Mixed-norm Sparse Representation for Multi View Face Recognition

Xin Zhang‡, Duc-Son Pham†, Svetha Venkatesh‡, Wanquan Liu†, Dinh Phung‡

‡ *PRaDA Deakin University*

† *Department of Computing Curtin University*

Abstract

Face recognition with multiple views is a challenging research problem. Most of existing works have focused on extracting shared information among multiple views to improve recognition. However, when the pose variation is too large or missing, 'shared information' may not be properly extracted, leading to poor recognition results. In this paper, we propose a novel method for face recognition with multiple view images to overcome the large pose variation and missing pose issue. By introducing a novel mixed norm, the proposed method automatically selects candidates from the gallery to best represent a group of highly correlated face images in a query set to improve classification accuracy. This mixed norm combines the advantages of both sparse representation based classification (SRC) and joint sparse representation based classification (JSRC). A trade off between the ℓ_1-norm from SRC and $\ell_{2,1}$-norm from JSRC is introduced to achieve this goal. Due to this property, the proposed method decreases the influence when a face image is unseen and has large pose variation in the recognition process. And when some face images with a certain degree of unseen pose variation appear, this mixed norm will find an optimal representation for these query images based on the shared information induced from multiple views. Moreover, we also address an open problem in robust sparse representation and classification which is using ℓ_1-norm on the loss function to achieve a robust solution. To solve this formulation, we derive a simple, yet provably convergent algorithm based on the powerful alternative directions method of multipliers (ADMM) framework. We provide extensive comparisons which demonstrate that our method outperforms other state-of-the-arts algorithms on CMU-PIE, Yale B and Multi-PIE databases for multi-view face recognition.

Keywords:
Multi-Pose Face Recognition, Sparse Representation Classification, ADMM, Group Sparse Representation, Multi-task Learning, joint dynamic sparse representation classification, unsupervised learning, convex optimization, robust face recognition

1. Introduction

Face recognition is one of non-intrusive biometrics. Due to the emerging demand in surveillance and security, it is an important research topic in pattern recognition [1]. According to previous literature surveys [2, 3], extensive studies have been done to resolve the face recognition issues, such as pose, illumination, expression and occlusion, etc. [4, 5, 6]. However, most of these methods are based on a single input image. They identify a subject by matching a single query face image with all gallery images one by one. In practice, it is common that the query face image is noisy or its pose may be missing in the gallery, thus working with a single face image is likely to be unreliable in real-world applications. On the other hand, multiple views of a same subject can be obtained easily with current technology. For instance, a sequence of face images from a subject with a large degree of pose variations may be observed over a time interval by a surveillance camera or multiple snapshots are captured by video camera networks at same time from different viewpoints. This will produce a large number of query images for recognition tasks. Since multiple view images are from the same subject under different time or viewpoint, there is likely some shared information across those face images. The existing face recognition techniques have not investigated the inter-correlation among the query images, therefore, exploiting the using of these shared information becomes an important work.

In the face recognition literature, several popular classifiers have been developed. The nearest neighbour (NN) is one of the most common and popular classifiers [7]. The NN classifies the query face image based on its closest neighbour in the gallery set. However, this classifier is sensitive to outliers. The NN classifier is generalized to nearest subspace (NS) [8]. Instead using a single image to perform classification, NS classifies a face based on the best linear representation in terms of all the gallery images in each class. Since the classification decision is made by all samples, NS is more robust than NN. Sparse representation-based classification (SRC) [6] seeks a balance between these two extreme cases, it represents a query image by adaptively choosing a minimum number of atoms (samples in gallery) from both within each class and across multiple classes. SRC has been shown more robust and effective than NN and NS on some common face recognition issues, such as occlusion and corruption. Encouraged by the SRC framework, a large number of its extensions have been proposed [9, 10, 11] and they have achieved state-of-art performance. However, they are limited to single query face image for recognition.

Recently, a growing interest [12, 13, 14, 15] in face recognition from an image set has emerged. Rather than using a single query image to perform recognition, multiple face images of the same subject are used as an input. In general, the system identifies a query subject based on a set of input images from known subjects in the gallery. The face images in both gallery and query sets may have large variations in pose, illumination, etc. By using multiple face images of the same subject in the query, the robustness of the recognition system has been improved significantly compared with single-input systems. In [16], an extended volume Local Binary Patterns is introduced to exploit the information among frames. It can achieve a good performance, but it requires sequential images from a video. Another approach to achieve this goal is by measuring the distance between the query set and each class in the gallery set. Inspired by label propagation [17], a graph-based classification for multiple view face recognition has been proposed [12]. It converts the face image set into a similarity graph, and then uses a class-wise graph matching procedure to compare this similarity graph with the graph generated by each class in the gallery. In [13], face images are represented as a feature vector in an affine feature space. They build an affine hull for each image set (query set and each class in the gallery). The geometric distances between the affine hull of query and of each class in the gallery are used to make the classification decision. A multi-class group Lasso is introduced in [18]. Images are represented by Local Binary Patterns [19], then the best suitable features are selected to measure the distance between each pair of sets. These methods treat each set of the gallery face images as a linear subspace, and use subspace distance to identify the query subject from subjects in the gallery independently. Thus, they have two limitations: (1) they cannot exploit information across multiple classes; (2) when there is a large difference between images in the same class, such as large pose variations, these methods can perform poorly.

Since SRC considers both within each class and between multiple class factors, a multiple test samples generalization of SRC is introduced, known as joint sparse representation-based classification (JSRC) [14, 20]. This method assumes that the query face images share the same sparsity pattern. The shared information can be exploited by using this assumption. Instead of solving the SRC problem for each query image, JSRC solves a set of query images from the same subject. It adaptively selects a minimum number of atoms from gallery images, these atoms can best represent every query images at same time. However, this assumption will not hold when there are large pose differences in the query images. For example, if a frontal face and a 90° right face exist in the query set at the same time, it is impossible to find an atom in the gallery to represent both of them at a same time accurately. In order to overcome this issue, joint dynamic sparse representation-based classification (JDSRC) was proposed in [15]. The authors in [15] argue that the same sparsity patterns is not necessary at the atom level, these patterns should be at the class level. To capture this model, they introduced a new concept of joint dynamic sparsity. This joint dynamic sparsity brings in flexibility to atom selection of JSRC. When the pose variation is large in the query images, JDSRC does not necessarily select the same atom for all poses as JSRC. Instead, JDSRC selects atoms from the whole class to represent all poses. Nevertheless, when a pose appears in the query but is missing in the gallery, JDSRC will be forced to select a 'similar' atom from the gallery to represent it. This may not lead to a robust solution. In addition, the JDSRC is achieved by an extension of simultaneous orthogonal matching pursuit [20] which is a naive greedy method and may be not convergent. Therefore, a new algorithm is needed to solve this challenging multi-view (multi-pose) problem.

In [21], the authors argue that the robustness of SRC based methods should be achieved by using the ℓ_1-loss function instead of the ℓ_2-loss in SRC. However, it was left as an open question, because solving via standard linear programming techniques is computationally expensive. The sparse representation is inspired from compressed sensing (CS). In the statistical signal processing community, the core CS problem is finding a sparse linear combination of signal atoms from an overcomplete dictionary [22, 23]. It was then applied to face recognition in [6]. As solving this CS problem is close to the Lasso in statistics in functional form, extensions to the basic sparse solution have been observed in related areas. A robust Lasso, which explicitly models the corruptions, is proposed and analysed in [24]. Statistically, this is more generic and provably better than the least entropy and error correction alternative discussed in a rejoinder [25] by the authors of SRC against the paper of Shi *et al.* [21]. However, this is obtained at the cost of an extra regularization parameter. In the related robust CS paper [26], a slightly different loss function, known as Huber's robust loss function is used. However, it requires the estimates of the Huber's parameters, which brings additional computational burden.

In this paper, we propose a novel mixed norm sparse representation classification (MSRC) method for multi-view face recognition. The proposed method has the similar ability to JDSRC, it allows some degree of flexibility in atom selection procedure of JSRC. On one hand, as SRC works with a single query image, it cannot exploit the shared sparsity pattern across query images. Thus, it will ignore the influence of large pose variations in the query images. On the other hand, JSRC struggles with shared information among query images, but it can easily be affected by the pose variations. Therefore, it is natural to strike a balance between them. Our MSRC achieves this goal. It exploits the correlation among the variance face images in the query and it also brings the flexibility to the atom selection to achieve an accurate and sparse representation. Moreover, to achieve more robustness, our MSRC uses the ℓ_1-loss instead of the general ℓ_2-loss. Indeed, the ℓ_1-norm loss function we use in this work, which is also an open question disscussed in [21], is also known in the robust statistics literature to be optimal for noise modelled as a Cauchy distribution.

The contributions of this work are as follows: (1) we derive a simple, provably convergent, and computationally efficient algorithm based on the framework of alternative direction method of multipliers (ADMM) [27]; (2) we establish a novel multi-view face recognition in a robust form to exploit the similarities between different images in the query images. (3) we provide extensive experiments to show the advantages of our proposed method against other state-of-the-arts.

The paper is organized as follows. We first briefly review the related works in Section 2. Then we derive our Multi-pose face recognition via sparse representation (MSRC) based on the ADMM framework in Section 3. Finally, we provide extensive experiments on CMU-PIE, Yale B and Multi-PIE datasets in Section 4. Section 5 concludes the paper.

The Matlab implementation of proposed method is publicly available at `https://sites.google.com/site/dspham/code`.

