

Sparse Representation Classification with Manifold Constraints Transfer

Baochang Zhang^{*,†}

Alessandro Perina^{*}

Vittorio Murino^{*}

Alessio Del Bue^{*}

^{*} Istituto Italiano di Tecnologia (IIT)
Pattern Analysis and Computer Vision (PAVIS)
Via Morego 30, 16136 Genova, Italy

[†] School of Automation Science and Electrical Engineering
Beihang University, Beijing, China

Abstract

The fact that image data samples lie on a manifold has been successfully exploited in many learning and inference problems. In this paper we leverage the specific structure of data in order to improve recognition accuracies in general recognition tasks. In particular we propose a novel framework that allows to embed manifold priors into sparse representation-based classification (SRC) approaches. We also show that manifold constraints can be transferred from the data to the optimized variables if these are linearly correlated. Using this new insight, we define an efficient alternating direction method of multipliers (ADMM) that can consistently integrate the manifold constraints during the optimization process. This is based on the property that we can recast the problem as the projection over the manifold via a linear embedding method based on the Geodesic distance. The proposed approach is successfully applied on face, digit, action and objects recognition showing a consistently increase on performance when compared to the state of the art.

1. Introduction

A constrained learning model allows one to incorporate domain-specific knowledge as constraints to balance the learned model given the implicit structure of the data [7]. From a machine learning perspective, it is of high significance to simplify the learning stage while improving the quality of the solutions. Imposing a data structure as a constraint is a new and flexible way to solve the optimization problems [5], and it has shown a lot of promise in improving performance of machine learning algorithms.

In this paper we revisit the Sparse Representation Classification (SRC) [29] framework by including the notion that data often lies on a specific manifold. The SRC approach

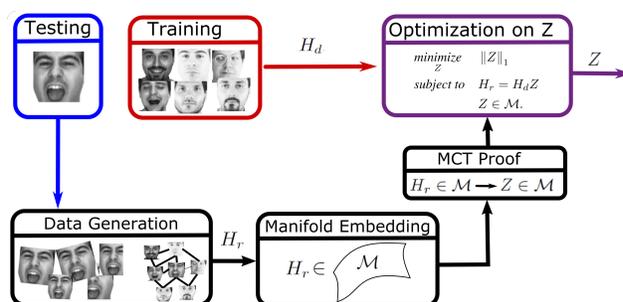


Figure 1. A scheme of the MCT approach.

instantiates recognition as a sparse optimization problem where the aim is to estimate an assignment matrix Z that associates each testing sample H_r to the corresponding training data H_d such that:

$$H_r = H_d Z.$$

This inference is in general done through the use of a dictionary computed directly from the training data. If such dictionary has certain properties, namely being uncorrelated, the theory of SRC has interesting links to the compressive sensing framework. This makes the problem solvable through a l_1 relaxation that can be dealt with efficient optimization techniques.

This paper provides new insights to the problem, in particular a relevant intuition that was neglected in previous work. It is well known in fact that often data lies on specific manifolds, especially when one has to identify a well-defined object category (e.g. faces, digits, animals, flowers, etc.). To this end, any recognition framework taking advantage of the implicit structure of the data can obtain improved results. However, the main crux is to efficiently embed manifold constraint in any optimization methods and, in the specific, in the SRC framework. Here we show that there exists a solution with high practicability that can include manifold

constraints in an alternating direction method of multipliers (ADMM).

Figure 1 shows a scheme of our approach. Given a testing sample and the training data, we can generate an augmentation of the data (data generation) by simulating geometrical transformations, e.g. rotation and translation, or by choosing k nearest neighbours from the training data. Then the data is embedded to a manifold using a Linear Local Embedding (LLE) approach. We then observe that, if testing and training data are linearly correlated through the assignment matrix Z , it is possible to transfer the manifold constraints to the latter matrix (MCT proof). This simplifies the inference mechanism of the algorithm and it provides an efficient implementation through ADMM (optimization on Z).

The rest of the paper is organized as follows. Sec. 2 presents the background on SRC methods and this paper contributions with respect to the state of the art. Sec. 3 details how manifold constraints can be efficiently embedded in a ADMM optimization framework while Sec. 4 shows the theoretical insights of the approach. Finally experiments over several datasets are discussed in Sec. 5 while Sec. 6 draws some conclusions.

2. Previous work and contributions

The SRC method was originally proposed for face recognition problems showing remarkable robustness to data corruptions such as occlusions and strong photometric variations [29]. If the signals in each class lie in a low-dimensional subspace and the different classes subspaces satisfy certain incoherence conditions, it is shown in [29] that all the non-zero coefficients in sparse coding will be associated with the dictionary atoms that belong to the same class as the test sample. Many variants of SRC have been proposed to increase performance. For example, the work in [26] presents a simply but effective coding scheme called Locality-constrained Linear Coding (LLC) in place of the coding scheme of traditional sparse representation. In [25], Wagner et al. proposed a sparse representation based method that could deal with face misalignment and illumination variation. In [17], they aim to minimize the distribution divergence between the labeled and unlabeled images, and incorporate this criterion into the objective function of sparse coding to make the new representations robust to the distribution differences. Another variation of SRC [19] has improved the fidelity term on the recovery process.

