm for solving LLRC**

**INPUT:** \( y, D, \lambda_1, \lambda_2 \)

**INITIALIZE:** \( w = 0, \Lambda = 0, \mu = 10^{-6}, \rho = 3, \tau = 10^{10}, \epsilon = 10^{-8} \)

**while** not converged **do**

1. **fix others and update** \( Z \) by

\[
\min_{Z} \frac{\lambda_2}{\mu} \|Z\|_* + \frac{1}{2} \|Z - (D \text{diag}(w) + \Lambda)\|_F^2.
\]

2. **fix others and update** \( w \) by

\[
\begin{align*}
2D^T D & + 2\lambda_2 \text{diag}(l) \odot \text{diag}(l) \\
p_1 &= \mu (D \odot D)^T 1 \\
p_2 &= 2D^T y + \mu (Z \odot D)^T 1 - (\Lambda \odot D)^T 1
\end{align*}
\]

3. **Update Lagrange and penalty parameters**

\[
\Lambda \leftarrow \Lambda + \mu (D \text{diag}(w) - Z)
\]

\[
\mu \leftarrow \min(\rho \mu, u)
\]

4. **Check for convergence**

\[
\|D \text{diag}(w) - Z\|_\infty < \epsilon
\]

**end while**

**OUTPUT:** \( w \)
3. Locality-constrained Low Rank Coding

3.1. Algorithm

We propose LLRC as a classification algorithm for data types that have a significantly low rank structure in the feature space. By low rank structure we mean that the data types that have a significantly low rank structure in feature space. By low rank structure we mean that the different classes in the given data lie approximately near linear independent subspaces. We use the word 'approximately' because real world data rarely lies near completely independent linear subspaces. This problem is dealt using a locality constraint. The idea is to find a weight vector \( w \) (that leads to a low reconstruction error) the elements of which when multiplied with the corresponding column of the dictionary matrix \( D \) results in a low rank matrix \( Dw \), hence the name Low Rank Coding. Mathematically, we propose to minimize the following objective function:

\[
\min \|y - Dw\|^2 + \lambda_1 \|Ddiag(w)\|_* + \lambda_2 \|l \odot w\|^2 \tag{4}
\]

where \( D = [d_1, d_2, d_3, \ldots, d_n] \in \mathbb{R}^{m \times n} \) is the dictionary containing the training samples as its column vectors, \( y \in \mathbb{R}^m \) is the given signal, \( w \in \mathbb{R}^n \) is the low rank code that linearly weights the training samples for signal reconstruction and finally \( l \in \mathbb{R}^n \) is a vector in which each element represents the exponential distance of each training sample \( d_i \) in the dictionary from the vector \( y \) in the same order. Thus the \( i \)-th element of vector \( l \) is given by:

\[
l_i = \exp \left( \frac{\|y - d_i\|^2}{\sigma} \right) \tag{5}
\]

For robustness, we also normalize the values of \( l \) to range between 0 and 1. The nuclear norm in (4) has been used as a convex surrogate of the rank operator. The matrix \( Ddiag(w) \) explicitly represents the vectors that are used to reconstruct the signal \( y \). Hence minimizing the rank of this matrix directly implies reconstructing \( y \) using only those training samples that belong to a low dimensional subspace. The choice of \( \lambda_1 \) and \( \lambda_2 \) on the structure of the data in feature space. The ratio \( \lambda_1/\lambda_2 \) is larger if different classes belong to separable linear subspaces and smaller otherwise.

Finally, the test sample is recognized as belonging to a class \( k \) which satisfies the following criteria:

\[
k = \arg \min_k \|y - D_k w_k\|^2 \tag{6}
\]

where \( D_k \) is that block of matrix \( D \) which contains samples from only class \( k \) as its column vector and \( w_k \) is the vector which contains element from vector \( w \) that corresponds to the column vectors in \( D_k \).