2. Related works

2.1. Face recognition via sparse representation

In sparse representation classification (SRC), given a set of gallery images $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ where each $\mathbf{a}_i \in \mathbb{R}^d$, one seeks a sparse combination of these images to represent an unknown image $\mathbf{y}$

$$\mathbf{y} \approx \sum_{i=1}^{N} x_i \mathbf{a}_i = \mathbf{A}\mathbf{x}. \tag{1}$$

Such a sparse solution can be found by solving following problem in [6]

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{y}. \tag{2}$$

Or alternatively, by solving

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 < \varepsilon. \tag{3}$$

According to CS theory [28, 29], (3) can be solved by convex optimization using ℓ_1-norm regularization

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1. \tag{4}$$

This convex optimization problem can be solved efficiently with many algorithms developed specifically for CS. Then SRC combines this sparse representation with nearest subspace classification. In other words, it computes the class-specific residual vector

$$\mathbf{r}_k = \mathbf{y} - \mathbf{A}_k \mathbf{x}_k, \tag{5}$$

where $\mathbf{A}_k$ is the sub-matrix of $\mathbf{A}$ that corresponds to all gallery images in class k, and $\mathbf{x}_k$ is the sub-vector of $\mathbf{x}$ with the corresponding sparse coefficients. Then the target $\mathbf{y}$ is classified according to the minimum ℓ_2-norm of the residual vectors

$$\text{class}(\mathbf{y}) = \arg\min_k \|\mathbf{r}_k\|_2^2. \tag{6}$$

2.2. Multi-pose face recognition via sparse representation

In many practical situations, it is desirable to recognize a number of unknown faces at the same time, such as, recognizing a person from a video sequence. Under this situation, the face images usually come from one subject with different poses (views). If we simply use SRC to perform multiple views face recognition, the sparse representation vectors will be generated individually (see Figure 1(a)). Information between different views is not involved in this scenario. Therefore, it is beneficial to exploit shared information across those face images. In this section, we will revise two works which have been proposed to extract the shared information.

We first proceed with some neccessary notations. Consider a gallery image set $\mathbf{A}$, which contains c classes. Each class $\mathbf{A_i}$ has N_i face images that may be captured with different poses. Suppose that $\mathbf{A} = [\mathbf{A}^1, \ldots, \mathbf{A}^c]$, and

$\mathbf{A}^i = [\mathbf{a}_1^i, \ldots, \mathbf{a}_{N^i}^i] \in \mathbb{R}^{d \times N^i}$, where d is the dimensionality of the images. Denote a test set of face images as $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M]$ where each $\mathbf{y}_i \in \mathbb{R}^d$. These M images may also be captured with different poses, but from the same person. The sparse representation coefficient matrix can be denoted as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$ with respect to $\mathbf{A}$.

2.2.1. Joint Sparse Representation Classification

In order to exploit the shared information across multiple views of the same subject, we rewrite the formulation of original SRC in the multi-task form as follows

$$\{\hat{\mathbf{x}}_i\}_{i=1}^M = \arg\min_{\mathbf{x}_i} \|\mathbf{x}_i\|_1 \tag{7}$$

$$\text{subject to} \sum_{i=1}^M \|\mathbf{A}\mathbf{x}_i - \mathbf{y}_i\|_2^2 < \varepsilon. \tag{8}$$

Recall that each $\mathbf{x}_i$ represents a pose image from $\mathbf{Y}$, and its rows represent the weights of corresponding gallery images. In addition, all images in $\mathbf{Y}$ comes from the same subject. Therefore, a joint sparse assumption can be applied to extract the shared information across all images in $\mathbf{Y}$ [20], which implies that multiple sparse representation vectors share the same sparsity pattern. For example, face images for each subject must contain some common features invariant to views. A same set of atoms may be used to represent for all views as shown in Figure 1(b). Therefore, by solving the following problem, the sparse representation vectors for multiple views can be found:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} ||\mathbf{X}||_{2,1} \tag{9}$$

$$\text{subject to } ||\mathbf{Y} - \mathbf{A}\mathbf{X}||_F^2 < \varepsilon \tag{10}$$

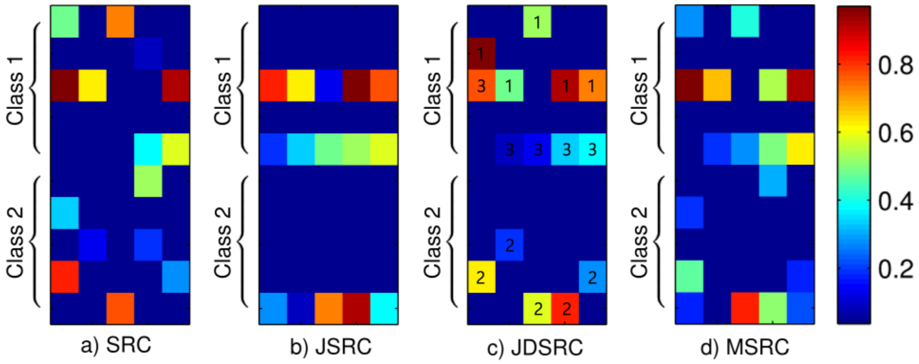

Figure 1: Four sparsity models for the coefficient matrix $\mathbf{X}$. The column represents the poses/images in the gallery as grouped by subjects, the row represents different images in the unknown set $\mathbf{Y}$. Each column denotes a sparse representation vector and each square denotes a coefficient. *a*) Independent sparsity (as in SRC): all coefficients are selected independently based on ℓ_1 regularization; *b*) Joint sparsity (as in JSRC): only few gallery images/poses are selected simultaneously by all test images; *c*) Joint dynamic sparsity (as in JDSRC): sparsity in terms of sorted active groups, active coefficients are *multiples* of the number of test images; *d*) Mixed Sparsity: a well trade-off balance of both *a*) and *b*) to adaptively select the suitable class-level and atom-level sparsity.

where $\|\bullet\|_{2,1}$ is defined as the sum of the ℓ_2-norm of all rows of a matrix and a Frobenius norm is used for reconstruction error. By introducing $\|\bullet\|_{2,1}$, the sparse representation matrix will have dense coefficients row-wise and sparse coefficients column-wise (see Figure 1(b)). This method is called JSRC. However, as stated by [15], the assumption that all the views share the same sparsity pattern is not applicable when solving multi-view face recognition with large variations. As face images could be captured from largely different angles, the shared information would be less when the difference between images in $\mathbf{Y}$ increase. Therefore, forcing the entire views to share the same set of atoms is not applicable to real-world multi-view face recognition.

2.2.2. Joint Dynamic Sparse Representation Classification

To overcome above issues with JSRC, it is argued that each view can be better represented by a different set of samples from the same class. The sparse representation vectors should share the same pattern across one subject, but not at the atom level [15] (see Figure 1(c)). Based on this assumption, they introduced the JDSRC model. Dynamic active sets are the core part of JDSRC, which allows it to exploit the joint dynamic sparsity prior for multi-view face recognition. The dynamic active sets are denoted as $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_s] \in \mathbb{R}$. Each dynamic active set $\mathbf{g}_i$ contains the (row-) indices of a set of coefficients which belong to the same class in coefficient matrix $\mathbf{X}$. Only one index is selected in each column of $\mathbf{X}$ for one dynamic active set. For example, $\mathbf{g}_i(j)$ refers the row index for j-th column of the coefficients matrix $\mathbf{X}$ in dynamic active set i. Based on these dynamic active sets, the following JDSRC model was developed in[15]

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} ||\mathbf{Y} - \mathbf{A}\mathbf{X}||_F^2 \tag{11}$$

$$\text{subject to } ||\mathbf{X}||_G \leq K, \tag{12}$$

where K is the sparsity level. Here, $||\mathbf{X}||_G$ is a combination of ℓ_2-norm and ℓ_0-norm based on dynamic active sets. The ℓ_2-norm is applied to the selected coefficients of each dynamic active set $\mathbf{g}_i$ individually, then the ℓ_0-norm is applied across all dynamic active sets. This joint dynamic sparsity regularization term is defined as follows

$$||\mathbf{X}||_G = ||[||\mathbf{x}_{\mathbf{g}_1}||_2, ||\mathbf{x}_{\mathbf{g}_2}||_2, \ldots]||_0, \tag{13}$$

where $\mathbf{x}_{g_i}$ indicates a set of coefficients that associated with dynamic active set $\mathbf{g}_i$:

$$\mathbf{x}_g = \mathbf{X}(g_s) = [\mathbf{X}(g_s(1), 1), \ldots, \mathbf{X}(g_s(M), M)]^T \in \mathbb{R}^M \tag{14}$$

To solve the problem with ℓ_0-norm and joint dynamic sparsity constraint, the authors of [15] proposed a greedy JDSRC algorithm which is similar to Simultaneous Orthogonal Matching Pursuit (SOMP) [20] and Compressive Simultaneous Orthogonal Matching Pursuit (CoSOMP) [30]. It consists of three steps: 1). select new candidates based on current residuals. 2). add these candidates to the selected atom set. 3). find new coefficients to reduce the size of an atom set to the specified sparsity level by using this atom set; 4). update the residuals based on this new atom set. These four steps are repeated until certain convergence conditions are satisfied [30]. In JDSRC, they introduce a new way to select atoms by using dynamic active sets. They first generate dynamic active sets by using the coefficients obtained from the previous step. They use all the atoms which have the largest absolute coefficients for each view from one class as one dynamic active set. Then they remove these large coefficients from the matrix, and select a new dynamic active set again. This procedure is repeated until there are no coefficients left in the coefficients matrix. After that, the ℓ_2-norm is applied to the coefficients from each dynamic active set, and the atoms from the active sets with K largest ℓ_2-norm will be selected as candidates for OMP optimization. Since the dynamic active set is selected from each class and the ℓ_2-norm is applied on dynamic active set separately, the sparse representations of JDSRC are forced to share the patterns only from the same class in order to exploit the common information for the same subject, this enhances the discriminative ability between different subjects.

However, there are a few issues with JDSRC. First, there is only one candidate selected for each view. This brings in two problems: 1) if the second largest atom in one view is much greater than the other view, it will miss a chance. 2) If a query view does not completely exist in the gallery for a subject, this candidate selection will lead to a false hit. Second, due to the complexity of the dynamic active set design, this JDSRC model cannot be solved by convex optimization. Therefore, a greedy method is often used. Since the convergence of the greedy method is not guaranteed, a robust and accurate solution may not be achieved.

3. Multi-view mixed norm robust sparse representation

3.1. Model formulation

In this section, we present our mixed-norm sparse representation classification (MSRC) method to overcome the issues with JDSRC. To start with, we recall from [21] that as the ℓ_2-norm used for residuals in SRC may not lead to a robustness solution, a ℓ_1-norm should be applied on the residuals:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^N} ||\mathbf{y} - \mathbf{A}\mathbf{x}||_1 \tag{15}$$

Therefore, the proposed method is designed to fulfil the following requirements:

- Shared information in the query set needs to be considered;

- A dynamic atom selection is needed to avoid large pose variations;

- A robustness solution has to be achieved;

- It has to be solvable by convex optimization.

