Sparse representation for face images

by

Xin Zhang

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Deakin University
May 2016
I am the author of the thesis entitled

Sparse representation for face images.

submitted for the degree of Doctor of Philosophy (Information Technology)

This thesis may be made available for consultation, loan and limited copying in accordance with the Copyright Act 1968.

'I certify that I am the student named below and that the information provided in the form is correct'

Full Name: XIN ZHANG
(Please Print)

Signed: .................................................................

Date: 27/09/2016
DEAKIN UNIVERSITY
CANDIDATE DECLARATION

I certify the following about the thesis entitled (10 word maximum)

Sparse representation for face images

submitted for the degree of Doctor of Philosophy (Information Technology)

a. I am the creator of all or part of the whole work(s) (including content and layout) and that where reference is made to the work of others, due acknowledgment is given.

b. The work(s) are not in any way a violation or infringement of any copyright, trademark, patent, or other rights whatsoever of any person.

c. That if the work(s) have been commissioned, sponsored or supported by any organisation, I have fulfilled all of the obligations required by such contract or agreement.

I also certify that any material in the thesis which has been accepted for a degree or diploma by any university or institution is identified in the text.

'I certify that I am the student named below and that the information provided in the form is correct'

Full Name: XIN ZHANG
(Please Print)

Signed: Signature Redacted by Library

Date: 27/09/2016

Deakin University CRICOS Provider Code 00113B
Published Work

Part of this thesis and some related work has been published in referred conference and journal papers or documented elsewhere. The list of these publications is provided below:


Abstract

One of the most challenging research problems in pattern recognition is to recognise a human face from images of multiple poses captured by a surveillance system. This thesis addresses this challenge by using sparse representation as a guiding principle. This principle has been found useful in face recognition under both supervised and unsupervised settings. In supervised learning, the well-known sparse representation classification method seeks a balance between nearest neighbour and nearest subspace to solve the face recognition problem. It finds the representation for an image as a linear combination of a small number of atoms in a dictionary, most of which are from the correct class. In unsupervised learning, the popular sparse subspace clustering method also finds the affinity matrix using a similar formulation, except with an extra equality constraint on the diagonal entries. The affinity matrix is then further analysed by spectral clustering to yield final result.

This thesis extends the above work to the case where face images are present under different poses due to the deployment of multiple cameras in many real-world computer vision applications. We propose a novel framework that generalises existing sparse representation-based methods in order to exploit the sharing information which is believed to exist between images of different poses. In this framework, we first cluster images belonging to several unknown subjects in the video sequence, then use these grouped images to perform the recognition.
To achieve this goal, we make contributions in the following four stages:

- In the first stage, a novel method is introduced for multi-view face recognition. A mixed norm is used to regularise the sparse representation process. By using this mixed norm, a trade-off between the $\ell_1$ norm and $\ell_{2,1}$ norm is achieved to optimally represent a given image with a group of highly correlated face images. This overcomes the large pose variations and missing pose issue in the existing multi-view face recognition literature.

- In the second stage, we further improve the performance of existing sparse representation classification methods by introducing an $\ell_p$ norm for classification, which is demonstrated to suppress outliers considerably.

- In the third stage, we achieve a better affinity matrix for multi-pose images by introducing the mixed norm to the regularisation step in sparse subspace clustering. Consequently, it promotes more optimal individual and group level sparsity, leading to a better clustering solution. A majority voting mechanism is then applied to the clustering result for each view.

- In the last stage, we propose a multi-view clustering method by searching for a unified latent structure for low frame rate cameras. This novel method integrates complementary information between views and obtains a global affinity matrix which best represents the relationship between clusters. This helps us to achieve a better performance on multi-view face images, especially when obtained from cameras operating at low frame rates.

We demonstrate that all the proposed methods outperform other state-of-the-art algorithms on CMU-PIE, Yale B and Multi-PIE databases under various settings. Moreover, we also provide theoretical insights of the proposed methods and discuss future research directions.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>JDSRC</td>
<td>Joint Dynamic Sparse Representation Classification</td>
</tr>
<tr>
<td>JSRC</td>
<td>Joint Sparse Representation Classification</td>
</tr>
<tr>
<td>GSC</td>
<td>Group Sparse Classification</td>
</tr>
<tr>
<td>MSSC</td>
<td>Multi-view Sparse Subspace Clustering</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>LPP</td>
<td>Locality preserving projections</td>
</tr>
<tr>
<td>LRR</td>
<td>Low Rank Representation</td>
</tr>
<tr>
<td>LRSC</td>
<td>Low Rank Subspace Clustering</td>
</tr>
<tr>
<td>LSA</td>
<td>Local Subspace Affinity</td>
</tr>
<tr>
<td>MSRC</td>
<td>Mixed-norm Sparse Representation Classification</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>SCC</td>
<td>Spectral Curvature Clustering</td>
</tr>
<tr>
<td>SRC</td>
<td>Sparse Representation Classification</td>
</tr>
<tr>
<td>SSC</td>
<td>Sparse Subspace Clustering</td>
</tr>
</tbody>
</table>
# Contents

1 Introduction ........................................ 1
   1.1 Aims and Approach .................................. 3
   1.2 Contributions ...................................... 5
       1.2.1 Novel methods ................................ 5
       1.2.2 Theoretical insights ......................... 6
       1.2.3 Experimental data and settings ............. 7
   1.3 Thesis structure ................................... 9

2 Background ......................................... 10
   2.1 General face recognition .......................... 10
   2.2 Multi-view face recognition ....................... 13
   2.3 Multi-view face clustering ....................... 18
   2.4 Related work ...................................... 20
       2.4.1 Sparse representation classification ........ 20
       2.4.2 Joint sparse representation classification .... 22
       2.4.3 Joint dynamic sparse representation classification .. 24
       2.4.4 Sparse subspace clustering ................... 26
   2.5 Alternating direction method of multipliers ....... 28
   2.6 Summary ........................................... 31

3 Mixed-norm sparse representation for multi-view face recognition 32
3.1 Multi-view mixed norm robust sparse representation ........ 36
  3.1.1 Model formulation ........................................ 36
  3.1.2 Algorithm .................................................. 42
3.2 Experiments ..................................................... 51
  3.2.1 Face recognition with different number of poses .......... 52
  3.2.2 Face recognition under different dimensions .......... 54
  3.2.3 Face recognition against unseen pose .................. 57
  3.2.4 Face recognition under pose difference ............. 58
  3.2.5 Face recognition with a large number of subjects ...... 59
  3.2.6 Computational efficiency ................................. 60
3.3 Summary .......................................................... 65

4 Optimal metric selection for sparse representation based classification ........................................... 66
  4.1 Group sparse classification ................................... 68
  4.2 Optimal metric selection ..................................... 69
  4.3 Experiments ..................................................... 73
    4.3.1 lp norm with Group Sparse Classification ............. 73
    4.3.2 Improved MSRC with lp norm .............................. 78
  4.4 Summary .......................................................... 80