A key feature of many learning approaches is indeed the careful design of the dictionary [6]. Originally, a SRC dictionary was constructed by directly including all the training samples [29]. Traditional dictionary learning methods have been specialized for sparse representation such as the Method of Optimal Direction (MOD) [13] and K-SVD [22]. In order to promote the discriminative power of dictionary,

recent works have augmented the reconstruction objective function with additional discrimination terms; e.g., maximum margin [28]. The latest classification scheme was proposed as a Fisher discrimination dictionary learning (FDDL) model [30] and it exploits the discriminative information in both the representation residual and the representation coefficients.

Given this review of SRC based method, to the best of our knowledge, this is the first attempt to perform sparse coding on data embedded into a manifold. The proposed method has two significant contributions:

- If data is distributed on a manifold, we obtain a framework based on alternating direction method of multipliers (ADMM) for several recognition problems.
- Given a linear (matrix) correlation between the optimized variables and the data from a manifold, we present a novel and general theory to transfer the manifold constraints from the data to the variables.

3. Manifold constraints in multipliers optimization methods

In this section, we present how manifold constraints can be used into a sparse representation framework for classification. First we show how to formulate the optimization problem in terms of an Augmented Lagrange Multipliers (ALM) minimization. Then a relaxed and more efficient solution can be obtained using the method of ADMM (Alternating Direction Method of Multipliers) that will lead finally to our solution using manifold constraints transfer (MCT).

3.1. SRC reformulation using manifold constraints

SRC defines the problem of recognition as the optimization of a matrix Z of size $N \times F$ which describe the input data matrix H_r of size $M \times F$ given a dictionary H_d of size $M \times N$ where N and F are respectively the training and test samples while M is the feature dimensionality. This can be formalized as:

$$\begin{aligned} & \underset{Z}{\text{minimize}} && \|Z\|_1 \\ & \text{subject to} && H_r = H_d Z, \end{aligned} \tag{P1}$$

where $\|Z\|_1$ corresponds to the sum of the l_1 -norm of each vector of Z , i.e., if $Z = [\mathbf{z}_1, \dots, \mathbf{z}_F]$, the value of $\|Z\|_1$ is given by $\sum_{i=1}^F \|\mathbf{z}_i\|_1$. In the SRC framework, the matrix Z represents an assignment of each data samples to the respective elements in H_d . The solution of (P1) allows to solve for the identity of the test sample given the feature descriptors, as stored in H_d . Data samples in H_r and H_d are represented by feature descriptors of length M , which

are custom for each application. Moreover, given the specificity of the data, H_r contains samples that are lying on a specific and unknown manifold \mathcal{M} , i.e. $H_r \in \mathcal{M}$.

In this paper, to solve problem (P1), we exploit the fact that input data lies on a manifold. More specifically, we propose to consider H_r as an unknown variable of the optimization by performing variable cloning i.e., $H_r \rightarrow \hat{H}_r \in \mathcal{M}$ and to explicitly enforce manifold constraints over the cloned variables \hat{H}_r . This aims to introduce explicitly the manifold constraints at the expenses of replicating a set of variables.

To reformulate the problem, we first rewrite the traditional sparse representation as a new optimization problem giving

$$\mathcal{F}(Z, H_r) = \|Z\|_1 + \lambda \cdot (H_r - H_d Z),$$

with λ as a Lagrange factor. It is important to note that the objective function $\mathcal{F}(Z, H_r)$ can be customized for different problems thus assuring the generality of the approach. Now, the original problem can be rewritten as:

$$\begin{aligned} & \underset{Z, \hat{H}_r}{\text{minimize}} && \mathcal{F}(Z, \hat{H}_r) \\ & \text{subject to} && H_r = \hat{H}_r \\ & && \hat{H}_r \in \mathcal{M}, \end{aligned} \quad (\text{P2})$$

Differently from traditional sparse optimisation, problem (P2) is not easily solvable because of the additional manifold constraint. To this end, we introduce here a generic method based on the Augmented Lagrange Multipliers (ALM). More formally we define:

$$\begin{aligned} L_\sigma(Z, H_r, \hat{H}_r; R) = & \\ \mathcal{F}(Z, \hat{H}_r) - \sum_{i=1}^F \mathbf{r}_i^\top (\mathbf{h}_{r,i} - \hat{\mathbf{h}}_{r,i}) + \frac{\sigma}{2} \sum_{i=1}^F \|\mathbf{h}_{r,i} - \hat{\mathbf{h}}_{r,i}\|^2 & \end{aligned} \quad (1)$$

The scalar $\sigma > 0$ is the weight of the penalty term, $R = [\mathbf{r}_1 \dots \mathbf{r}_N]$ denotes the Lagrange multipliers and the sub-matrix \hat{H}_r is given by

$$\hat{H}_r = [\hat{\mathbf{h}}_{r,1}, \dots, \hat{\mathbf{h}}_{r,i}, \dots, \hat{\mathbf{h}}_{r,F}].$$

The optimization algorithm for the function in Eq. (1) is presented in Alg. 1 where we show the iterative steps of the method.