As discussed, if images in the query set are from the same subject, there is likely some shared information across all images which can help to identify the subject. It is natural to apply the $\ell_{2,1}$-norm on the representation matrix to achieve a dense solution in each row and with a minimum number of rows (Figure 1(b)). It has been shown that this can achieve good performance when the images in the query set are highly correlated with each other (with a small pose variation) in JSRC. However, when JSRC encounters a large pose variation, this naive application could not achieve satisfactory performance. The $\ell_{2,1}$-norm thus reduces classification performance. On the other hand, the multi-task version of the original SRC will typically select atoms in an image-versus-image manner. The representation matrix is constructed based on the best representation of the input images, which does not exploit the shared information (Figure 1(a)). Although this characteristic could not help finding the shared pattern, it would not be confused by the increased pose variation. Therefore, we propose a new model to combine the $\ell_{2,1}$ and normal ℓ_1-norm to solve them in the same time. The final decision is not based on any individual factor, it is an overall view (Figure 1(d)).We note that the difference between the sparisty patterns (c) and (d) as shown in Figure 1 is subtle. The sparsity pattern (c) as found in JDSRC is also group-wise. However, each group for each subject class is not restricted to a row in pattern (b) of JSRC. Rather, it may span across multiple rows depending on how active groups are selected. Important properties of JDSRC's active groups are: non-overlapping, having equal sizes, and having exactly one coeffícent in each column. Only the top active groups are finally chosen through global optimization. Thus, the number of active coefficients per subject class is always a *multiple* of the number of images in the test set. On the contrary, the sparsity pattern (d) of MSRC is not group-wise, but a well traded-off balance between group-wise and element-wise. This is achieved through a novel combination of both SRC and JSRC, which gives the strength of both methods and eliminates their weaknesses.

To describe our model, we first extend the original SRC to the following robust and stable formulation for sparse representation:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^N} ||\mathbf{y} - \mathbf{A}\mathbf{x}||_1 + \lambda\|\mathbf{x}\|_1 \tag{16}$$

Here, the regularization parameter λ specifies the desired sparsity. Clearly, the robust sparse formulation is even more general than (15), because (15) is a special case when one sets $\lambda = 0$. Thus, solving this formulation allows one to obtain a solution for (15) easily.

Then, this formulation needs to be converted into a multi-task version. The individual robust sparse representation problems are

$$\begin{aligned} \hat{\mathbf{x}}_1 &= \arg\min_{\mathbf{x}_1\in\mathbb{R}^N} \{\|\mathbf{y}_1 - \mathbf{A}\mathbf{x}_1\|_1 + \lambda\|\mathbf{x}_1\|_1\}, && (17) \\ &\dots && \\ \hat{\mathbf{x}}_T &= \arg\min_{\mathbf{x}_T\in\mathbb{R}^N} \{\|\mathbf{y}_T - \mathbf{A}\mathbf{x}_T\|_1 + \lambda\|\mathbf{x}_T\|_1\}. && (18) \\ & && (19) \end{aligned}$$

We collect the variables in matrix quantities

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_1, \dots, \mathbf{x}_T] && (20) \\ \mathbf{Y} &= [\mathbf{y}_1, \dots, \mathbf{y}_T] && (21) \end{aligned}$$

then we can write all the single tasks more conveniently in a matrix form as

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_1 + \lambda\|\mathbf{X}\|_1. \tag{22}$$

Here, the ℓ_1-norm for matrices is defined as $\|\mathbf{X}\|_1 = \sum_{i,j} |X_{ij}|$. The ℓ_1-norm used on the residuals in the first term could prevent the bad influence of noise in image pixels. Thus, a more robust solution can be delivered by this formulation. In addition, as we mentioned above, the second term allows us to select the best representation in an atoms level. Thus, the large pose variations do not affect this representation.

Next, we introduce information sharing between different face views. Recall that each column of the coefficient matrix $\mathbf{X}$ represents one view of a subject, and each row represents the weights of the corresponding gallery images in all views of that same subject. We apply the same hypotheses with JSRC, the shared information appears in each face image for one subject onto the previous formulation (22). A $\ell_{2,1}$-norm is used on the coefficients matrix $\mathbf{X}$ to exploit the shared information.

To capture this modelling, we propose the following mixed-norm solution

$$\hat{\mathbf{X}} \quad = \quad \arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_1 + \lambda\|\mathbf{X}\|_{\star}^{\gamma} \tag{23}$$

where the mixed norm is defined as

$$\|\mathbf{X}\|_{\star}^{\gamma} = \gamma\|\mathbf{X}\|_1 + (1-\gamma)\|\mathbf{X}\|_{2,1}. \tag{24}$$

Here, the block regularizer $\| \bullet \|_{2,1}$ is defined as the sum of the ℓ_2-norm of all rows of a matrix. It is known from statistics that such ℓ_2-norm promotes *dense* solutions. The parameter γ controls the trade-off between absolute sparse ($\gamma = 1$) and absolute dense in the row with minimum number of rows

($\gamma = 0$). When $\gamma = 1$, this reduces to the multi-task version of robust SRC. When $\gamma = 0$, it is a special robust version of JSRC. For γ between 0 and 1, the formulation automatically adapts to the underlying statistics. The proposed formulation is termed mixed-norm sparse representation classification (MSRC).

Since the coefficient matrix is found by the mixed norm constraint, it will have certain advantages: 1). The proposed method could perfectly exploit the shared information across the images in the query set by the $\ell_{2,1}$-norm on $\mathbf{X}$. 2). By introducing the ℓ_1-norm on $\mathbf{X}$, the proposed method could overcome "miss chance" and "false hit" issues in JDSRC. When the second largest atom in one view is much greater than the largest atom in the other view, this "second" largest atom may be captured by the ℓ_1-norm in (24). When some pose images in the query do not exist in the gallery, the overall weights for these images will automatically decrease. Therefore, the valid distance is defined by the remaining high correlated face images in the query.

To illustrate the proposed method, we consider a synthetic example, wherein there are two subject classes A and B, each with 4 images of varying poses, and a test set of 4 images also with varying poses. The test set has the groundtruth of subject class A. Figure 2 shows spatially the pose distribution of all images. Note that the placement of the images is not meant to be exact as it is only a conceptual sketch. As can be seen, the gallery images of subject A has 4 different poses concentrated around the frontal pose, whilst the 4 gallery images of subject B spread out in the pose space. The 4-image test set to be recognized also has widespread poses. In SRC, images in the test set tend to select the nearest gallery images as their representatives in

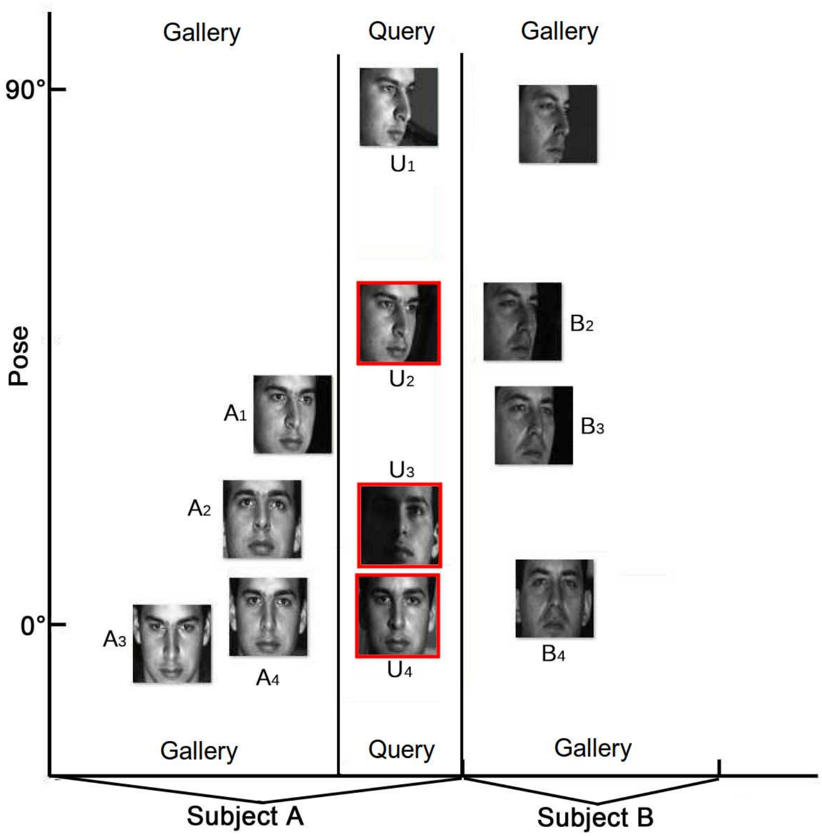

Figure 2: A conceptual illustration of the spatial pose distribution of the images in the training and test sets of a synthetic example.

a rough proximity sense. Thus, the spatial distribution in Figure 2 aids in the explanation of the coefficients obtained by SRC, JSRC, and MSRC respectively as shown in Figure 3. When the intra-class variation is too large in query set (as shown in this figure), SRC and JSRC do not perform well. The closest neighbours for U_1 and U_2 are from subject B and for U_3 and U_4 are from subject A, SRC typically selects atoms based on the best representation from subjects A and B, which is a single input image. Thus, if one simply performs majority voting from individual SRC solutions for each image in a pose set and does not consider other input images, this may lead to a failure because of pose similarity. On the other hand, JSRC finds the best candidates as A_1, B_2 and B_3 as it treats all the images in the query set as a whole space. Thus, it may not perform well in this case, because JDSRC is forced to select some candidates to represent U_1, even there is no similar pose in the gallery for subject A. Due to large pose variations in $U1$ and $U2$, both SRC and JSRC have a difficulty in distinguishing between class A and B, which is evident through the values of their coefficients (Figure $(3.a)$ and $(3.b)$), whose sums are similar between the two classes. On the other hand, the coefficients of MSRC reveal better discrimination as they carry the strength of both SRC and JSRC (Figure $(3.c)$). Its coefficients corresponding to class A are enhanced in a sense of the total sum when compared to those of either SRC or JSRC. In particular, its coefficients corresponding to the gallery image $A1$ become more dominant as they are also selected in SRC and JSRC. Likewise, its coefficients corresponding to class B are degraded in a sense of the total sum when compared to those of either SRC or JSRC. The coefficients corresponding to the gallery image $B1$ become less dominant as they are only selected in SRC but not JSRC. Note that although the coefficients corresponding to the gallery image $B2$ and $B3$ are also enhanced, their values are not sufficient enough when compared with those corresponding to class A.