5 Mixed-norm regularised face image clustering .................. 83
  5.1 Mixed-norm sparse subspace clustering ...................... 86
    5.1.1 Model formulation ........................................ 87
    5.1.2 Algorithm .................................................. 88
  5.2 Experiments ..................................................... 95
    5.2.1 Single-view clustering ................................. 95
    5.2.2 Multi-view clustering ................................. 97
  5.3 Summary .......................................................... 100
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Four sparsity models for the coefficient matrix $X$</td>
<td>35</td>
</tr>
<tr>
<td>3.2</td>
<td>Illustration of the spatial pose distribution</td>
<td>40</td>
</tr>
<tr>
<td>3.3</td>
<td>Coefficients of SRC, JSRC and MSRC</td>
<td>43</td>
</tr>
<tr>
<td>3.4</td>
<td>Illustration of Lemma 3.1.1</td>
<td>46</td>
</tr>
<tr>
<td>3.5</td>
<td>Convergence of mixed-norm SRC</td>
<td>49</td>
</tr>
<tr>
<td>3.6</td>
<td>Pose variations in CMU-PIE face database</td>
<td>50</td>
</tr>
<tr>
<td>3.7</td>
<td>Recognition rate under different number of views for CMU-PIE</td>
<td>52</td>
</tr>
<tr>
<td>3.8</td>
<td>Recognition rate under different number of views for Yale B</td>
<td>53</td>
</tr>
<tr>
<td>3.9</td>
<td>Recognition rate under different dimensionality for CMU-PIE</td>
<td>55</td>
</tr>
<tr>
<td>3.10</td>
<td>Recognition rate under different dimensionality for Yale B</td>
<td>56</td>
</tr>
<tr>
<td>3.11</td>
<td>Face recognition against different scales for multi-PIE</td>
<td>61</td>
</tr>
<tr>
<td>3.12</td>
<td>Face recognition against different scales for multi-PIE</td>
<td>62</td>
</tr>
<tr>
<td>3.13</td>
<td>Computational comparison with dimension $d=64$ for CMU-PIE</td>
<td>63</td>
</tr>
<tr>
<td>3.14</td>
<td>Computational comparison with dimension $d=64$ for Yale B</td>
<td>64</td>
</tr>
<tr>
<td>4.1</td>
<td>Residual values of top candidates in SRC</td>
<td>71</td>
</tr>
<tr>
<td>4.2</td>
<td>Why does the $\ell_p$ norm suppress outliers?</td>
<td>72</td>
</tr>
<tr>
<td>4.3</td>
<td>Recognition rates with different $\ell_p$ norm</td>
<td>74</td>
</tr>
<tr>
<td>4.4</td>
<td>Dimension reduction based on PCA</td>
<td>76</td>
</tr>
<tr>
<td>4.5</td>
<td>With various number of groups</td>
<td>77</td>
</tr>
<tr>
<td>4.6</td>
<td>Multi-View ($M = 5$) with various number of subjects</td>
<td>80</td>
</tr>
<tr>
<td>4.7</td>
<td>Multi-View ($M = 7$) with various number of subjects</td>
<td>81</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Face recognition against unseen pose for CMU-PIE ........... 58
3.2 Face recognition against unseen pose for Yale B ............... 58
3.3 Face recognition against large pose variations ............... 59

4.1 Classification performance comparison for $\ell_p$ norm ........ 79

5.1 Clustering error on the Yale B data set ....................... 96

6.1 Clustering performance (%) under different number of images. . . 118
List of Algorithms

5.1 Solving Eq. (5.8) for mixed-norm sparse subspace clustering. . . . 95
6.1 Solving Eq. (6.5) for multi-view subspace representation. . . . . 109
Chapter 1

Introduction

In today’s world, the need to identify individuals is becoming increasingly important, in particular for the information security and criminal investigation. Due to the advancements in computing capability over the past few decades, new technology becomes available to verify or identify the human subject. This technology is called biometrics which is used to recognize the identity of a person based on some physiological characteristics, such as fingerprints, irises or face. One of the most efficient biometrics is face recognition. There are two main advantages of face recognition compared with other biometrics. First, face recognition can be easily integrated into existing surveillance systems. Second, face recognition is non-intrusive, the user cooperation is not required. This property does not just bring the comfortability to its user, and it also enables some large-scale applications for public needs, such as combating passport fraud, identifying missing children, supporting law enforcement, etc.

Since the benefits of face recognition are significant, some face recognition research works have been demonstrated a satisfactory performance, such as Bartlett et al. (2002); He et al. (2005); Cai et al. (2007a,b); Pham and Venkatesh (2008). However, these methods are limited to controlled environments, which do not
have the challenges seen in the real world. There are a few issues in those works. The first issue is the need to determine a subject’s distinctive features under large illumination variations. In the real world, human faces are captured in various lighting conditions, and this will bring significant degree of noises to the face images. The second issue arises when the appearance of the human face changes due to facial expressions. This change will distort the face image and make the recognition harder to find the correct solution. Occlusion of human face is the third issue which needs to be resolved. In reality, people often wear sunglasses or a scarf, and these accessories usually occlude parts of the human face. The information under the occlusion is missing when a face is captured by ordinary cameras. In the end, pose variations also bring difficulties to the face recognition process. When the face images of one subject are captured in the different views, some information is missing during the facial pose transformation.

In recent years, a large number of research works have been done to address these issues. In Han et al. (2013), they provide a comparative study on illumination preprocessing in face recognition. Those preprocessing approaches that can effectively eliminate most of the effects of illumination and preserves enough information that is required by recognition. To overcome the facial expression problem, a method that applies a spatial representation of local binary pattern on face images was proposed in Khan et al. (2013). Researchers in Wright et al. (2009) investigated robust face recognition based on sparse representation. They pointed out that their sparse representation-based classification (SRC) framework can work effectively against occlusion and corruption. Although these state-of-the-art face recognition techniques show fairly high recognition rates under uncontrolled lighting, facial expressions, occlusions, etc., they are only able to handle small pose variations (e.g. 15° rotation) Tan et al. (2006).

Due to the observation that the general face recognition approaches are sensitive to pose variations, many works have been proposed to handle pose variations
1.1 Aims and Approach

explicitly by using 3D data. However, these 3D data (depth information) are captured by some external devices. In practice, these devices are not available in many situations and the capturing process may be intrusive. In addition, as the growth of the deployment of surveillance systems, tens of thousands of face images are generated each second. The human subject captured in these images are presented in different views. Each view usually presents a face pose. Therefore, it is an open problem of how to fill the gap between controlled environments and the pose variations with overwhelmed images for face recognition.

There are two major challenges for this task. One such a challenge is recognising a person from a video sequence. Under this situation, the face images of one subject usually come with different views (poses). We classify subjects based on multiple images as multi-view problems. Since all these images are captured from the same subject, it is natural to believe that some shared information or connection exists in these images. Investigating this shared information is becoming both important and necessary. The other challenge is pre processing the face images obtained from millions of frames from surveillance systems. When multiple unknown human subjects are captured in a video sequence, it is crucial to label each face shown in each frame with respect to these unknown humans. Especially, when multiple cameras work together to create a correspondent scene, investigating the common information across images captured by these cameras becomes more and more important.