Eq. (2) in Alg. 1 requires a minimization over $\hat{\mathbf{h}}_{r,i} \in \mathcal{M}$ with $i = 1, \dots, F$. In [10, 33] the manifold constraints are enforced in an ALM strategy by using a matrix projection which efficiently computes the solution over several given manifold (e.g. Stiefel, unit sphere). Differently, in our case, data is embedded into a manifold, not known a priori (i.e.

Algorithm 1 - ALM solution with manifold constraints

- 1: Set $l = 0$ and $\epsilon_{\text{best}} = +\infty$
- 2: Initialize $\sigma^{[0]}$, $R^{[0]}$, $\gamma > 1$ and $0 < \eta < 1$
- 3: Initialize $Z^{(0)}$ and $H_r^{(0)}$
- 4: **repeat**
- 5:

$$\begin{aligned} & (\tilde{Z}^{[l+1]}, \tilde{H}_r^{[l+1]}) = \\ & \underset{\hat{\mathbf{h}}_{r,i} \in \mathcal{M}, i=1, \dots, F}{\text{argmin}} && L_{\sigma^{[l]}}(Z, H_r, \hat{H}_r; R^{[l]}) \end{aligned} \quad (2)$$

- 6: $\epsilon = \left\| H_r^{[l+1]} - \hat{H}_r^{[l+1]} \right\|^2$
 - 7: **if** $\epsilon < \eta \epsilon_{\text{best}}$
 - 8: $R^{[l+1]} = R^{[l]} - \sigma^{[l]} \left(H_r^{[l+1]} - \hat{H}_r^{[l+1]} \right)$
 - 9: $\sigma^{[l+1]} = \sigma^{[l]}$
 - 10: $\epsilon_{\text{best}} = \epsilon$
 - 10: **else**
 - 10: $R^{[l+1]} = R^{[l]}$
 - 11: $\sigma^{[l+1]} = \gamma \cdot \sigma^{[l]}$
 - 12: **endif**
 - 13: $l \leftarrow l + 1$
 - 14: **until** Some stopping criterion
-

H_r), and thus it is customized to the specific recognition problem at hand.

Now we formalise the manifold following the notation of [4, 9] by introducing a neighbour-preserving embedding which we use to find an estimation on a sub-manifold (a subset of a manifold). Such formalisation is similar to [9, 8] that first calculates the weights in the process of dimension reduction by LLE [23]. In particular, the embedding generated by [9, 8] is exactly based on a sub-manifold given by a small set of samples. Therefore, instead of finding a whole manifold (LLE) to the data, we alternatively construct a sub-manifold (check Sec. 5 for details). We just consider the Geodesic distance information as in [8] contained in a sub-manifold to find a projection based on its ‘‘true’’ neighbours so to avoid perturbations from samples far from the input data.

3.2. Neighbour-preserving embedding based on the Geodesic distance

Let \mathcal{M} be the sample set representing a manifold and let m' be the embedding of \mathcal{M} via a mapping function $\Phi(\cdot)$.

Definition 1. *The map function $\Phi : m' \rightarrow \mathcal{M}$ in the neighbour-preserving embedding method based on the Geodesic distance is defined as follows:*

1. $\Phi(m', \mathcal{M}) = \sum_{j=1}^K (1 - W_j) \mathcal{M}_j$ where W_j is the Geodesic distance of the sample m' and the j^{th} sample

in a sub-manifold set.

2. As shown in [8], the map Φ is a linear embedding based on the Geodesic distance. For a given point onto the sub-manifold, the bigger weights are reasonably set to its nearest points in the recovery process as defined in the previous point of the definition.

From Definition 1 the input sample can be projected onto a well-designed sub-manifold via an embedding function by fully exploiting the neighbour structure information [9]. As shown in LLE [23, 27], a local point on a sub-manifold can be represented by a small and compact set of nearest neighbours (i.e. K) to approximate ISOMAP. Later in [4], it has been shown that Geodesic distance used in ISOMAP is another effective way to locate the neighbours for a linear embedding. Our idea has similarities to [9] with the main difference being its simplicity, suitable for recognition problems. The manifold structure later is, for the first time, exploited to find a reasonable constraint for the original sparse representation problem.

This is the only part of the algorithm where the constraint manifold \widehat{H}_r plays a role and replacing \widehat{H}_r amounts to compute the proposed manifold embedding. Finally, ALM adds an extra computational cost to (P2) because variables are added given the cloning mechanism. Based on the Definition 1, ADMM can be alternatively used to solve our problem and in the following we give details about the formalization of the this optimization procedure.

3.3. Alternating Direction Method of Multipliers (solving for Eq. 2)

The ADMM algorithm allows to solve the optimization problem of Eq. (2) iteratively for one of the variables while keeping the others fixed. More in detail, we solve first over $Z^{[l+1]}$ given $(\widehat{H}_r^{[l]}, Z^{[l]})$ and then for $\widehat{H}_r^{[l+1]}$ given $Z^{[l+1]}$.