3.2. Multi-task mixed norm algorithm

We now discuss a solution for the formulation above. We follow the ADMM framework in the convex optimization literature [27] (Readers who are not familiar with the basics of ADMM are referred to [27] for a background). By utilising the ADMM framework, we show that our algorithm is computionally efficient. Note that this problem is convex in $\mathbf{X}$ and hence there exists a global minimum. Thus, it completely avoids the convergence problem in JDSRC which solved by a greedy search.

For simplicity, we denote $\alpha = \lambda\gamma$ and $\beta = \lambda(1-\gamma)$, and we can express the problem as

$$\begin{aligned} \min_{\mathbf{X},\mathbf{V},\mathbf{Z},\mathbf{T}} \quad & \|\mathbf{V}\|_1 + \alpha\|\mathbf{Z}\|_1 + \beta\|\mathbf{T}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{AX} - \mathbf{Y} - \mathbf{V} = \mathbf{0} \\ & \mathbf{X} - \mathbf{Z} = \mathbf{0}. \\ & \mathbf{X} - \mathbf{T} = \mathbf{0}. \end{aligned} \tag{25}$$

Note that an additional variable $\mathbf{T}$ is introduced to the single-task case to effectively decouple the block regularization. Thus, we can consider the aug-

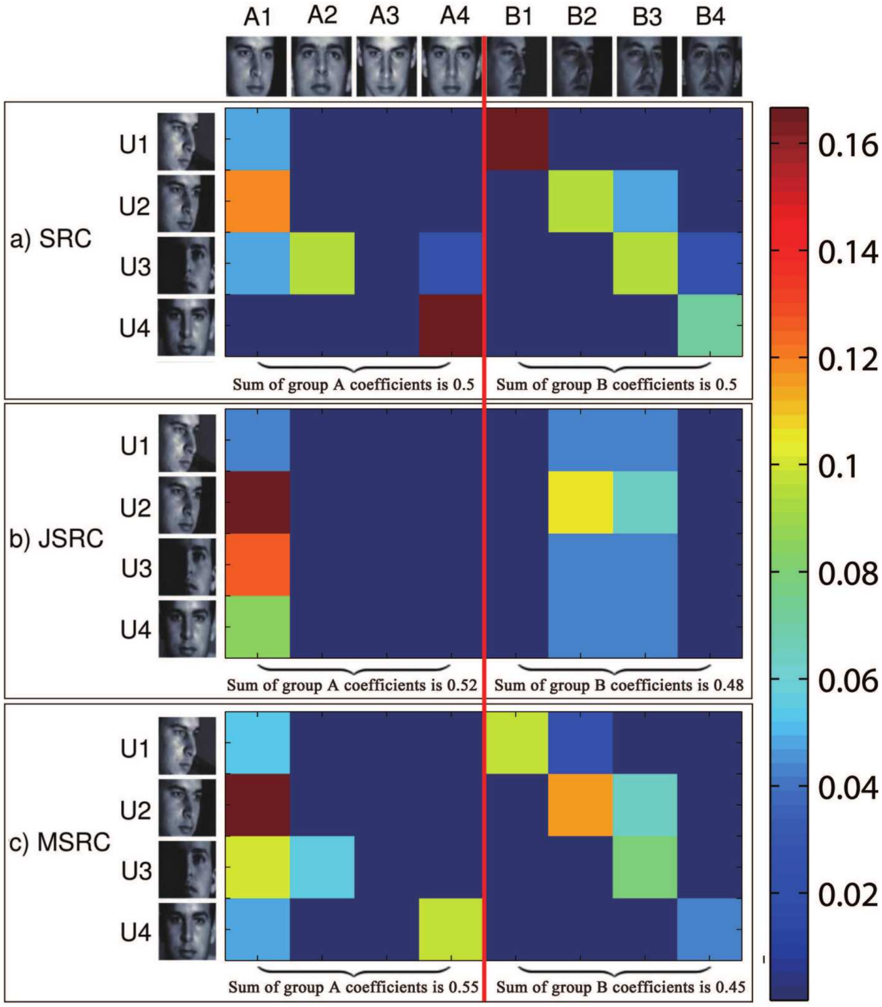

Figure 3: Coefficients of SRC[6], JSRC[14, 20] and MSRC for Figure 2. The coefficients in each set are positive and sum to one. In (a), SRC selects atoms based on similarity between each image. Thus, the coefficients are distributed across all classes. In (b), JSRC favours atoms based on similarity across all poses in the query set, thus there are only few columns (rows) being selected. In (c), coefficients are likely to be enhanced if they appear in both SRC and JSRC. Otherwise, they are likely to be suppressed if they appear only in one of the two methods. Thus, the total sum of coefficients of MSRC for group A is significant more than the total sum of coefficients for group B.

mented Lagrangian

$$\begin{aligned}\mathcal{L} \;=\;& \|\mathbf{V}\|_1 + \alpha\|\mathbf{Z}\|_1 + \mathsf{tr}[\mathbf{W}_1^T(\mathbf{AX}-\mathbf{V}-\mathbf{Y})] \\ & + \frac{\eta_1}{2}\|\mathbf{AX}-\mathbf{V}-\mathbf{Y}\|_2^2 + \beta\|\mathbf{T}\|_{2,1} \\ & + \mathsf{tr}[\mathbf{W}_2^T(\mathbf{X}-\mathbf{Z})] + \frac{\eta_2}{2}\|\mathbf{X}-\mathbf{Z}\|_F^2 \\ & + \mathsf{tr}[\mathbf{W}_3^T(\mathbf{X}-\mathbf{T})] + \frac{\eta_3}{2}\|\mathbf{X}-\mathbf{T}\|_F^2. \end{aligned} \tag{26}$$

Here, $\mathsf{tr}[\bullet]$ denotes the trace of a matrix and we omit the arguments $(\mathbf{X}, \mathbf{V}, \mathbf{Z}, \mathbf{T}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3)$ of the Lagrangian for notational simplicity. As with ADMM, we scale dual variables $\mathbf{U}_i = \mathbf{W}_i/\eta_i, i = 1, 2, 3$, to obtain a simpler form

$$\begin{aligned}\mathcal{L} \;=\;& \|\mathbf{V}\|_1 + \alpha\|\mathbf{Z}\|_1 + \beta\|\mathbf{T}\|_{2,1} \\ & + \frac{\eta_1}{2}\|\mathbf{AX}-\mathbf{V}-\mathbf{Y}+\mathbf{U}_1\|_F^2 \\ & + \frac{\eta_2}{2}\|\mathbf{X}-\mathbf{Z}+\mathbf{U}_2\|_F^2 \\ & + \frac{\eta_3}{2}\|\mathbf{X}-\mathbf{T}+\mathbf{U}_3\|_F^2 + \text{const.} \end{aligned} \tag{27}$$

where the constant is independent of the primal variables $\mathbf{X}, \mathbf{V}, \mathbf{Z}$.

Again, the updates for the variables are easily computed under the ADMM principle. For $\mathbf{X}$, we find the update from

$$\begin{aligned}\mathbf{X}^{k+1} \;=\;& \arg\min_{\mathbf{X}} \frac{\eta_1}{2}\|\mathbf{AX}-\mathbf{V}^k-\mathbf{Y}+\mathbf{U}_1^k\|_2^2 \\ & + \frac{\eta_2}{2}\|\mathbf{X}-\mathbf{Z}^k+\mathbf{U}_2^k\|_2^2 \\ & + \frac{\eta_3}{2}\|\mathbf{X}-\mathbf{T}^k+\mathbf{U}_3^k\|_2^2, \end{aligned} \tag{28}$$

which yields the exact solution

$$\mathbf{X}^{k+1} \;=\; \mathbf{H}^{-1}q, \tag{29}$$

where $\mathbf{H}^{-1} = (\eta_1\mathbf{A}^T\mathbf{A} + (\eta_2 + \eta_3)\mathbf{I})^{-1}$ can be computed once, and $q = (\eta_1\mathbf{A}^T(\mathbf{V}^k + \mathbf{Y} - \mathbf{U}_1^k) + \eta_2(\mathbf{Z}^k - \mathbf{U}_2^k) + \eta_3(\mathbf{T}^k - \mathbf{U}_3^k))$ is the update term.

As can be seen in (29), the update step of $\mathbf{X}$ is computationally expensive. Here, the matrix under inversion has dimensions $N \times N$ where $\mathbf{A} \in \mathbb{R}^{d\times N}$. In the case $d < N$, i.e., the feature dimension is less than the number of images in the gallery, such a direct matrix inversion can be inefficient. A much more efficient approach is to use Cholesky decomposition to achieve the goal. It is known from linear algebra that if $\mathbf{H}$ is a positive definite matrix then it admits the factorization $\mathbf{H} = \mathbf{L}\mathbf{L}^T$ and thus $\mathbf{H}^{-1}\mathbf{q}$ can be efficiently computed by solving $\mathbf{L}\mathbf{X}_1 = \mathbf{q}$ first, then $\mathbf{L}^T\mathbf{X} = \mathbf{X}_1$, which can be written as $\mathbf{X} = \mathbf{L}^T \setminus (\mathbf{L} \setminus \mathbf{q})$.