1.1 Aims and Approach

This thesis aims at addressing the above challenges. On one hand, we cast the process of recognising a person from a video sequence as a problem that finds a person’s identity from a given dictionary with a set of images. These images all belong to the person and are obtained from the video sequence. We call this type
1.1 Aims and Approach

of problem as multi-view face recognition. On the other hand, we treat the process of grouping each face shown in every frame as the face clustering problem. And when the corresponding data available, we recognise this type of problem as face clustering with multi-view images. Since most of existing methods have focused on the single-view problem and the multi-view data usually consider having more information, the objectives of this thesis can be stated as follows:

- Design efficient techniques for multi-view face recognition;
- Develop efficient face clustering methods which can cluster subjects from unlabelled multi-view images.

To achieve these objects, we further divide the main work into several stages:

1. Introducing a mixed-norm method to combine the advantages of SRC and JSRC. An efficient face recognition technique against large pose variation can be proposed based on this mixed norm;
2. Analyzing the $\ell_2$ norm used in the classification step of the original SRC framework;
3. Developing a new efficient sparse subspace clustering for face image with the mixed norm for normal videos.
4. Investigating the complementary information obtained from multi-view images to improve the performance of face clustering for low frame rate videos.

The first two stages are focused on multi-view face recognition. We propose a novel recognition method from a set of correlated face images and improve its performance by analysing the $\ell_2$ norm used in SRC. The last two stages deliver two multi-view face clustering methods for high- and low-frame rate cameras respectively.
1.2 Contributions

In this thesis, all mentioned objectives are achieved, resulting in several contributions to the field. We describe some of our significant contributions to methodological, theoretical and experimental perspectives.

1.2.1 Novel methods

All methods proposed in this work are novel and significant. They either achieve outstanding performance when comparing with the state of the arts for face recognition or provide new theoretical insights. The following methods are proposed in this thesis.

1. Mixed-norm sparse representation for multi-view face recognition extracts both shared and local information among views to improve face recognition performance. A mixed-norm regulariser is introduced to construct this new model that utilises the shared and local information. We demonstrate that this formulation outperforms rivals due to the advantage of exploiting the inter-correlation among the multi-view face images and the flexibility of atom selection.

2. Optimal metric selection for improved multi-pose face recognition with Group Information addresses the limitation of sparse representation based classification. We use an $\ell_p$ norm for the metric norm of the residual vectors to obtain better solutions. We proposed a new framework for sparse representation methods via utilising the $\ell_p$ norm. The experiments show that this framework can significantly improve the face recognition performance.

3. Mixed-norm sparse subspace clustering is a novel clustering method which is regularised by a mixed norm. This mixed norm gives us the flexibility in modelling the actual properties of the underlying data. The experimental
1.2 Contributions

results on the benchmark data sets: Yale B and Multi-PIE show a great success of this mixed-norm regulariser.

4. Multi-view face clustering for low frame videos can efficiently use the complementary information obtained from other views when the number of images per cluster is relatively small. The subspace representation for each view is constructed first. At the same time, a global subspace representation is developed based on the complementary information. The power of the proposed method is shown by comparing with the state-of-the-art clustering method on Multi-PIE face data set.

1.2.2 Theoretical insights

Theoretical insights are significant contributions as they inspire future research. This thesis makes three main theoretical insights to the field of computer vision: (1) the development of the mixed-norm regularisation, (2) the use of the $\ell_p$-norm in the SRC framework, (3) the latent structure in multi-view data. These contributions and their significance are detailed as follows.

In terms of the multi-view recognition problem, we provide a strong analytical evidence on the advantages of the mixed-norm approach compared to others. As we know the essential point of dealing with multi-view data is how to exploit the shared information. We argue that the local information also needs to be considered in this process. Furthermore, we have given an explicit formulation to capture local and shared information simultaneously. In this mixed-norm formulation, we combine the $\ell_1$ and $\ell_{2,1}$ norms with a balancing parameter. We explain how this formulation addresses the issue of considering the local information when extracting the shared information.

We show that the performance of the current sparse representation classification technique is limited by the non-optimal choice of $\ell_2$ metric norm for measuring the
1.2 Contributions

residual vectors obtained in the classification stage. We show that the residual vectors of face recognition images are not perfectly Gaussian distributed. We argue that a properly selected $\ell_p$ norm for nearest subspace classification in SRC can address this limitation. We also explain theoretically and numerically why such metric can suppress outliers.

Lastly, we prove that there exists a latent structure in the affinity matrix of each view despite the fact that multi-view face images for the same subject do not lie in the same subspace. This latent structure allows us to exploit the common information which is carried by different views of images. With a properly designed algorithm, the performance of multi-view face clustering can be enhanced at a low cost.

1.2.3 Experimental data and settings

Experimental data and settings are necessary for evaluating the proposed algorithms and providing an unbiased comparison.

Three publicly available databases are considered in this work. They are CMU-PIE Sim et al. (2002), Yale B Georghiades et al. (2001) and Multi-PIE Gross et al. (2010). 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 were taken under 13 different poses, 43 illumination conditions, and with 4 different expressions for each people. The Yale B database contains 5760 images of 10 subjects each seen under 576 viewing conditions (9 poses with 64 illumination conditions). A system of 15 cameras was used to take images in Multi-PIE. Thirteen cameras were placed at head height and spaced at $15^\circ$ intervals. The Multi-PIE contains 300 images for each people under 15 view points and 20 illumination conditions. In total, more than 750,000 images of 337 people were captured in Multi-PIE. These databases provide face images with
large variations in poses, expressions, and illuminations. Especially, Multi-PIE also has a significantly large number of subjects.

In general, we crop the human face according to the nose tip with eyes and mouth properly aligned. The areas under the chin and above the hair are not considered. The cropped image is down-sampled to $32 \times 32$ pixel for computational efficiency.

In multi-view face recognition, two sets of images are required. One is the gallery set which is also known as the dictionary. The gallery set contains the images of all subjects with known identity. In most of the time, the gallery set is generated with randomly selected images for each subject. The other one is the query set which is usually obtained from the cameras. It contains the images of an unknown subject. As we mentioned before, we would like to explore the usefulness of shared information among different views. Therefore, we use faces that have various poses to create the query set. Our query set consists of images from different views of the same subject.

In face clustering, the variations of face images are generally caused by illuminations and some small changes in expressions. This is because the images used for clustering are obtained from frames. Due to the short time between each frame, it is impossible to capture significant pose variation in few tens of frames. We randomly select face images with different illuminations and slightly various expressions for each subject to create the data set. When face clustering comes with multi-view data, the pose difference will be created by the placement locations of the cameras. In this case, a testing data set is generated from a group of single-view clustering data sets with different poses.

Through these experiment settings, we would like to obtain reasonable experimental settings which can simulate the real-world scenarios as close as possible.
1.3 Thesis structure

In Chapter 2, related background knowledge is presented to support understanding of this thesis. Some definitions of sparse representation classifications, sparse subspace clustering, and multi-view face recognition are presented. Some sparse representation based methods are reviewed. We also detail some sparse representation methods which are directly related to this thesis.

Chapter 3 proposes the mixed-norm sparse representation for multi-view face recognition. By introducing a new mixed-norm regularisation to sparse representation classification model, a new method is derived. We contrast the difference between our proposal and other state-of-the-art methods. The experiment clearly shows the advantages of our algorithms under different scenarios.