For (P1), \widehat{H}_r is solved following Def. 1, giving

$$\widehat{H}_r^{[l]} = \Phi(H_d Z^{[l]}, \mathcal{M}) = \sum_{j=1}^K (1 - W_j) \mathcal{M}_j, \quad (3)$$

where \mathcal{M} is specific given the test data. Then $Z^{[l+1]}$ is further solved by letting $Z = Z^{[l+1]}$, $\widehat{H}_r = \widehat{H}_r^{[l]}$, $\sigma = \sigma^{[l]}$ and $R = R^{[l]}$. Solving Eq. (2) in Algorithm 1 corresponds to solve for

$$L_\sigma(Z, H_r, \widehat{H}_r; R) = \mathcal{F}(Z, \widehat{H}_r) + \frac{\sigma}{2} \sum_{i=1}^F \left\| \mathbf{h}_{r,i} - \left(\widehat{\mathbf{h}}_{r,i} + \frac{1}{\sigma^{[l]}} \mathbf{r}_i \right) \right\|^2 \quad (4)$$

where $R = R^{[l-1]} - \sigma \left(H_r - \widehat{H}_r^{[l-1]} \right)$ and with the column vector \mathbf{r}_i given by

$$R = [\mathbf{r}_1, \dots, \mathbf{r}_i, \dots, \mathbf{r}_F].$$

At iteration l , $\mathbf{r}_i^{[l-1]}$ can be considered as a constant so that the objective can be accordingly changed with a smaller set of variables as:

$$L_\sigma(Z, H_r, \widehat{H}_r) = \mathcal{F}(Z, \widehat{H}_r) + \sigma' \sum_{i=1}^F \left\| \mathbf{h}_{r,i} - \widehat{\mathbf{h}}_{r,i} \right\|, \quad (5)$$

where $\sigma' = \frac{\sigma}{2} \left(1 + \frac{1}{\lambda} \right)$ or $\sigma' = \sigma$. Considering $\sigma' \leq \sigma$, we still use σ to balance the original objective and the recovery error. In the next section, we also show that the l_2 norm is can be replaced by the l_1 norm. Then the solution for $\widehat{H}_r^{[l+1]}$ (given $Z^{[l+1]}$) is exactly $\Phi(H_d Z^{[l+1]}, \mathcal{M})$.

4. Manifold Constraint Transfer (MCT) for sparse representation

Given the solution presented in the previous section, we argue that the original objective is still complex to solve for a recognition problem due to the dimensionality of the cloned variables involved. However, by considering the relationship between the optimized variable and the input data from a manifold, we can further relax the minimization objective to be extremely efficient. In particular, the relationship between H_r and Z is a basic part of our optimization problem. To do that, we provide a solution which can guarantee that a new optimization method can solve (P2) by taking advantage of the linear relation between Z and the involved data H_r .

Given the previous problem solved with the ADMM algorithm, we demonstrate that the minimized objective function can be reformulated as:

$$\begin{aligned} & \underset{Z}{\text{minimize}} && \|Z\|_1 \\ & \text{subject to} && H_r = H_d Z \\ & && Z \in \mathcal{M}. \end{aligned} \quad (P3)$$

This last optimization problem (P3) provides the same solution of (P2) but it is characterized by a simpler cost function given by

$$L_\sigma(Z) = \|Z\|_1 + \sigma \sum_{i=1}^F \|\mathbf{z}_i - \mathbf{z}'_i\|_1, \quad (6)$$

where Z' is a cloned variable from Z for which $Z' \in \mathcal{M}$ and the vector \mathbf{z}_i is given by

$$Z = [\mathbf{z}_1, \dots, \mathbf{z}_i, \dots, \mathbf{z}_F].$$

In this way we detach the optimization of the l_1 norm cost from the manifold constraints that are effecting the cloned variable only. The manifold is defined again as a linear embedding as shown in Def. 1. Moving from (P2) to (P3), we can observe that the manifold constraint is transferred from the data to the variable, if there exists a linear correlation ($H_r = H_d Z$), which is the main theoretical contribution of MCT. Details about the proof are reported in the following.

Proof. Considering that the following inequality always holds true:

$$\|\mathbf{h}_{r,i} - \widehat{\mathbf{h}}_{r,i}\|_2 \leq \|\mathbf{h}_{r,i} - \widehat{\mathbf{h}}_{r,i}\|_1,$$

we change the previous minimization to the optimization of the upper bound of the original objective giving:

$$L_\sigma(Z, H_r, \widehat{H}_r) = \mathcal{F}(Z, H_r) + \sigma \sum_{i=1}^F \|\mathbf{h}_{r,i} - \widehat{\mathbf{h}}_{r,i}\|_1 \quad (7)$$