For variable $\mathbf{V}$, the update step solves

$$\mathbf{V}^{k+1} \quad = \quad \arg\min_{\mathbf{V}} \|\mathbf{V}\|_1 + \frac{\eta_1}{2}\|\mathbf{Q}^k - \mathbf{V}\|_2^2, \tag{30}$$

where $\mathbf{Q}^k = \mathbf{A}\mathbf{X}^{k+1} - \mathbf{Y} + \mathbf{U}_1^k$. Likewise, for $\mathbf{Z}$ the update step solves

$$\mathbf{Z}^{k+1} \quad = \quad \arg\min_{\mathbf{Z}} \alpha\|\mathbf{Z}\|_1 + \frac{\eta_2}{2}\|\mathbf{P}^k - \mathbf{Z}\|_2^2, \tag{31}$$

where $\mathbf{P}^k = \mathbf{X}^{k+1} + \mathbf{U}_2^k$.

As $\| \bullet \|_1$ is absolute value, the first terms in both (30) and (31) are not differentiable. However, we still can solve them directly. A soft-thresholding shrinkage operator can be used to find the solutions in element-wise. Therefore, the solutions for $\mathbf{V}$ and $\mathbf{Z}$ are defined as follows

$$\mathbf{V}^{k+1} \quad = \quad \mathsf{S}_{1/\eta_1}(\mathbf{Q}^k), \tag{32}$$

$$\mathbf{Z}^{k+1} \quad = \quad \mathsf{S}_{\alpha/\eta_2}(\mathbf{X}^{k+1} + \mathbf{U}_2^k), \tag{33}$$

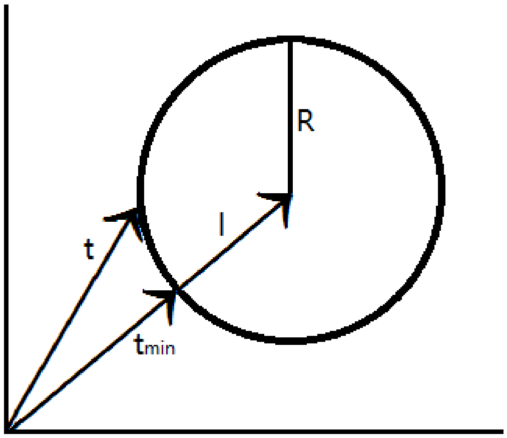

Figure 4: Illustration of Lemma 3.1. The circle (or sphere in high dimension space) which centered at $\mathbf{l}$ with radius R is the set of all feasible points for $\mathbf{t}$. When we constraint $\mathbf{t}$ with minimizes ℓ_2-norm, we will have the only solution t_{min}.

where this soft-thresholding shrinkage operator is defined as

$$\mathsf{S}_\tau(\mathbf{X}) = \{\mathbf{t} : t_i = \operatorname{sign}(X_{i,j}) \max(|x_{i,j}| - \tau, 0)\}. \tag{34}$$

For the last primal variable $\mathbf{T}$, the update steps are only slightly different

$$\mathbf{T}^{k+1} = \arg\min_{\mathbf{T}} \frac{\eta_3}{2}\|\mathbf{X}^{k+1} + \mathbf{U}_3^k - \mathbf{T}\|_F^2 + \beta\|\mathbf{T}\|_{2,1}. \tag{35}$$

To solve this problem, suppose that $\mathbf{t}_i$ and $\mathbf{l}_i$ are the ith row vectors of $\mathbf{T}$ and $\mathbf{X}^{k+1} + \mathbf{U}_3^k$ respectively, then we decompose the problem as

$$\mathbf{T}^{k+1} = \arg\min_{\mathbf{t}_i, i=1,\ldots} \sum_i \frac{\eta_3}{2}\|\mathbf{l}_i - \mathbf{t}_i\|_2^2 + \beta\|\mathbf{t}_i\|_2. \tag{36}$$

Thus, we can find each row of $\mathbf{T}$ separately by exploiting the following result

Lemma 3.1. *The solution of the optimization problem*

$$\min_{\mathbf{t}} \frac{\eta_3}{2}\|\mathbf{l} - \mathbf{t}\|_2^2 + \beta\|\mathbf{t}\|_2$$

is $\mathbf{t} = \kappa\mathbf{l}$*, where* $\kappa = \max(1 - \frac{\beta}{\eta_3\|\mathbf{l}\|_2}, 0)$ *if* $\|\mathbf{l}\|_2 > 0$ *and* $\kappa = 0$ *if* $\|\mathbf{l}\|_2 = 0$.

This result can be easily proved by a geometrical argument as shown in Figure 4. Indeed, suppose that $\mathbf{t}^*$ is the solution of the problem then we consider all feasible $\mathbf{t}$ such that $\|\mathbf{l}-\mathbf{t}\|_2 = \|\mathbf{l}-\mathbf{t}^*\|_2 = R$. Then, it is observed that the set of those feasible points is the sphere centred at $\mathbf{l}$ with radius R. Among all those feasible points, the solution must be the one that minimizes $\|\mathbf{t}\|_2$, which is the intersection of the sphere and the line from the origin to the centre of the sphere. Then it follows that the solution must be of the form $\mathbf{t} = \kappa\mathbf{l}$ with $1 \geq \kappa \geq 0$. Then straightforward manipulations easily lead to the result.

Finally, the updates for dual variables are

$$\mathbf{U}_1^{k+1} = \mathbf{U}_1^k + \mathbf{A}\mathbf{X}^{k+1} - \mathbf{V}^{k+1} - \mathbf{Y}, \tag{37}$$

$$\mathbf{U}_2^{k+1} = \mathbf{U}_2^k + \mathbf{X}^{k+1} - \mathbf{Z}^{k+1}, \tag{38}$$

$$\mathbf{U}_3^{k+1} = \mathbf{U}_3^k + \mathbf{X}^{k+1} - \mathbf{T}^{k+1}. \tag{39}$$

3.3. Stopping criterions and convergence

The original ADMM is designed for two primal variables, it solves

$$\begin{aligned} \min \quad & f(\mathbf{x}) + g(\mathbf{z}) \\ \text{s.t.} \quad & A\mathbf{x} + B\mathbf{z} = b. \end{aligned} \tag{40}$$

where both $f(\mathbf{x})$ and $g(\mathbf{z})$ are convex functions. However, there are three primal variables in our problem (25). Since the proposed method does not have the explicit form of the original ADMM framework, we now show that it can be easily converted to that standard form, and thus the proposed method naturally inherits the convergence property established in ADMM theory. Indeed, we rewrite the proposed formulation as follows

$$\begin{aligned} \min \mathbf{X}, \mathbf{Z}, \mathbf{T} & \quad f(\mathbf{X}) + g(\mathbf{Z}) + h(\mathbf{T}) \\ f(\mathbf{X}) &= \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_1, \\ g(\mathbf{Z}) &= \alpha\|\mathbf{Z}\|_1, \\ h(\mathbf{T}) &= \beta\|\mathbf{T}\|_{2,1}, \\ \text{s.t.} & \quad \mathbf{X} - \mathbf{Z} = \mathbf{0} \\ & \quad \mathbf{X} - \mathbf{T} = \mathbf{0}. \end{aligned} \tag{41}$$

We now reduce it to two variables by combining $g(\mathbf{Z})$ and $h(\mathbf{T})$ into a function of $\mathbf{Z}' = [\mathbf{Z};\ \mathbf{T}]$

$$k(\mathbf{Z}') = g(\mathbf{Z}) + h(\mathbf{T}). \tag{42}$$

As both g and h are convex and that $\mathbf{Z}$ and $\mathbf{T}$ are sub-blocks of $\mathbf{Z}'$, it follows that k is also convex in $\mathbf{Z}'$. Next, we combine two equality constraints as

$$\begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} \mathbf{X} - \begin{bmatrix} \mathbf{Z} \\ \mathbf{T} \end{bmatrix} = \mathbf{0}, \tag{43}$$

or equivalently $\mathbf{C}_x\mathbf{X} + \mathbf{C}_{z'}\mathbf{Z}' = \mathbf{0}$ where $\mathbf{C}_x = [\mathbf{I}; \mathbf{I}]$ and $\mathbf{C}_{z'} = -\mathbf{I}$. Thus, the proposed formulation can be expressed in the same form as the original ADMM as follows

$$\begin{aligned} \min_{\mathbf{X}, \mathbf{Z}'} & \quad f(\mathbf{X}) + k(\mathbf{Z}') \\ \text{s.t} & \quad \mathbf{C}_x\mathbf{X} + \mathbf{C}_{z'}\mathbf{Z}' = \mathbf{0} \end{aligned} \tag{44}$$

and thus it inherits all desirable properties of ADMM.

According to [27], the reasonable termination criterions for the proposed method are when the Frobenius norms of the residual vectors for primal and dual are sufficiently small,

$$\begin{aligned}
\mathbf{S}_1^k &= \eta_1(\mathbf{V}^{k+1} - \mathbf{V}^k) \leq \varepsilon^{S_1}, & (45)\\
\mathbf{S}_2^k &= \eta_2(\mathbf{Z}^{k+1} - \mathbf{Z}^k) \leq \varepsilon^{S_2}, & (46)\\
\mathbf{S}_3^k &= \eta_3(\mathbf{T}^{k+1} - \mathbf{T}^k) \leq \varepsilon^{S_3}, & (47)\\
\mathbf{R}_1^k &= \mathbf{A}\mathbf{X}^k - \mathbf{Y} - \mathbf{V}^k \leq \varepsilon^{R_1}, & (48)\\
\mathbf{R}_2^k &= \mathbf{X}^k - \mathbf{Z}^k \leq \varepsilon^{R_2}, & (49)\\
\mathbf{R}_3^k &= \mathbf{X}^k - \mathbf{T}^k \leq \varepsilon^{R_3}. & (50)
\end{aligned}$$

These tolerances can be chosen using an absolute and relative criterion, such as

$$\begin{aligned}
\varepsilon^{S_1} &= \sqrt{n}\varepsilon^{abs} + \varepsilon^{rel}\|\eta_1\mathbf{U}_1^k\|_F, & (51)\\
\varepsilon^{S_2} &= \sqrt{n}\varepsilon^{abs} + \varepsilon^{rel}\|\eta_2\mathbf{U}_2^k\|_F, & (52)\\
\varepsilon^{S_3} &= \sqrt{n}\varepsilon^{abs} + \varepsilon^{rel}\|\eta_3\mathbf{U}_3^k\|_F, & (53)\\
\varepsilon^{R_1} &= \sqrt{n}\varepsilon^{abs} + \varepsilon^{rel}\max\{\|\mathbf{A}\mathbf{X}^k - \mathbf{Y}\|_F, \|-\mathbf{V}^k\|_F\}, & (54)\\
\varepsilon^{R_2} &= \sqrt{n}\varepsilon^{abs} + \varepsilon^{rel}\max\{\|\mathbf{X}^k\|_F, \|-\mathbf{Z}^k\|_F\}, & (55)\\
\varepsilon^{R_3} &= \sqrt{n}\varepsilon^{abs} + \varepsilon^{rel}\max\{\|\mathbf{X}^k\|_F, \|-\mathbf{T}^k\|_F\}. & (56)
\end{aligned}$$

where $\varepsilon^{abs} > 0$ is for absolute tolerance and $\varepsilon^{rel} > 0$ is for the relative tolerance. The n indicates number of faces in the gallery set. The relative tolerance ε^{rel} might be choosen from 10^{-3} or 10^{-4} based on application practice [27]. In this paper, we choose 10^{-4} for both absolute and relative tolerances.