Chapter 4 discusses the usage of the $\ell_2$ norm in the original sparse representation classification method. We propose a different metric norm in the classification stage and explain theoretically and numerically why such a metric norm works. Experimental results support our model at the end.

Chapter 5 extends the sparse subspace framework to be used with the mixed norm. We first discuss the coefficient distribution with this mixed norm. We then provide the algorithm and convergence proof of proposed algorithm. The experiment is conducted on both CMU-PIE and Multi-PIE databases.

Chapter 6 investigates the complementary information among different views of the same subject. An efficient face clustering method is obtained based on this modelling. Experiments shown an outstanding performance when compared to other state-of-the-art methods.

In the end, Chapter 7 concludes this thesis in the perspective of sparse representation based face recognition and face clustering. Some interesting future directions are described.
Chapter 2

Background

The scope of research in this thesis involves two main branches of pattern recognition against face data: face recognition and face clustering. This chapter reviews the literature that is related to this work from each of these areas. This chapter is presented as follows. A review of the various approaches for multi-view face recognition is outlined in Section 2.1. In Section 2.3, a discussion on various techniques and algorithms used to cluster face images is given. Section 2.4 reviews related works for both face recognition and clustering in details, including the sparse representation classification and sparse subspace classification frameworks that this thesis extends. Section 2.5 describes the powerful alternative direction method of multipliers, which is used to prove convergence of the proposed algorithms in this thesis. Finally, a summary of this chapter is presented in Section 2.6.

2.1 General face recognition

In most face recognition systems, we usually have a gallery set which contains the images for each known subject, and a query image that does not have a
2.1 General face recognition

subject label. Our goal is to recognise this query image based on the gallery set. According to previous literature surveys Zhao et al. (2003); Zhang and Gao (2009), extensive studies have been done in Sirovich and Kirby (1987); Belhumeur et al. (1997); Wright et al. (2009) to resolve the general face recognition issues, such as pose, illumination, expression and occlusion, etc.

First, we will introduce several popular classifiers. The nearest neighbour (NN) is one of the most common and popular classifiers in Duda et al. (2001). 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 generalised to the nearest subspace (NS) in Ho et al. (2003). Instead of 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 classification (SRC) which proposed in Wright et al. (2009) 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. These classifiers are heavily used in various face recognition methods.

Eigenfaces is one of the earliest and successful methods for face recognition proposed by Sirovich and Kirby (1987) based on the principal component analysis (PCA). They first construct a projection matrix by applying PCA on a training image data set. Then the eigenfaces for the nearest neighbour classification can be obtained by using the matrix to project the testing image into subspaces. It achieves a satisfactory result compared with traditional template matching approaches, but it still has a limitation. Since Eigenfaces extract features based on the dominant factors, when the major variation in the data set is caused by illumination or other reasons instead of subject identities, the total scatter matrix
2.1 General face recognition

(covariance matrix) in the PCA cannot lead to the proper subspaces.

Unlike Eigenfaces which extracts features based on large data variations, Fisherface which is introduced by Sirovich and Kirby (1987) extracts features based on two criteria simultaneously: the extracted feature can maximise the difference among classes and minimise the variance between each face image within one class. Instead of using PCA, fisher’s linear discriminate analysis (LDA) was introduced to fulfil these requirements. This enhanced approach can achieve outstanding performance on many data sets.

Geometric approaches have also been extensively studied for a long period Rouhi et al. (2012). Geometric approaches based on facial component features such as eyes, nose, mouth and a shape of the head etc. Landmark points and relations among these components are used to extract the spatial features for improve face recognition performance. In Zhou and Wei (2006), Gabor wavelet features are generated by Gabor wavelet transformation and AdaBoost uses these features to perform face verification. They achieve a good performance on frontal face images. In Aguerrebere et al. (2007), a Face Bunch Graph is obtained from landmark points based on Gabor wavelet features for face recognition. However, all these methods are designed for improving face recognition and they achieve good performance with clean front images. They are not able to handle the multi-view problems such as occlusions or corruptions.

In Wright et al. (2009), the authors exploited the discriminative nature of sparse representation to perform classification. They demonstrated that human face images lie in subspaces, and faces of the same person lie in its local linear subspace. Any testing (or query) face image can be sparsely represented as a linear combination of the training (or gallery) set. A sparse representation classification is derived from this assumption. This SRC seeks a sparse linear representation for a given image, and labels it based on the residuals between the given image and reconstructions from the sparse representation of each class. This method achieves
2.2 Multi-view face recognition

great success in face recognition, especially where corruptions and occlusions are present.

Gaussian faces Lu and Tang (2014) proposed a novel approach based on discriminative Gaussian process latent variable model for face verification. The primary target for this method is exploiting additional data from multiple source-domains to improve generalisation performance of face verification in an unknown target-domain. In order to achieve this goal, the authors combined Gaussian process and a efficient form of kernel fisher discriminant analysis. Extensive experiments on Labelled Face in Wild datasets show an impressive performance.

2.2 Multi-view face recognition

The methods in previous section have shown a satisfactory performance. However, they 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.
2.2 Multi-view face recognition

Recently, a growing interest 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 Hadid et al. (2007), an extended volume of Local Binary Patterns is introduced to exploit the information among frames. It can achieve 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.

In Beymer (1994), a view-based face recognition system is designed. It uses template matching with images based single-view representation. Each input view is registered to a known person’s template by using locations of the eyes and nose. Fifteen gallery face images are used to cover a range of pose variations with approximately ±40° in yaw and ±20° in tilt. Then the typical template matching algorithm is used to match the templates around the eyes, nose and mouth. This method requires a large number of gallery images, and thus it is likely to fail when some poses are missing in gallery. Moreover, the high computational complexity and the low accuracy of template matching are the drawbacks of this method.

A panoramic view method is introduced in Singh et al. (2007). Instead of using a group of images, a panoramic view of a subject is built from a frontal view and rotated views in three steps:

- View alignment: a coarse affine alignment is applied on the face images with different poses;
- Image segmentation: 8×8 pixel boundary blocks for the segmentation are detected using phase correlation. These blocks are the connection regions...
2.2 Multi-view face recognition

for the two views in next step;

- Image stitching: the final panoramic image is stitched by a multi-resolution splinting, which connects the boundary blocks for two views.

A frontal image plus left and right views in 40° rotations have been shown the optimal combination of gallery images. The reason is that this combination covers the most face changes in horizontal rotations. After that, a support vector machine with some feature extraction is used to classify the query image based on the panoramic images for each gallery subject.

The above methods are view-based matching and their requirements are usually too strict. The gallery has to cover all the pose images for each subject. Another way to deal with the pose variation problem is by using a pose transformation. The active shape model is originally proposed in Cootes et al. (2002). It is a powerful tool to describe deformable object images. It is applied to face recognition across poses in Shan et al. (2006). Given a set of training images for one subject, where the feature points are manually marked, a shape and texture can be represented by applying principal component analysis (PCA) to the shape and texture distributions. Then, a new face can be represented by a vector, which controls the shape and texture variations in both shape and texture eigen spaces. This method achieves good recognition rates on large head pose angles in near real-time. However, manually marking points for training set is time consuming and unreliable in real-world applications.