Here, we also replaced $\mathcal{F}(Z, \widehat{H}_r)$ with $\mathcal{F}(Z, H_r)$, because $H_r = \widehat{H}_r$. We can further prove that the variable Z should be constrained by the same manifold as the data samples, if there is a linear relationship between Z and H_r . As $\widehat{H}_r = H_d Z$, then $\widehat{H}_r^{[l+1]}$ is spanned by $H_d Z^{[0:l]}$. Thus, based on Def. 1, $\widehat{\mathbf{h}}_{r,i}$ can be estimated by linear embedding given by

$$\widehat{\mathbf{h}}_{r,i} = \sum_{j=1}^K H_d \mathbf{z}_i^{[0:l]} (1 - W_j), \quad (8)$$

where K is the number of samples. The manifold space is spanned by $H_d Z^{[0:l]}$ and iteratively the structure in the original space $H_d Z^{[0]}$ can be fully exploited in the process mentioned above. We now define a new variable Z' with $\mathbf{z}'_i = \sum_{j=1}^K \mathbf{z}_i^{[0:l]} (1 - W_j)$, which is calculated based on the same manifold embedding as that of \widehat{H}_r .

From the definition, we can say that Z' , given each vector in Z , is a linear embedding of the variable $Z^{[0:l]}$, considering that H_r and Z are linearly related and the geodesic distance on \widehat{H}_r is equal to the one in Z .

Again using the constraint $\widehat{H}_r = H_r$, we have that:

$$H_r - \widehat{H}_r = H_d(Z - Z'), \quad (9)$$

where the minimization on $\|H_r - \widehat{H}_r\|_1$ is solved by $\|H_d(Z - Z')\|_1$. According to Lemma 1, as shown in the Appendix, the following relation holds:

$$\|H_r - \widehat{H}_r\|_1 = \|H_d(Z - Z')\|_1 \leq \|H_d\|_1 \|Z - Z'\|_1, \quad (10)$$

and we alternatively minimize the upper bound of $\|H_r - \widehat{H}_r\|_1$. Given that $\|H_d\|_1$ is a constant, we can simplify the expression as:

$$\min \|H_d\|_1 \|Z - Z'\|_1$$

The above equation is equal to minimizing the upper bound of $\|H_d\|_1 \|Z - Z'\|_1$ giving:

$$\min \|Z - Z'\|_1$$

Consequently, the final objective for (P3) is as:

$$L_\sigma(Z) = \mathcal{F}(Z, H_r) + \sigma \sum_{i=1}^F \|\mathbf{z}_{r,i} - \mathbf{z}'_i\|_1 \quad (11)$$

where $Z' \in \mathcal{M}$. From above, the cloned variable Z' is given by a projection on a manifold spanned by Z and the structure of the original space is preserved in the projection. \square

Thus for each sample from a given class (manifold), we can optimize the above objective separately just based on an initialized manifold sample space. The details about the formulation of each step are shown in Algorithm 2.

Algorithm 2 Manifold Constraint Transfer (MCT) based on ADMM

- 1: set $k = 0$ and $\epsilon_{\text{best}} = +\infty$
- 2: Initialize $\sigma^{[0]} = 0.25$
- 3: Initialize $Z^{[0]}$ based on [29].
- 4: **repeat**
- 5:

$$\left(Z^{[k+1]} \right) = \underset{Z'}{\text{argmin}} L_{\sigma^{[k]}}(Z^{[k]}, Z'^{[k]}) \quad (12)$$

As solved in the appendix

- 6: $\epsilon = \|Z^{[k+1]} - Z^{[k]}\|^2$
 - 7: **if** $\epsilon < \eta \epsilon_{\text{best}}$
 - 8: $\sigma^{[k+1]} = \sigma^{[k]}$
 - 9: $\epsilon_{\text{best}} = \epsilon$
 - 9: **else**
 - 10: $\sigma^{[k+1]} = k\sigma^{[k]}$
 - 11: **endif**
 - 12: $k \leftarrow k + 1$
 - 13: **until** some stopping criterion
-

Summarising, our method is based on a sparse optimization process that is minimizing $\|Z\|_1$ with the constraints $H_r = H_d Z$ and $Z \in \mathcal{M}$. The manifold constraint on sparse representation comes from the fact that:

1. input samples are assumed to lie on a manifold;
2. the optimized variable lies on a manifold as well.

Solving for (P3) requires to initialize the sub-manifold space first, which we will discuss in the next section.

For what concern the computational complexity, the computation of Geodesic distance is based on the Dijkstra's algorithm, whose complexity is $\mathcal{O}(V \log(V) + E)$ where V, E represent the number of the samples and edges respectively. Also, the added pair of distances can be updated online. The number of the neighbours in the Geodesic distance calculation is empirically set to $K = 7$. For each testing face in the AR database, MCT can be run within 1 second on the Intel Xeon E5 2.0GHz DualCore with 8 GB RAM.

5. Experiments

We test our approach with several recognition problems in the literature in order to demonstrate the flexibility of the method. In details, experiments are related to face (AR database, Yale B), digits (USPS), objects (Oxford flowers) and action (UCF50, UCF sport) recognition problems. The experimental validation of the proposed method follows the one defined in [30]. Given the different recognition problems, there is the need to initialise the sub-manifold with respect to each dataset. In particular we augmented the data H_r for the testing images by:

Scheme 1 translating (1 and 2 pixels in each of the four directions) and rotating the images (1 and 2 degrees) for the problem of face and digit recognition.