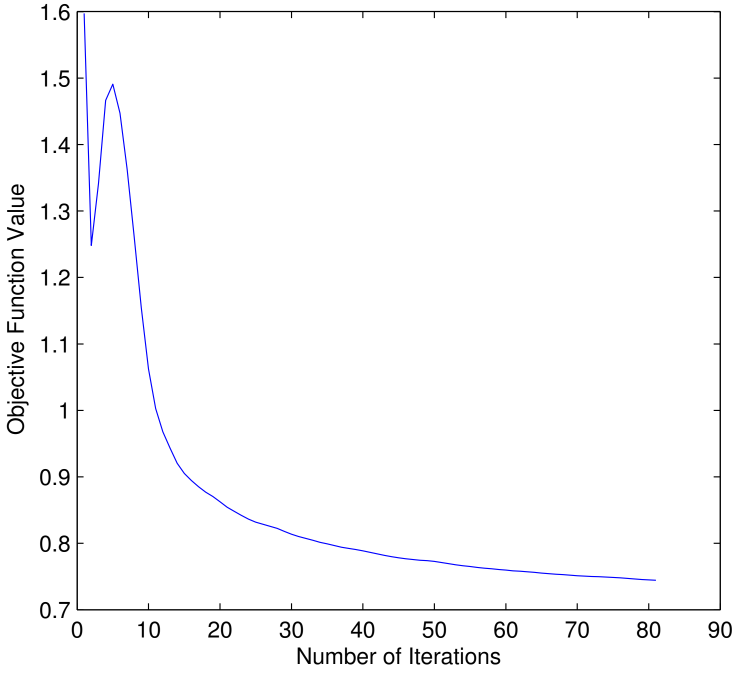

Figure 5: Convergence of proposed method.The objective function is converged after 30 iterations.

Although the ADMM framework can be slow to converge to high accuracy, when we setup the proper stopping criterions, the proposed ADMM-based method can converge to modest accuracy within a few tens of iterations (see Figure 5). This behavior makes our method can deal with large-scale problem in a short time. In next section, the experiment results show that this level of accuracy is sufficient enough for face recognition with multipe views.

3.4. Recognition and classification

Once the sparse representation matrix $\mathbf{X}$ is found for all views $\mathbf{Y}$, the classification is delivered by computing the fitness of query set with respect

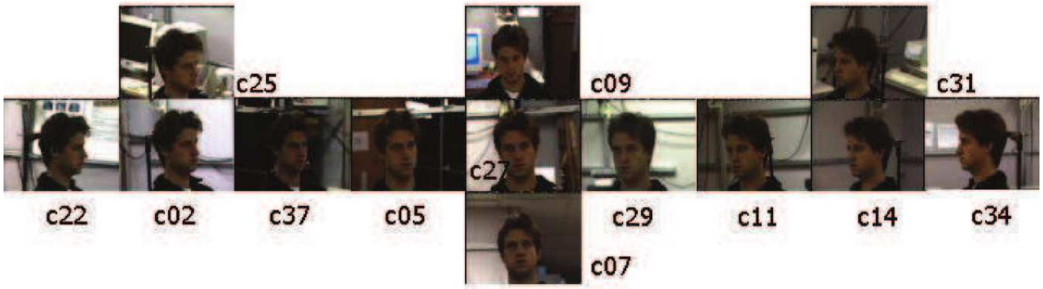

Figure 6: An example of the pose variations in CMU-PIE face database [31]. The pose variations are from 90° (left) to 0° (frontal) and then to -90° (right). They are separated by about 22.5° in horizontal. Pitch angles are involved for the other 4 pose variations

to the sparse solution. By following [15], there is only one decision made simultaneously on the class label for the whole query set based on $\mathbf{X}$ by combining the residuals for each image in the query set. Denote $\mathbf{A}^k$ is the subset of the gallery images corresponding faces of class k, and $\mathbf{X}^k$ is the corresponding coefficient subset for all query images. The fitness for class k is represented by the residual matrix $\mathbf{F}_k = \mathbf{Y} - \mathbf{A}^k\mathbf{X}^k$

$$\text{class}(\mathbf{Y}) = \arg\min_k \|\mathbf{F}_k\|_F^2. \tag{57}$$

The class label of $\mathbf{Y}$ is assigned to the class with minimum reconstruction error under Frobenius norm $\| \bullet \|_F$.

4. Experiments

In this section, we present extensive experiments on CMU-PIE [31], Yale B [32] and Multi-PIE [33] face databases. CMU-PIE is one of the most popular evaluation benchmarks in the face recognition literature. The CMU-PIE database consists of 41,368 images of 68 individuals. The face images are taken under 13 different poses, 43 illumination conditions, and with 4 different expressions for each people. Examples of CMU-PIE images are shown in Figure 6. The Yale B database contains 5760 images of 10 subjects each seen under 576 viewing conditions (9 poses with 64 illumination conditions). The Multi-PIE database contains 750,000 images of 337 people. These images were taken from 4 different seasons with 15 different poses under 19 illumination conditions. Since experiments are setting up for evaluating the multi-view images recognition, only images with neutral expressions under different illumination and different poses are used. Only basic preprocessings are performed before comparison, such as aligning and cropping face images, histogram equalization to make input data robust to lighting conditions, and PCA is used to resize the images to suitable working dimensions.

Whilst the main method for comparison is JDSRC outlined in [21], we also include the original SRC [6] and JSRC [14, 20]. In addition, some popular base line face recognition techniques are also evaluated, including principal component analysis (PCA) and linear discriminant analysis (LDA)[4]. Since SRC is originally used for single task, we follow [15] to use majority voting for SRC classification step and do the same for JSRC [20]. For notational convenience, we denote the mixed-norm sparse representation classification as MSRC. In this work, we follow the standard cross validation procedure in machine learning to select the regularization parameter λ for SRC and JSRC. This is achieved by further dividing the training set into a smaller training set and a validation set. For the proposed MSRC method, we also follow the same procedure, wherein all the training, validation, and test sets are exactly the same as those used for SRC and JSRC. The only minor differ-

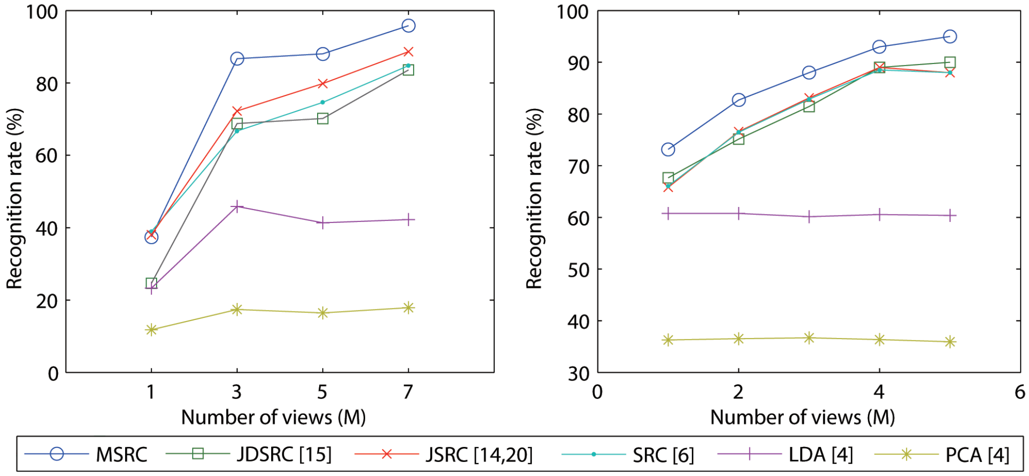

Figure 7: Recognition rate under different number of views with dimension d=64 for CMU-PIE (Left) and Yale B (Right)

ence is that MSRC has both the regularization parameter λ and the mixed norm parameter γ (ranged between 0 and 1). This means the computation slightly increases because the search is done on two dimensions. From the interpretation of the role of γ in controlling the pose variation and our intensive numerical studies, we suggest that the computational increase in cross validation due to γ might be reduced by a preliminary estimation of the pose variation. We notice that when there is large pose variation in the test set, a larger γ is preferred and vice versa. This suggests that if we use a reliable pose detection method such as [34], we may have a good estimate of the pose variation and hence a fixed γ can be set without sacrificing an increase in computation due to cross validation. However, we shall address this issue in a great depth in future works.

4.1. Face recognition with different number of poses

In this experiment, all methods are evaluated under different number of views. In order to show the performance of those methods under multiple face poses, we follow the experiment settings in [15] for CMU-PIE datasets. Images in the training set are selected based on a pose subset [0°, ±22.5°, ±45°, ±67.5°, ±90°]. Only one face image is selected for each subject with each pose in the training. In the testing set, **M** poses are selected to compose the query set for each subject. And we also use only one image for each pose. Since we randomly select from all 13 poses, the selected pose may not exist in the training set. Then, we use a similar settings for Yale B datasets. This makes our experiments more realistic and challenging.