Inspired by label propagation in Zhou et al. (2004), Kokiopoulou and Frossard (2010) proposed a graph-based classification for multi-view face recognition. This graph-based classification is a specific classification problem in label propagation and it constraints the unlabelled data to one single class. The authors formulate a simple and optimal algorithm for this single-class problem. This method can be naturally used for multiple-observation face recognition. In this work, the authors only use the smoothness constraint in label propagation to formulate the local
2.2 Multi-view face recognition

structure (manifold or graph). Since all the unknown samples are from same class, a class-wise distance which is measured by $\ell_2$ norm between the unknown class and each labelled class is used to classify the unknown samples. Although this manifold-based smoothing constrained method shows a great potential of multi-view recognition, it only treats the comparison between each sets individually. Thus, when the pose variation is large, it may not be able to handle well.

There are multiple images in each class in gallery, and there is only one image set which contains multiple images from same subject in query set. Based on this situation, the authors of Cevikalp and Triggs (2010) approximate each image sets in both gallery and query with a convex model, which is either an affine or a convex hull. They build an affine hull for each image set (query set and each class in gallery). The geometric distances between the affine hull of query and of each class in gallery are used to make the classification decision. They first find the affine linear subspaces for each pair of the query set and gallery set. The $\ell_1$ norm with some boundary conditions is applied on the coefficients of these subspaces to seek a more robust hull fitting. Instead of using all samples, a smaller number of samples is selected from each set to represent this set. This allows the subspaces to focus on the closest part and prevent outliers and deliver a robust and accurate representation. Once the coefficients are optimised, the distance between the two sets can be calculated. Although this method provides a novel way to compare the two sets to handle the multi-view face recognition, and it is more robust against outliers, it still has two limitations: (1) it cannot exploit information across multiple classes; (2) when there is a large difference between images in the same class, such as large pose variations, this method would perform poorly.

Since SRC considers both within each class and between multiple class factors, Tropp et al. (2006) introduced a multiple test samples generalisation of SRC, known as joint sparse representation-based classification (JSRC). 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. The original SRC is rewritten into a multi-task form by combining different single tasks together. An $\ell_{2,1}$ norm is applied on coefficients matrix. It adaptively selects a minimum number of atoms from gallery images, these atoms can best represent every query image at the 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 $90^\circ$ 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. Therefore, forcing the entire view share the same sets of atoms is not applicable to real-world multi-view face recognition.

To overcome this issue, joint dynamic sparse representation-based classification (JDSRC) was proposed in Zhang et al. (2012). The authors in Zhang et al. (2012) argue that whilst 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 known as joint dynamic sparsity. This joint dynamic sparsity brings 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. The dynamic active sets are the core part of JDSRC, which allows JDSRC to exploit the joint dynamic sparsity prior for multi-view face recognition. 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 Tropp et al. (2006) which is a naive greedy method and may not be convergent.
2.3 Multi-view face clustering

Face clustering is another important problem in computer vision. It has drawn considerable attention in the research community Elhamifar and Vidal (2013); Ho et al. (2003); Liu et al. (2013c); Zhang et al. (2015b). The aim of this task is to cluster face images by identity under external illumination conditions. It has been proven that the frontal face images with illumination changes are linearly separable Basri and Jacobs (2003). Therefore, linear subspace separation can be applied to this problem. Early approaches for these problems assume that the underlying data is modelled by a union of subspaces - $K$-subspaces Ho et al. (2003), mixture of probabilistic PCA (MPPCA) Tipping and Bishop (1999), generalised PCA Vidal et al. (2005) and spectral methods Lauer and Schnorr (2009); Lu et al. (2014). Recently, some more advanced approaches are proposed, they can be split in two main categories: sparse modelling and low-rank recovery. By using a dictionary as the data itself, a sparse or low-rank structure can be found. They construct an affinity graph on this sparse or low-rank structure. This affinity graph can be used for spectral clustering. These methods usually have a convex formulation which can provide a global optimal solution.

In Elhamifar and Vidal (2013), sparse subspace clustering (SSC) was introduced. It expresses each data point as a sparse linear combination of others. An $\ell_1$ norm regularisation is used to seek this sparse representation. Later on, a weighted formulation of SSC was proposed Pham and Venkatesh (2012). It shows that SSC can be significantly improved by exploiting geometric relation between data points as constraints. On the low-rank direction, the sparsity of singular values is promoted via a trace norm regularisation in Liu et al. (2010). In Favaro et al. (2011), they cast the clustering problem as finding a low-rank representation of the data based on the data itself.

The methods mentioned above were designed for faces of a single view. In many
real-world scenarios, computer vision applications have access to data which are taken in different representations or views Blum and Mitchell (1998); Zhang et al. (2015a). For example, images from different viewpoints for the same subject are captured when multiple cameras are placed in one location. These images can provide complementary information. It has been shown that this complementary information can improve the performance of classification or clustering Sun (2013). Since images obtained in multiple views correspond to different poses, they are no longer linearly separable. This may cause face clustering methods, in particular SSC, that are not designed to handle multi-view data to perform rather sub-optimally, since they are based on the linear subspace separation assumption.

Methods that address the multi-view problems beyond face clustering have been observed in the literature. In Kumar and Daumé (2011), a co-training spectral clustering is introduced. They use the spectral embedding from one way to constrain for the other view. By applying this approach, they iteratively solve the eigenvectors for two views in spectral clustering. A low-rank Markov chain multi-view clustering method is proposed in Xia et al. (2014). They decompose the Markov chains of each individual view into combinations of one universal low-rank matrix and residual matrices. Each residual matrix represents the difference between the universal low-rank matrix and a Markov chain for each individual view. By formulating a joint matrix factorisation process with a constraint that pushes clustering solution of each view towards a common consensus, a multi-view non-negative matrix factorisation is proposed in Liu et al. (2013b). They use a loss function to measure the disagreement between views and consensus in non-negative matrix factorisation. However, all these methods are designed for documents rather than face clustering.
2.4 Related work

The original sparse representation classification is proposed by Wright et al. (2009), by representing a face image with other images in the gallery, it shows an outstanding performance. However, SRC is designed to against single-view face recognition. On the other hand, joint sparse representation based classification (JSRC) Tropp et al. (2006) and joint dynamic sparse representation based classification was proposed in Zhang et al. (2012) are two well-developed methods for multi-view face problems. JSRC 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 addresses a set of query images from the same subject. Instead of finding the shared information at the atom level, it could be better to determine the sparsity pattern at the class level. A new concept of joint dynamic sparsity was introduced.

In the face clustering area, the idea of sparse representation also shows its power. An advanced face clustering method was proposed by Elhamifar and Vidal (2013). An affinity matrix which is used for spectral clustering could be obtained through the sparse representation process. A better clustering performance is achieved due to the robustness of sparse representation.

In this section, we will briefly discuss these state-of-the-art face recognition and clustering methods.

2.4.1 Sparse representation classification

In Wright et al. (2009), the authors exploited the discriminative nature of sparse representation to perform classification. They assumed that any test sample (such as face image) can be represented as a linear combination of training samples from each class. This representation is naturally sparse, and involves only
2.4 Related work

a small fraction of the overall dictionary. Based on this assumption, they introduced the sparse representation-based classification (SRC). Moreover, they further indicated that the sparse representation is found by solving a compressed sensing problem, which is close to the Lasso in statistics in functional form Zhao and Yu (2006); Tibshirani (1996).