Scheme 2 finding the 10 nearest neighbors from the validation set for the flower recognition and from the unlabelled test set in the case of face recognition.

Scheme 3 For action recognition problems, we add 20 different Gaussian noises [32] with 0 mean and increasing standard deviations (i.e. 0.0010, 0.0015, ...). Then we choose the 10 samples closer to the input to generate a sub-manifold. Although our procedure is not exactly the same as that in [32], ours is specially designed for a classification problem.

After obtaining the samples with one of the above schemes, we computed the sub-manifold space for the unsolved original variable using SRC [14] with the very same features of [30].

5.1. Face recognition.

Several algorithms in the literature leverage the fact that images of faces lie on a manifold. This provides an exemplar recognition problem since we explicitly use manifold constraints. For this task, we considered three datasets: the AR database [20], the extended Yale Face database B [2] and the LFW database [36, 38]. All these datasets are very challenging, due to their severe illumination and occlusion variations. We compare our method with other state-of-the-art sparsity face reconstruction methods, including SRC [14] and its latest variants FDDL [30]. We also compared two baseline methods, namely nearest subspace classifier (NSC) and linear support vector machine (SVM) implemented by [30].

The AR database consists of over 4,000 frontal images from 126 individuals. For each individual, 26 pictures were taken in two separated sessions. As in [30], we chose a subset consisting of 50 male subjects and 50 female subjects. For each subject, 7 images with illumination and expression changes from Session 1 were used for training, and the other 7 images with the same condition from Session 2 are used for testing.

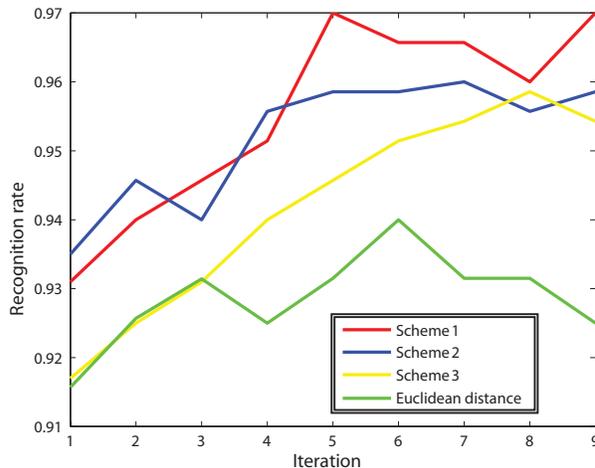


Figure 2. Performance over the sub-manifold initialization and iteration on the AR database

The Extended Yale B database consists of 2,414 frontal-face images from 38 individuals (about 64 images per subject) captured under various laboratory-controlled lighting conditions. For each subject, we randomly selected 20 images for training with the remaining images for testing.

The LFW database has 5749 images of different individuals in an unconstrained environment. LFW-a is specifically designed for face recognition problem, which is a version of LFW after alignment using a face alignment software. Given the protocol in [38, 39], we used the subjects in LFW-a having no less than ten samples thus obtaining a dataset with 158 subjects. For validating the results, after cropping the images to 121×121 and then resized to 32×32 , we randomly selected 5 samples for training and another 2 samples were used for testing.

Comparative results, reported in Table 1, show that the proposed method achieves higher performance than competing methods (SRC and other variants). We also investigated the performance of the three different sub-manifold initialization methods, across the iterations, on the AR dataset. Fig. 2 reports our findings and shows that best result is achieved for 5-6 iterations in all the cases. Despite all the data augmentation improved the performance in the AR database, **Scheme 1** should be preferred among the others. In the same figure, we also reported the standard Euclidean distance (minus 1) as the weight measure in Def. 1, which shows that it can also be used to improve the performance. However when the sample size is increased, MCT is significantly better. From an optimization viewpoint, the true neighbours could help our algorithm to converge to a “true” point, since the samples far from the input are given smaller weights in the process of embedding generation based on the Geodesic distance. On the other hand the Euclidean distance maybe lead to bigger weights for the samples actually

far from the inputs [4].

Dataset	SRC	NN	SVM	FDDL	MCT
AR	88.0	71.0	87.0	92.0	97.0
Yale-B	90.0	61.0	88.0	91.9	96.1
LFW	44.1	44.2	43.3	42.0	44.9

Table 1. Accuracy (%) for the face recognition task. All the results use the same parameters as in [30]

5.2. Digit recognition.

We performed handwritten digit recognition on the widely used USPS database [12]. It is composed by 7291 training images and 2007 test images. We compared our method with FDDL, SRC, DLSI [11], and LLC [26]. In addition, as baselines, the results of some problem-specific methods (i.e., the standard Euclidean KNN and SVM with a Gaussian kernel) are also listed. Again, the proposed method achieves the better performance as shown in Table 2.