In Figure 7, we compare the classification accuracy of the proposed method with others for both CMU-PIE (Left) and Yale B (Right). As can be seen, traditional subspace methods cannot reach satisfactory classification rates, but all SRC based methods can work well in multiple-views scenario. If there is just one testing image, none of methods can perform well. We note that all methods perform better in Yale B than CMU-PIE when there is only one training image. The reason for this is that the Yale B only has 10 subjects, which is much less than CMU-PIE. When more views are added in, the performance of SRC based methods is increased. Especially, our proposed MSRC reached a satisfactory rate when M=3, and it achieved 95.82% when we have 7 views in the test set. Clearly, this outperforms the closest competitor - JSRC by about 7% for CMU-PIE. It can be further improved by adding more number of views for Yale B. Furthermore, we note that both JSRC and MSRC have a similar recognition rate with M=1 on CMU-PIE. When the number of views increases, JSRC cannot achieve the same performance as MSRC.

According to [15], when the difference between different views becomes larger and larger, the assumption of JSRC that all views can be represented by the same set of atoms becomes more and more inaccurate. The images containing large pose variations will bring in inaccurate factors to query set. JSRC tries to find a solution across this poor query set. Thus, it is not able to find an optimal result. However, the proposed method has a degree of freedom to remove a few images, which have low correlation with other images. This makes MSRC find a more accurate representation than JSRC. Moreover, both SRC and JSRC are supposed to perform better than JDSRC. This might sound contradicting to dynamic atoms selection and what reported in [15]. But, a closer inspection reveals that the authors in [15] used greedy algorithms for solving the sparse problems, which is known to be inferior to the convex optimization algorithm used by this work, and perhaps that leads to a different result. Overall, MSRC gains the advantages from both dynamic atoms selection and superior convergence properties of specialized ADMM.

4.2. Face recognition under different dimensions

In this experiment, we investigate how performance depends on different feature dimensions. To do so, we reduce the original image to d = [32, 64, 128, 256] for CMU-PIE, which is effective for SRC based face recognition [6]. Following [15], we use the same training set from previous experiments and randomly select 5 views for the testing set. For Yale B, since the pose variation is not as large as CMU-PIE, we reduce the original image to d = [8, 16,

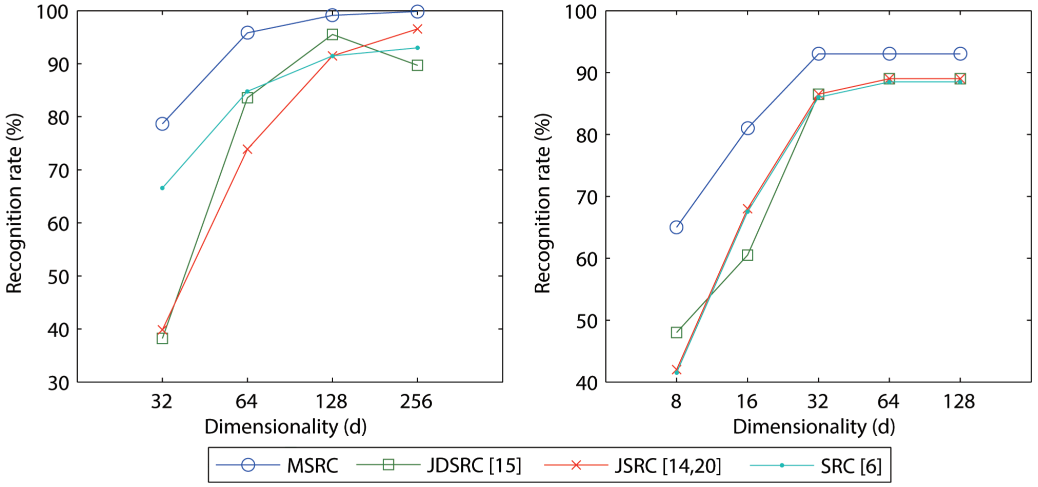

Figure 8: Recognition rate under different dimensionalities with number of views M=5 for CMU-PIE (Left) and $M = 4$ for Yale B (right)

32, 64, 128]. The comparison results are shown in Figure 8. As can be seen, MSRC achieves the highest recognition rates across all image dimensions for both CMU-PIE and Yale B. Both JSRC and SRC have competitive accuracy, though less than MSRC. The performance of all these three methods is superior to JDSRC in most image dimensions. When the data dimension $\geq$ 64 in CMU-PIE and 32 in Yale B, performance of all methods becomes saturated. However, JDSRC drops after $d = 128$ in CMU-PIE, this may be caused by the low accuracy of its greedy algorithm. In conclusion, MSRC is not sensitive to feature dimensions when dimension $\geq$ 64 for CMU-PIE or 32 for Yale B. This means that the face image with 64 feature dimension is a good choice for our MSRC. Also, it can achieve satisfactory performance with much lower computational complexity.

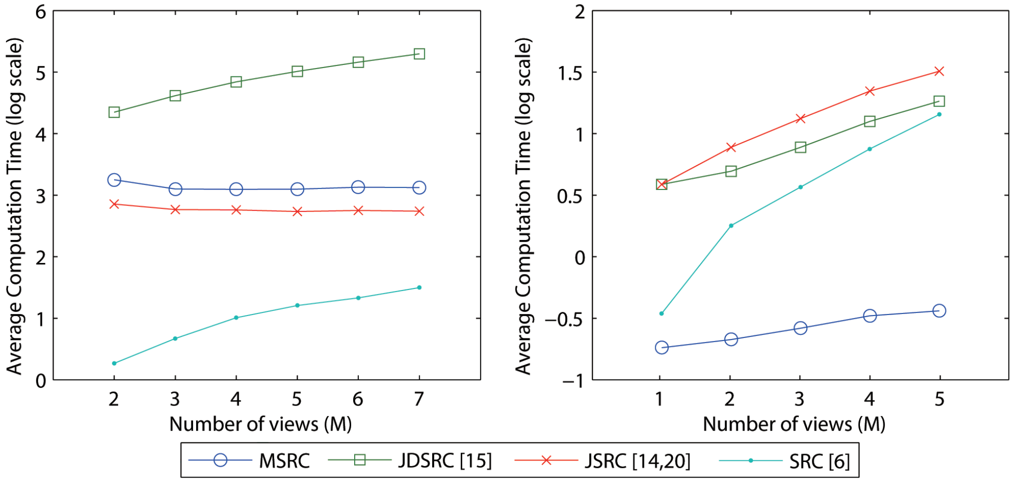

Figure 9: Computational comparison with dimension d=64 for CMU-PIE (Left) and Yale B (Right). The time axis is shown in log scale

4.3. Computational efficiency

In this section, we demonstrate how the specialized ADMM algorithm for robust sparse representation provides a computational advantage over other methods. We generate the training and testing sets based on random pose selection, and then reduce the dimension to $d = 64$. In Figure 9, we record the average computation time of completing this experiment in the log scale with 10 randomly selected testing sets for each number of views for both CMU-PIE and Yale B. The Yale B set has 10 subjects, which is much less than 64 subjects in CMU-PIE. This means the size of the problem in Yale B is much smaller than CMU-PIE. As can be seen in Figure 9, all methods take less time to complete on Yale B than CMU-PIE. Among these, our proposed method achieves the best time complexity. When the number of views increases, the scaled time rises, but the increased scaled time is minor compared with others. On CMU-PIE, SRC with majority voting performs best. This is caused by losing the ability to extract the shared information. When the size of problem increases, it takes advantages of less computation complexity. However, MSRC still achieved satisfactory performance, especially when compared with JDSRC. In addition, the completion time of MSRC remains almost unchanged when the number of views increases in the experiment on CMU-PIE. Overall, the proposed MSRC achieves adequate performance for both CMU-PIE and Yale B, and its computation complexity is insensitive to the increase of number of views.

4.4. Face recognition against unseen pose

We next examine the effectiveness of recognition against unseen pose. In order to achieve this goal, a pose appearing in the testing set may not appear in the training set. Therefore, we randomly select images from all poses to create the training set. Randomly selected images for the testing sets are from three different groups:

1. images with the same poses observed in the training set;
2. images with completely different poses from the training set;
3. images selected randomly from both seen and unseen poses of the training set.

This setting allows us to investigate the effect of unseen poses in our method. Experiment results are reported in Table 1 for CMU-PIE and 2 for Yale B. As shown in Table 1, all methods perform well with the same poses from the training set except PCA and LDA. The reason for poor performance of traditional subspace methods might attribute to the fact that there is only 1 image for each poses each subject in the training. However, this would not affect SRC-based methods. When unseen poses are present in the testing set, the performance of all methods drop as shown in “Mixed” column of Table 1. In this situation, MSRC still remains at 95.82% (only 3% decrease). When images in the testing set completely come from unseen poses, most of the methods cannot achieve satisfactory performance except MSRC, which can still reach 73.88%. Table 2 shows a similar story: SRC-based methods perform better than general subspace techniques and our proposed method outperforms others with at least 4% increase for unseen pose cases. On the other hand, the experiment results on Yale B are similar to those on CMU-PIE. The proposed method outperforms all other methods for all three cases. Overall, the proposed MSRC is much more insensitive to unseen poses. This makes MSRC more suitable to real-world applications.

view(s)	Unseen	Mixed	Same
PCA[4]	12.59%	17.91%	16.67%
LDA[4]	24.40%	42.26%	41.29%
SRC[6]	55.37%	84.78%	93.28%
JSRC[14, 20]	58.06%	88.66%	95.52%
JDSRC[15]	48.51%	83.58%	92.99%
MSRC	**73.88%**	**95.82%**	**98.21%**

Table 1: Face recognition against unseen pose for CMU-PIE

view(s)	Unseen	Mixed	Same
PCA[4]	36.20%	36.10%	35.50%
LDA[4]	52.70%	56.90%	52.70%
SRC[6]	84.50%	88.50%	89.50%
JSRC[14, 20]	86.00%	89.00%	89.00%
JDSRC[15]	84.00%	89.00%	89.00%
MSRC	**90.00%**	**93.00%**	**94.50%**

Table 2: Face recognition against unseen pose for Yale B

4.5. Face recognition under pose difference

In this section, we investigate how the performance of the proposed method and other methods under large pose variations in multi-view face recognition scheme. Since Yale B only has 9 poses, which is not sufficient to perform this experiment, we only use CMU-PIE in this experiment. It is organised as follows. Face images of all 13 poses from Figure 6 are used for the training set, but only one image is randomly selected for each pose for each subject. We then create 4 different pose groups: [0°, ±22.5°], [0°, ±45°], [0°, ±67.5°] and [0°, ±90°]. There are 3 images (one image for each pose) in each group. The testing sets are generated by randomly selecting from these 4 pose groups. As can be seen in Table 3, traditional subspace methods perform poorly, this is consistent with previous experimental results. However, all SRC-based methods achieve satisfactory performance. We observe that when the pose difference increases, the performance of SRC decreases. JSRC performs slightly better than SRC across all pose variations. Also, JDSRC outperforms both JSRC and SRC. Since JDSRC uses dynamic selected atoms, it would not select the same set of atoms for all views as JSRC. This makes JDSRC more suitable to multiple-views scenarios. Overall, our proposed method reaches the highest recognition rates under each testing pose group. It achieves 95.52% for [0°, ±22.5°] group with 8% improvement compared to its best competitor.