In sparse representation classification, given a set of gallery images $A = [a_1, \ldots, a_N]$ where each $a_i \in \mathbb{R}^d$, one seeks a sparse combination of these images to represent an unknown image $y$

$$y \approx \sum_{i=1}^N x_i a_i = Ax.$$  \hspace{1cm} (2.1)

Such a sparse solution can be found by solving following problem in Wright et al. (2009)

$$x = \arg \min_x \|x\|_1 \text{ subject to } Ax = y,$$  \hspace{1cm} (2.2)

another formulation considering noise is

$$x = \arg \min_x \|x\|_1 \text{ subject to } \|Ax - y\|_2 < \varepsilon.$$  \hspace{1cm} (2.3)

According Compressed Sensing theory Candès et al. (2006); Donoho (2006), this is equivalent to its Lagrangian form

$$x = \arg \min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1.$$  \hspace{1cm} (2.4)

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

$$r_k = y - A_k x_k,$$  \hspace{1cm} (2.5)

where $A_k$ is the sub-matrix of $A$ that corresponds to all gallery images in class $k$, and $x_k$ is the sub-vector of $x$ with the corresponding sparse coefficients. Then the target $y$ is classified according to the minimum $\ell_2$ norm of the residual vectors

$$\text{class}(y) = \arg \min_k \|r_k\|_2^2.$$  \hspace{1cm} (2.6)
2.4 Related work

2.4.2 Joint sparse representation classification

Joint sparse representation based classification is a nature generalisation of SRC for multi-task face recognition, and it can be regarded as a special form group Lasso. A joint sparse assumption is applied to extract the shared information across all images. In Tropp et al. (2006), the authors state that the multiple sparse representation vectors share the same sparsity pattern. The authors of Obozinski et al. (2010) also point out that by penalising the sum of $\ell_2$ norm on the coefficients across all classes, similar sparsity pattern are encouraged. This leads to an improvement in classification performance. Later on, an extension has been introduced which applies this $\ell_{2,1}$ norm on multi-task sparse representation classification for face recognition Yuan and Yan (2010). When the face images in query set are from the same subject, they always share some information regardless of pose variations. This behaviour makes JSRC overcome the missing pose issue. If a pose image does not exist in gallery set, the $\ell_{2,1}$ norm can force the algorithm select the candidates based on the overall influence of all query images.

Consider a gallery image set $A$, which contains $c$ classes. For each class $A_i$, it has $N_i$ face images that may be captured with different poses. Then we write $A = [A^1, \ldots, A^c]$, and $A_i = [a^i_1, \ldots, a^i_{N_i}] \in \mathbb{R}^{d \times N_i}$, where $d$ is the dimensionality of images. Given a set of face images $Y = [y_1, \ldots, y_M]$ where each $y_i \in \mathbb{R}^d$. These $M$ images may also be captured with different poses, but from same person. The sparse representation coefficient matrix is denoted as $X = [x_1, \ldots, x_M]$ with respect to $A$.

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

follows

\[
\{\hat{x}_i\} = \arg \min_{x_i} \|x_i\|_1, \text{ for } i = 1, \ldots, M \tag{2.7}
\]

subject to

\[
\sum_{i=1}^{M} \|Ax_i - y_i\|_2 < \varepsilon. \tag{2.8}
\]

Recall that each \(x_i\) represents a pose image from \(Y\), and its rows represent the weights of corresponding gallery images. In addition, all images in \(Y\) comes from the same subject. Therefore, a joint sparse assumption can be applied to extract the shared information across all images in \(Y\) according to Tropp et al. (2006), 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 3.1(b). Therefore, by solving the following problem, the sparse representation vectors for multiple views can be found:

\[
\hat{X} = \arg \min_X \|X\|_{2,1} \tag{2.9}
\]

subject to

\[
\|Y - AX\|_F^2 < \varepsilon \tag{2.10}
\]

where \(\|\cdot\|_{2,1}\) is defined as the sum of the \(\ell_2\) norm of all rows of a matrix and a Frobenius norm is used for reconstruction error. By introducing \(\|\cdot\|_{2,1}\), the sparse representation matrix will have dense coefficients row-wise and sparse coefficients column-wise (see Figure 3.1(b)). This method is called JSRC. However, as stated by Zhang et al. (2012), 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 \(Y\) increase. Therefore, forcing the entire views to share the same set of atoms is not applicable to real-world multi-view face recognition.
2.4 Related work

2.4.3 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 atoms level Zhang et al. (2012) (see Figure 3.1(c)). Based on this assumption, they introduced the joint dynamic sparse representation classification model (JDSRC). 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 $G = [g_1, \ldots, g_s] \in \mathbb{R}$. Each dynamic active set $g_i$ contains the row indices of a set of coefficients which belong to the same class in coefficient matrix $X$. Only one index is selected in each column of $X$ for one dynamic active set. For example, $g_i(j)$ refers the row index for $j$-th column of the coefficients matrix $X$ in dynamic active set $i$. Based on these dynamic active sets, the following JDSRC model was developed in Zhang et al. (2012)

$$\hat{X} = \arg \min_X ||Y - AX||_F$$

subject to $||X||_G \leq K$, (2.12)

where $K$ is the sparsity level. Here, $||X||_G$ is a combination of $\ell_2$ norm and $\ell_0$ norm based on dynamic active sets. The $\ell_2$ norm is applied to the selected coefficients of each dynamic active set $g_i$ individually, then the $\ell_0$ norm is applied across all dynamic active sets. This joint dynamic sparsity regularisation term is defined as follows

$$||X||_G = || \|x_{g_1}\|_2, \|x_{g_2}\|_2, \ldots ||_0,$$ (2.13)

where $x_{g_i}$ indicates a set of coefficients that associated with dynamic active set $g_i$:

$$x_g = X(g_s) = [X(g_s(1), 1), \ldots, X(g_s(M), M)]^T \in \mathbb{R}^M$$ (2.14)

To solve the problem with $\ell_0$ norm and joint dynamic sparsity constraints, the authors of Zhang et al. (2012) proposed a greedy JDSRC algorithm which is
2.4 Related work

similar to simultaneous orthogonal matching pursuit Tropp et al. (2006) and compressive simultaneous orthogonal matching pursuit Duarte et al. (2009). As with any orthogonal matching pursuit (OMP) Tropp and Gilbert (2007) based method, they follow these steps:

- Select new candidates based on current residuals;
- Add these candidates to the selected atom set;
- Find a new coefficient to reduce the atom set size to the specified sparsity level by using this atom set;
- Update the residuals based on this new atom set.

These four steps are repeated until certain conditions are satisfied Duarte et al. (2009). In JDSRC, they introduce a new way to select atoms by using dynamic active set. They first generate dynamic active sets by using coefficients. 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 $\ell_2$ norm is computed for the coefficients from each dynamic active set, and the atoms from the active sets with $K$ largest $\ell_2$ norm will be selected as candidates for OMP optimisation. Since the dynamic active set is selected from each class and the $\ell_2$ norm is applied on dynamic active sets 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 discrimination between different subjects.