SRC	LLC	DLSI	SVM	FDDL	MCT
4.48	6.05	3.98	4.2	3.69	2.99

Table 2. Error rates (%) for the digit recognition task on the USPS dataset

5.3. Action recognition

We then considered the UCF sport and the larger UCF50 datasets¹. The 140 videos clips in the UCF sport action dataset were collected from various broadcast sports channels, covering 10 sport action classes: driving, golfing, kicking, lifting, horse riding, running, skateboarding, swinging-(prommel horse and floor). The UCF50 dataset has 50 action categories for a total of 6680 realistic videos collected from YouTube. For both datasets we used action bank features as in [24].

We followed the experimental settings of [21, 1, 30] when testing UCF sport and we compared our method with FDDL, SRC [21], [1, 24]. The recognition rates are listed in Table 3 showing that our method provides better performance than all the other competing methods. We then evaluated the proposed method on the UCF50 action dataset by replicating the experimental settings of [24]. The results are shown in Table 3 where our method provides a better performance than SRC but slightly worst then FDDL.

5.4. Object categorization.

The last test performed is the multi-class object categorization problem on the 17 Oxford Flowers dataset [18] and

¹<http://crcv.ucf.edu/data/>

Dataset	[21]	[1]	[24]	SRC	FDDL	MCT
UCF Sport	83.6	86.6	90.7	92.9	94.3	95.1
UCF50	-	-	57.9	59.6	61.1	60.9

Table 3. The comparative results (accuracies %) on the action recognition tasks

we adopted the default experimental settings as pointed out by the authors², including training, validation, test splits and the provided features. For a fair comparison with the state-of-the-art methods such as MTJSRC [31], we also extended the original features as done in their experiments. We used the histogram intersection similarity of the recently proposed Frequent Local Histogram (FLH [3]) feature to generate new features. Table 4 lists the best results of NSC, SVM, MTJSRC-CG, SRC and FDDL. The proposed method could always improve the original SRC (which directly uses training samples as the dictionary) and compared similarly to FDDL.

NSC	SVM	MTJSRC-CG	SRC	FDDL	MCT
79.3	88.6	88.4	88.4	91.7	91.73

Table 4. The comparative results (accuracies %) for the Oxford flowers dataset using the FLH features.

6. Discussions and future work

This paper focuses on a new insight into sparse representation from the perspective of manifold constraint. Based on a simple neighbour-preserving embedding method, an ADMM scheme is proposed to solve our problem. Finally, based on a novel and general result, we transfer the data manifold constraints to the optimized variables. This paper actually proposes a new problem: “how to transfer the data structure into the optimization process”, and we show that the manifold constraints can be transferred from the data to the optimized variables if these are linearly correlated. The proposed approach is successfully applied on face, digit, action and general objects recognition showing a consistently increase on performance compared to the state of the art.

Acknowledgements

B. Zhang acknowledge the support of the Natural Science Foundation of China, under Contracts 61272052 and 61473086, and the Program for New Century Excellent Talents of the University of Ministry of Education of China.

²www.robots.ox.ac.uk/~vgg/data/flowers

Appendix

To solve the optimization problem constrained by a manifold, the ADMM method for Eq. (12) is described as:

1. Based on $Z^{[t]}$, we solve the following problem based on the inner point method [16]:

$$(Z^{[t+1]}) = \arg \min L_{\sigma}(Z^{[t]})$$

2. After $(Z^{[t+1]})$ is solved, we do

$$Z^{[t+1]} = \Phi(Z^{[t+1]}, Z^{[0:t]})$$

Lemma 1: If T is a matrix, the vector \mathbf{x} is transformed by T , and have :

$$\|T\mathbf{x}\|_1 \leq \|T\|_1 \|\mathbf{x}\|_1$$

Lemma 1 is actually easy to be proven, and a simple proof is described here.

Firstly, we define the matrix 1-norm as follows:

$$\|T\|_1 = \sum_{i=1}^n |\mathbf{t}_i|_1 \leq \max\{|\mathbf{t}_i|_1\}_{i=1}^n \quad (13)$$

where $T = [\mathbf{t}_1, \dots, \mathbf{t}_n]$.

$T\mathbf{x}$ is defined as:

$$T\mathbf{x} = \sum_i^n \mathbf{t}_i x_i$$

and we have:

$$\|\sum_i^n \mathbf{t}_i x_i\|_1 \leq \sum_i^n \|\mathbf{t}_i\|_1 \|x_i\|_1$$

Thus, based on Eq. (13), we have that:

$$\|T\mathbf{x}\|_1 \leq \|T\|_1 \sum_i^n \|x_i\|_1 \leq \|T\|_1 \|\mathbf{x}\|_1.$$