	22.5°	45°	67.5°	90°
PCA[4]	24.43%	24.68%	24.38%	23.43%
LDA[4]	59.75%	60.00%	58.11%	54.13%
SRC[6]	82.24%	80.30%	80.15%	69.25%
JSRC[14, 20]	84.48%	84.63%	83.28%	74.18%
JDSRC[15]	87.16%	84.63%	87.16%	78.21%
MSRC	**95.52%**	**94.33%**	**93.73%**	**88.96%**

Table 3: Face recognition against large pose variations

4.6. Face recognition under in different scales

To examine how the compared methods scale with a large number of subjects, we use the Multi-PIE dataset to perform two sets of experiment. To make our experiments more realistic and challenging, we mix all images from 4 different seasons altogether for 337 subjects. Images in each training set are selected based on following poses [0°, ±30°, ±60°, ±90°]. Three face images are selected for each subject with each pose in the training. For testing, M poses are randomly selected from all poses to compose the query set for each subject ($M = 5$ and $M = 7$). We also use three face images for each pose in testing set. Four pairs of training and testing sets are created from randomly selected subjects. These pairs of dataset correspond to 64, 136, 204, 272 subjects.

The left subplot of Figure (10) shows that the performance of SRC, JSRC and JDSRC deteriorates significantly when the number of subjects increases. However, the proposed MSRC is more robust against this large number of subjects. From the right subplot of Figure (10), we notice that all methods are benefited from an increasing in the number of views. Their performance are observed to improve overall, there is less sharp drop in recognition accuracy when number of subjects increases. Due to the flexible atom selection, both MSRC and JDSRC outperform SRC and JSRC for $M = 5$ and $M = 7$. However, the lack of guaranteed convergence of JDSRC makes it hard to find a robust and accurate solution. In general, the proposed method achieves a robust performance against different scales of datasets because it has an advantage of a dynamic atom selection and fast convergence.

5. Conclusion

In this paper, we have proposed a mixed-norm sparse representation classification, which has been demonstrated to outperform rivals. This is due to the advantage of exploiting the inter-correlation among the multiple face images in the query, the flexibility of atom selection and the robustness brought by an ℓ_1-loss function. Furthermore, this MSRC is built on the powerful ADMM framework, which results in a very simple, yet provably convergent, algorithm, where further improvement in both performance and computation can be made. We have demonstrated the power of the ADMM framework in deriving numerical algorithm to solve the proposed formulation. We also

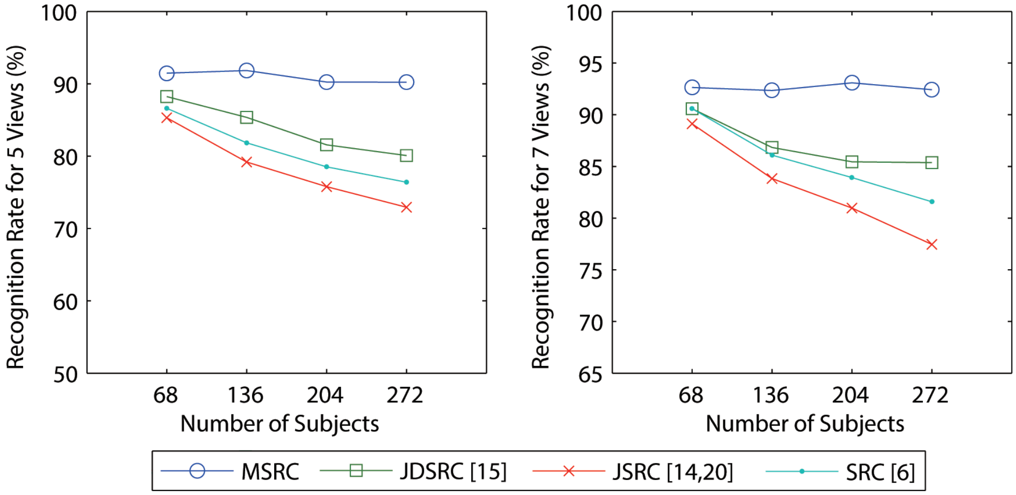

Figure 10: Face recognition against different scales. Number of subjects increases from 68 to 272 with different number of views M=5 (Left) and M=7 (Right).

have extensively studied and compared our MSRC with other methods on the CMU-PIE and Yale B datasets. The results indeed show the superior performance of the proposed methods under different number of views, various dimensionality, view differences, and computational time and scalability.

References

[1] T. S. Huang, G. Hua, M.-H. Yang, E. Learned-Miller, Y. Ma, M. Turk, D. J. Kriegman, Introduction to the special section on real-world face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2011) 1921–1924.

[2] W. Zhao, R. Chellappa, P. J. Phillips, A. Rosenfeld, Face recognition: A literature survey, Acm Computing Surveys (CSUR) 35 (2003) 399–458.

[3] X. Zhang, Y. Gao, Face recognition across pose: A review, Pattern Recognition 42 (2009) 2876–2896.

[4] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 711–720.

[5] X. Niyogi, Locality preserving projections, in: Proceedings of the Conference on Neural Information Processing Systems, volume 16, The MIT Press, 2004, p. 153.

[6] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 210–227.

[7] R. Duda, P. Hart, D. Stork, Pattern Classification, John Wiley & Sons, 2001.

[8] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, D. Kriegman, Clustering appearances of objects under varying illumination conditions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 1, IEEE, 2003, pp. I–11.

[9] R. He, W.-S. Zheng, B.-G. Hu, X.-W. Kong, A regularized correntropy framework for robust pattern recognition, Neural Computation 23 (2011) 2074–2100.

[10] M. Yang, L. Zhang, J. Yang, D. Zhang, Robust sparse coding for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 625–632.

[11] X. Zhang, D. Pham, W. Liu, S. Venkatesh, Optimal metric selection for improved multi-pose face recognition with group information, in: Proceedings of International Conference on Pattern Recognition (ICPR), 2012, pp. 1675–1678.

[12] E. Kokiopoulou, P. Frossard, Graph-based classification of multiple observation sets, Pattern Recognition 43 (2010) 3988–3997.

[13] H. Cevikalp, B. Triggs, Face recognition based on image sets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 2567–2573.

[14] A. Rakotomamonjy, Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms, Signal processing 91 (2011) 1505–1526.

[15] H. Zhang, N. M. Nasrabadi, Y. Zhang, T. S. Huang, Joint dynamic sparse representation for multi-view face recognition, Pattern Recognition 45 (2012) 1290–1298.

[16] A. Hadid, M. Pietikäinen, S. Z. Li, Learning personal specific facial dynamics for face recognition from videos, in: Analysis and Modeling of Faces and Gestures, Springer, 2007, pp. 1–15.

[17] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: Proceedings of Advances in neural information processing systems, 2004, pp. 321–328.

[18] L. Zini, N. Noceti, G. Fusco, F. Odone, Structured multi-class feature

selection with an application to face recognition, Pattern Recognition Letters (2014).

[19] T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns, in: Computer vision-eccv 2004, Springer, 2004, pp. 469–481.

[20] J. A. Tropp, A. C. Gilbert, M. J. Strauss, Algorithms for simultaneous sparse approximation. part i: Greedy pursuit, Signal Processing 86 (2006) 572–588.

[21] Q. Shi, A. Eriksson, A. van den Hengel, C. Shen, Is face recognition really a compressive sensing problem?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 553–560.

[22] E. J. Candes, J. K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications on pure and applied mathematics 59 (2006) 1207–1223.

[23] E. J. Candes, T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, IEEE Transactions on Information Theory 52 (2006) 5406–5425.

[24] N. Nguyen, N. Nasrabadi, T. Tran, Robust lasso with missing and grossly corrupted observations, preprint (2011).

[25] J. Wright, A. Ganesh, A. Yang, Z. Zhou, Y. Ma, Sparsity and robustness in face recognition, arXiv preprint arXiv:1111.1014 (2011).

[26] D. Pham, S. Venkatesh, Improved image recovery from compressed data contaminated with impulsive noise, IEEE Transactions on Image Processing 21 (2012) 397–405.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning 3 (2011) 1–122.

[28] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Transactions on Information Theory 52 (2006) 489–509.

[29] D. Donoho, Compressed sensing, IEEE Transactions on Information Theory 52 (2006) 1289–1306.

[30] M. F. Duarte, V. Cevher, R. G. Baraniuk, Model-based compressive sensing for signal ensembles, in: Proceedings of 47th Annual Allerton Conference on Communication, Control, and Computing., IEEE, 2009, pp. 244–250.

[31] T. Sim, S. Baker, M. Bsat, The cmu pose, illumination, and expression (pie) database, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 2002, pp. 46–51.

[32] A. S. Georghiades, P. N. Belhumeur, D. Kriegman, From few to many: Illumination cone models for face recognition under variable lighting and pose, Pattern Analysis and Machine Intelligence, IEEE Transactions on 23 (2001) 643–660.

[33] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image and Vision Computing 28 (2010) 807–813.

[34] E. Murphy Chutorian, M. M. Trivedi, Head pose estimation in computer vision: A survey, Pattern Analysis and Machine Intelligence, IEEE Transactions on 31 (2009) 607–626.