However, there are a few issues with JDSRC. Firstly, the candidates in a dynamic active set are selected by one 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 exist in the gallery for a subject,
2.4 Related work

this will force candidate selection to a false hit. Secondly, due to the complexity of the dynamic active set design, this JDSRC model is not able to be solved by convex optimisation techniques. Therefore, a greedy method is often used. Since greedy methods often have poor convergence properties, a robust and accurate solution may not be achieved.

2.4.4 Sparse subspace clustering

In unsupervised face recognition, is a powerful method which was proposed in Elhamifar and Vidal (2013). It was originated from the motion segmentation problem and has shown an outstanding performance on face clustering problem. Consider a set of $N$ face images in a $D \times N$ data matrix $A = [a_1, a_2, \ldots, a_N]$, where $a_i \in \mathbb{R}^D$ denotes one of the face images which is captured by video cameras. Assume that, they belong to $k$ subjects or they can form into $k$ clusters. Denote $\mathcal{S}_i$ as the index-set of the subspace (cluster) that point $a_i$ belongs to. Then SSC seeks a linear representation

$$a_i = \sum_{j \neq i} c_{ij} a_j = \sum_{j \in \mathcal{S}_i, j \neq i} c_{ij} a_j + \sum_{j \notin \mathcal{S}_i} c_{ij} a_j. \quad (2.15)$$

Ideally, the coefficients in the second summation of the right term are zeros, giving rise to a sparse representation. By capturing the linear representation of all points in $C$, we have

$$A = AC, \quad \text{diag}(C) = 0. \quad (2.16)$$

Denote $\|C\|_1 = \sum_{i,j} |c_{ij}|$, the $\ell_1$ norm of a matrix $C$, then when SSC recovers the sparse solution $C$ by minimising the following, subject to the above constraints

$$\|C\|_1 + \lambda \|A - AC\|_F^2. \quad (2.17)$$

Here, $\lambda > 0$ and $\ell_1$ norm is known to promote sparse solutions.
2.4 Related work

When data is corrupted or incomplete, SSC starts with the following modelling

\[ \mathbf{A} = \mathbf{AC} + \mathbf{E}, \quad \text{diag}(\mathbf{C}) = \mathbf{0}, \]  

(2.18)

where the variable matrix \( \mathbf{E} \) accounts for the corrupted observations in the data. SSC then recovers \( (\mathbf{C}, \mathbf{E}) \) by solving

\[
\begin{align*}
\arg \min_{\mathbf{C}, \mathbf{E}} & \quad \|\mathbf{C}\|_1 + \lambda\|\mathbf{E}\|_1 \\
\text{s.t.} & \quad \mathbf{A} = \mathbf{AC} + \mathbf{E}, \quad \text{diag}(\mathbf{C}) = \mathbf{0}.
\end{align*}
\]

When the data is incomplete or missing, the problem can be posed as a special case of the corruption problem, where the missing entries are filled with random numbers. SSC uses a different approach to fill up the missing entries in which the missing locations of the data point \( \mathbf{a}_i \) and the corresponding rows in \( \mathbf{A} \) are removed. Then, SSC solves (2.17) using the modified \( (\mathbf{a}_i^*, \mathbf{A}^*) \). Then, the sparse solution \( \mathbf{c}_i^* \) is used to recover incomplete \( \mathbf{a}_i \) as \( \mathbf{a}_i^{\text{complete}} = \mathbf{Ac}_i^* \) and also the recovered \( \mathbf{A}^* \) is used for clustering.

Once the coefficient matrix \( \mathbf{C} \) is obtained, the next step is to do final clustering. A symmetric affinity graph \( \bar{\mathbf{C}} \) is constructed by

\[
\bar{\mathbf{C}} = \left( \mathbf{C} + \mathbf{C}^T \right)/2.
\]

(2.19)

Now a standard Von Luxburg (2007) is used to perform the clustering. We first compute the Laplacian of \( \bar{\mathbf{C}} \) as

\[
\mathbf{L}_\mathbf{C} = \mathbf{I} - \mathbf{D}^{-1/2}\bar{\mathbf{C}}\mathbf{D}^{-1/2},
\]

(2.20)

where \( \mathbf{I} \) is an identity matrix of appropriate dimension. \( \mathbf{D} \) is a diagonal matrix with \( D_{ii} = \sum_{j=1}^N \bar{c}_{ij} \).

Then, we compute the first \( k \)-eigenvectors \( \mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_k] \) of \( \mathbf{L}_\mathbf{C} \). In the end, the \( k \)-means approach is used to cluster all data points into \( k \) clusters according to \( \mathbf{U} \).
### 2.5 Alternating direction method of multipliers

The alternating direction method of multipliers (ADMM) is a simple but powerful algorithm in convex optimisation, and it is particularly useful in machine learning Boyd et al. (2011). It was first introduced in the mid-1970s, and studied through the 1980s and mid-1990s. ADMM was developed for large-scale and massive optimisation problems. This ADMM framework is an algorithm that can solve a decomposable dual ascent problem with superior convergence properties of the method of multipliers. It converts the problem into an augmented Lagrangian form which was introduced to optimisation in the late 1960s Hestenes (1969); Powell (1967) and solves this problem by differentiation. Generally, ADMM solves the problems in the following form

\[
\begin{align*}
\text{minimise} & \quad f(x) + g(z) \\
\text{subject to} & \quad Ax + Bz = c
\end{align*}
\]  

(2.21)

where \( x \in \mathbb{R}^n, z \in \mathbb{R}^m, A \in \mathbb{R}^{p \times n}, B \in \mathbb{R}^{p \times m}, \) and \( c \in \mathbb{R}^p. \) When \( f \) and \( g \) are convex, we can form the augmented Lagrangian as follows

\[
L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + (\rho/2)||Ax + Bz - c||^2_2.
\]

Since the function \( f(x) \) and \( f(z) \) are separable, we can split the x-minimisation and z-minimisation into 2 separable problems. Therefore, ADMM consists of three parts which can be optimised iteratively

\[
x^{k+1} := \text{argmin}_x L_\rho(x, z^k, y^k)
\]

(2.22)

\[
z^{k+1} := \text{argmin}_z L_\rho(x^{k+1}, z, y^k)
\]

(2.23)

\[
y^{k+1} := y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)
\]

(2.24)

where \( \rho > 0. \) ADMM solves for \( x \) in (2.22), then it computes \( z \) in (2.23). After that, the dual variable is updated by (2.24). Since it updates \( x \) and \( z \) in an alternating way, it is called alternating direction. ADMM can be written in a
2.5 Alternating direction method of multipliers

scaled form, which is more convenient, by following steps. Define \( r = Ax + Bz - c \),

\[
y^T r + (\rho/2)||r||^2_2 = (\rho/2)||r||^2_2 + 2\sqrt{\rho/2} \frac{1}{\sqrt{2\rho}} y^T r + (1/2\rho)||y||^2_2 - (1/2\rho)||y||^2_2
\]

\[
= (\rho/2)||r + (1/\rho)y||^2_2 - (1/\rho)||y||^2_2
\]

\[
= (\rho/2)||r + u||^2_2 - (\rho/2)||u||^2_2
\]  (2.25)

where \( u = (1/\rho)y \) is the *scaled dual variable*. When we use the scaled dual variable, the augmented Lagrangian can be expressed as

\[
L_p(x, z, y) = f(x) + g(z) + (\rho/2)||Ax + Bz - c + u||^2_2 - (\rho/2)||u||^2_2. \tag{2.26}
\]