References

- [1] A. Yao, J. Gall, and L. Van Gool A hough transform-based voting framework for action recognition. *In Proc. IEEE Computer Vision and Pattern Recognition, 2010.* 7
- [2] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, 2001. 6
- [3] B. Fernando, E. Fromont, and T. Tuytelaars Effective use of frequent item set mining for image classification. *In Proc. European Conf. Computer Vision, 2012.* 7
- [4] C. Varini, A. Degenhard, and T. Nattkemper Isolle: Locally linear embedding with geodesic distance. *In Proc. Knowledge Discovery in Database: PKDD, 2005.* 3, 4, 7
- [5] G. Cabanes and Y. Bennani. Learning topological constraints in self-organizing map. *in proc. of ICONIP, 2010.* 1
- [6] S. Cai, W. Zuo, L. Zhang, X. Feng, and P. Wang. Support vector guided dictionary learning. *ECCV 2014.* 2
- [7] M. Chang, L. Ratinov, and D. Roth. Guiding semi-supervision with constraint-driven learning. *ACL*, pages 280–287, 6 2007. 1
- [8] J. Chen, X. Chen, J. Yang, S. Shan, R. Wang, and W. Gao. Optimization of a training set for more robust face detection., *Pattern Recognition*, 42(11), Pages 2828–2840, 2009. 3, 4
- [9] J. Chen, R. Wang, S. Yan, S. Shan, X. Chen, and W. Gao. Enhancing human face detection by resampling examples through manifolds., *IEEE Transactions on Systems, Man, and Cybernetics-Part A*, pages vol. 37, no. 6, pp. 1017–1028, 2007. 3, 4
- [10] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear modeling via augmented lagrange multipliers (balm). *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(8):1496–1508, 2012. 3
- [11] I. Ramirez, P. Sprechmann, and G. Sapiro Classification and clustering via dictionary learning with structured incoherence and shared features. *In Proc. IEEE Computer Vision and Pattern Recognition, 2010.* 7
- [12] J.J. Hull A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1994 7
- [13] K. Engan, S.O. Aase, and J.H. Husoy Method of optimal directions for frame design. *In Proc. IEEE Int. Conf. Acoust.Speech, Signal Process, 1999.* 2
- [14] K. Huang and S. Aviyente Sparse representation for signal classification. *In Proc. Neural Information and Processing Systems, 2006.* 6
- [15] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Trans. PAMI*, pages pp:1409–1422, 2012.
- [16] K. Koh, S.-J. Kim, and S. P. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research, 2007.* 8
- [17] M. Long, G. Ding, J. Wang, J. S. Y. Guo, and P. S. Yu. Transfer sparse coding for robust image representation. *Proc. IEEE Computer Vision and Pattern Recognition, CVPR 2013.* 2
- [18] M. Nilsback and A. Zisserman A visual vocabulary for flower classification. *In Proc. IEEE Computer Vision and Pattern Recognition, CVPR 2006.* 7
- [19] M. Yang, L. Zhang, J. Yang, and D Zhang. Robust sparse coding for face recognition. *In Proc. IEEE Computer Vision and Pattern Recognition, 2011.* 2
- [20] A. Martinez and R. Benavente. The ar face database. CVC technical report. *CVC Technical Report 24, 1998.* 6
- [21] Q. Qiu, Z. Jiang, and R. Chellappa Sparse dictionary-based representation and recognition of action attributes. *In Proc. of Conf. Computer Vision, 2011.* 7

- [22] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. *In Proc. IEEE Computer Vision and Pattern Recognition, 2010.* 2
- [23] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000. 3, 4
- [24] S. S. and C. J.J. Action bank: A high-level representation of activity in video. *In Proc. IEEE Computer Vision and Pattern Recognition, 2012.* 7
- [25] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, , and Y. Ma. Towards a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell., 2012* 2
- [26] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. *In Proc. IEEE Computer Vision and Pattern Recognition, 2010.* 2, 7
- [27] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao. Manifold-manifold distance and its application to face recognition with image sets. *IEEE Transactions on Image Processing, 2012.* 4
- [28] Z. Wang, J. Yanga, N. M. Nasrabadi, and T. S. Huang. A max-margin perspective on sparse representation-based classification. *In Proc. of ICCV, 2013.* 2
- [29] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell., 2009.* 1, 2, 5
- [30] M. Yang, L. Zhang, X. Feng, and D. Zhang. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, 2014. 2, 6, 7
- [31] X. Yuan and S. Yan. Visual classification with multitask joint sparse representation. *In Proc. IEEE Computer Vision and Pattern Recognition, 2010.* 7
- [32] Salah Rifai, Yoshua Bengio, Yann N. Dauphin, Pascal Vincent. A Generative Process for Sampling Contractive Auto-Encoders *in proc. of ICML, 2012.* 6
- [33] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear Factorization via Augmented Lagrange Multipliers *in 11th European Conference on Computer Vision, ECCV 2010.* 3
- [34] YiGang Peng, Arvind Ganesh, John Wright, Wenli Xu, Yi Ma. RASL: Robust Alignment by Sparse and Low-Rank Decomposition for Linearly Correlated Images *IEEE Trans. Pattern Anal. Mach. Intell., 2012.*
- [35] T. Lin and H. Zha, Riemannian Manifold Learning, *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, no. 5, pp. 796-809, 2008
- [36] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst
- [37] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? *In ICCV, 2011.* 6
- [38] P. Zhu, L. Zhang, Q. Hu, and S. C. K. Shiu. Multi-scale patch based collaborative representation for face recognition with margin distribution optimization. *ECCV, 7572:822–835, 2012.*
- [39] Jiqing Wu, Radu Timofte, Luc Van Gool, Learned Collaborative Representations for Image Classification. *WACV 2015: 456-463* 6