### Table 1: Recognition rates on AR database

<table>
<thead>
<tr>
<th>Dim</th>
<th>30</th>
<th>150</th>
<th>300</th>
<th>540</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN</td>
<td>60.9</td>
<td>69.4</td>
<td>70.6</td>
<td>71.0</td>
</tr>
<tr>
<td>LLC</td>
<td>73.0</td>
<td>87.0</td>
<td>90.9</td>
<td>91.1</td>
</tr>
<tr>
<td>SRC</td>
<td>71.1</td>
<td>90.1</td>
<td>91.1</td>
<td>91.7</td>
</tr>
<tr>
<td>LLRC [ours]</td>
<td>74.0</td>
<td>89.4</td>
<td>92.3</td>
<td>93.0</td>
</tr>
</tbody>
</table>

3.2. Optimization Procedure

We solve the optimization problem using alternating direction method of multipliers (ADMM) by converting the problem (4) to the following equivalent formulation:

\[
\min_{w,Z} \|y - Dw\|^2 + \lambda_1 \|Z\|_* + \lambda_2 \|l \odot w\|^2 \\
\text{s.t. } Z = Ddiag(w) \tag{7}
\]

The above problem can now be solved using the following augmented lagrangian formulation [2]:

\[
\min_{w,Z} \|y - Dw\|^2 + \lambda_1 \|Z\|_* + \lambda_2 \|l \odot w\|^2 + \\
\text{tr}(\Lambda^T(Ddiag(w) - Z)) + \frac{\mu}{2} \|Ddiag(w) - Z\|_F^2 \tag{8}
\]

where \( \text{tr}(\cdot) \) is the trace operator, \( \|\cdot\|_F \) is the Frobenius norm, \( \Lambda \) is the Lagrange multiplier and \( \mu \) is the penalty parameter. Algorithm 1 shows the step by step procedure to solve the optimization problem (8). Step 1 in the algorithm can be solved using the singular value thresholding operator as given in [1]. The closed form of \( w \) in step 2 can be obtained by simple algebraic calculations.

4. Experimental Results

We evaluate our algorithm on two face databases - AR database [4] and ORL database [5]. We use Principal Component Analysis (PCA) [6] for dimensionality reduction and perform experiments for different feature dimensions. We compare our results with SRC [8], LLC [7] and k-Nearest Neighbor (kNN).

1) AR database: The AR database consists of more than 4000 images of 126 individuals with 26 images per individual taken in 2 sessions. We perform our evaluations on a subset of this dataset consisting of 50 male and 50 female subjects. For each subject, we choose 7 images each from session 1 and session 2 with only illumination and expression changes for training and testing respectively. The images are resized to 32 \( \times \) 32, converted to grayscale, normalized and the pixels are concatenated to form a vector. We perform experiments with feature dimensions 30, 150, 300 and 540.

We vary the parameters for every algorithm we compare with and show their best results for fair comparison. The typical values of \( \lambda_1 \) and \( \lambda_2 \) are around 0.01.
and 10 respectively for LLRC. As can be seen in table 1, LLRC consistently outperforms the other algorithms except at feature dimensions 150 at which SRC outperforms LLRC by 0.7%. The highest accuracy reached by LLRC is 93% with 504 dimensional features as against 91.7% of SRC and 91.1% of LLC.

We perform another experiment using this database in which we fix the test dataset as previously and vary the number of training samples from 1 to 7. The performance of LLRC (figure 2) is consistently better than the other algorithms compared with.

In Figure 3, each row represents the training samples from the liner subspace found by LLRC to represent the test sample. The results show that most members of the found subspace correctly belong to the same class as the test sample. However, the fourth row shows an interesting result that LLRC found a subspace of laughing female faces to represent the test sample.

Figure 4 shows that both the locality and the low rank constraints used by LLRC are essential for good performance as compared to either of them applied alone. Locality constraint only guarantees the proximity of reconstructing training samples to the test sample. But a subsequent low rank constraint leads to pruning of these training samples to a subset that belongs to a single subspace corresponding to one class.

2) ORL database: The ORL database consists of 400 images of 40 individuals such that there are 10 images per person with varying pose and illumination. The images are resized to 32 × 32, converted to grayscale, normalized and the pixels are concatenated to form a vector. We randomly split the database into two halves and for each subject, we keep 5 samples for training and the other 5 for evaluation. We perform experiments with feature space of dimensions 20, 80, 140 and 300.

The typical values of \( \lambda_1 \) and \( \lambda_2 \) for LLRC were chosen to be around 0.5 and 100. As can be seen from table 2, LLRC outperforms the other algorithms at all values of feature space dimensions and achieves a highest accuracy rate of 96.5%.

5. Conclusion and Future Work

We presented a novel low rank approach for face recognition and achieved higher accuracy than existing algorithms. It is worth mentioning that neither the locality constraint nor the low rank scheme alone gives as good a result as compared to when both the constraints are applied together. This is because face images of different subjects do not lie in completely independent linear subspaces and a locality constraint helps in choosing the correct subspace for correct classification. The future work includes extending this idea to deal with the cases of corrupt and occluded images.

References