Then the parts of ADMM can be written as

\[
x^{k+1} := \text{argmin}_x (f(x) + (\rho/2)||Ax^k + Bz^k - c + u^k||^2_2), \tag{2.27}
\]

\[
z^{k+1} := \text{argmin}_z (g(z) + (\rho/2)||Ax^{k+1} + Bz - c + u^k||^2_2), \tag{2.28}
\]

\[
u^{k+1} := u^k + Ax^{k+1} + Bz^{k+1} - c. \tag{2.29}
\]

The stopping criterion is when the \( \ell_2 \) norm of dual residual and primal residual vectors are relative small, such as

\[
||s_1^{k+1}||_2 \leq \varepsilon_1, \quad s_1^{k+1} := \rho_1(x^{k+1} - x^k) \tag{2.30}
\]

\[
||s_2^{k+1}||_2 \leq \varepsilon_2, \quad s_2^{k+1} := \rho_2(z^{k+1} - z^k) \tag{2.31}
\]

\[
||r^{k+1}||_2 \leq \varepsilon_r, \quad r^{k+1} := Ax^{k+1} + Bz^{k+1} - c. \tag{2.32}
\]

In fact, we further extend this framework to multiple separated functions, such as

\[
\text{minimise} \quad f_i(x_i) \tag{2.33}
\]

subject to \( \sum_{i=1}^{n} A_i x_i = c \)
2.5 Alternating direction method of multipliers

In this case, we can easily obtain the following optimisation steps

\[
x_{i}^{k+1} := \arg\min_{x_i} f_i(x_i) + \left(\rho/2\right) \|A_i x_i^k + \sum_{j=1, j \neq 1}^n A_j x_j^k - c + u_i^k \|^2_2
\]  

(2.34)

\[
\ldots
\]

\[
x_{n}^{k+1} := \arg\min_{x_j} f_n(x_n) + \left(\rho/2\right) \|A_n x_n^k + \sum_{j=1, j \neq n}^n A_j x_j^k - c + u_n^k \|^2_2
\]  

(2.35)

\[
u_{i}^{k+1} := u_i^k + A_i x_i^{k+1} + \sum_{j=1, j \neq 1}^n A_j x_j^{k+1} - c
\]  

(2.36)

\[
\ldots
\]

\[
u_{n}^{k+1} := u_n^k + A_n x_n^{k+1} + \sum_{j=1, j \neq n}^n A_j x_j^{k+1} - c.
\]  

(2.37)

and stopping criterion

\[
\|s_{1}^{k+1}\|_2 \leq \varepsilon_{s_1}, \quad s_{1}^{k+1} := \rho_1 (x_{1}^{k+1} - x_{1}^{k})
\]  

(2.38)

\[
\ldots
\]

\[
\|s_{n}^{k+1}\|_2 \leq \varepsilon_{s_n}, \quad s_{n}^{k+1} := \rho_2 (x_{n}^{k+1} - x_{n}^{k})
\]  

(2.39)

\[
\|r_{1}^{k+1}\|_2 \leq \varepsilon_{r_1}, \quad r_{1}^{k+1} := A_1 x_1^{k+1} + \sum_{j=1, j \neq 1}^n A_j x_j^{k+1} - c
\]  

(2.40)

\[
\ldots
\]

\[
\|r_{n}^{k+1}\|_2 \leq \varepsilon_{r_n}, \quad r_{n}^{k+1} := A_n x_n^{k+1} + \sum_{j=1, j \neq n}^n A_j x_j^{k+1} - c.
\]  

(2.41)

Whilst the ADMM framework may be slow to converge with high accuracy, it can converge to the modest accuracy within a few tens of iterations. This behaviour makes ADMM suitable for large-scale problems. Besides, some machine learning algorithms have been implemented in the ADMM framework and show exceptional performance in accuracy and time complexity, such as Lasso, group Lasso, support vector machine, etc. Boyd et al. (2011).
2.6 Summary

This chapter has presented a review of the literature which is related to this thesis. It begins with a short discussion of general face recognition approaches, such as (PCA, LDA, SRC). Then, we compared and briefly revised some multi-view face recognition methods. Following this is a brief overview of multi-view face clustering methods. Then, we explained face recognition methods: SRC, JSRC, JDSRC and face clustering methods: SSC in details. Finally, we introduced the ADMM framework which allows us to derive our algorithms in a convex optimisation form.

In the following chapters, we explore how face recognition and face clustering can be enhanced in the case of multi-view face recognition.
Chapter 3

Mixed-norm sparse representation for multi-view face recognition

In many practical situations, it is desirable to recognising a number of unknown faces at the same time, such as, recognising a person from a video sequence. Under this situation, the face images usually come from one subject with different poses (views). There will be some shared information among multiple views which can be used to improve face recognition performance. However, on one hand, if we simply use SRC to perform multiple views face recognition, the sparse representation vectors will be generated individually (see Figure 3.1(a)). Information between different views is not involved in this scenario. On the other hand, when the pose variation is too large or missing, ‘shared information’ may not be properly extracted, leading to poor recognition results.

As we discussed in Section 2.2, this shared information has been investigated by few popular methods. By assuming that the query face images share the same sparsity pattern, JSRC was introduced by Tropp et al. (2006). Instead of
finding the sparse solution in general SRC for each individual image, JSRC tries to seek a set of query images from the same subject. An $\ell_{2,1}$-norm regulariser is used to select the atoms. However, this assumption is not valid when there are large variations in the query set. This makes it inefficient to deal with multi-view problem. The detail of JSRC is presented in Section 2.4.2. To overcome this issue, Zhang et al. (2012) propose JDSRC. They claim that the same sparsity pattern should be at the class level, not the atom level. A new concept of joint dynamic sparsity is introduced. Instead of selecting the same atom for all poses as JSRC, JDSRC selects atoms from the whole class to represent all poses. Section 2.4.3 explains JDSRC in details. Nevertheless, there still are limitations of JDSRC. Due to the property of the joint dynamic sparsity, it is forced to select a ‘similar’ atom to represent a missing pose in gallery. This may not result in a robust solution. Moreover, the algorithm of JDSRC is achieved by an extension of simultaneous orthogonal matching pursuit Tropp et al. (2006) which is a naive greedy method and may not converge. Therefore, a new algorithm is needed to solve this challenging multi-view (multi-pose) problem.

In Shi et al. (2011), the authors argue that the robustness of SRC based methods should be achieved by using the $\ell_1$-loss function instead of the $\ell_2$ 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 to find a sparse linear combination of signal atoms from an over-complete dictionary Candes et al. (2006); Candes and Tao (2006). It was then applied to face recognition in Wright et al. (2009). As solving this CS problem is similar 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 Nguyen et al. (2011). Statistically, this is more generic and provably better than the least entropy and error correction alternative discussed in a rejoinder Wright et al.
(2011) by the authors of SRC against the paper of Shi et al. (2011). However, this is obtained at the cost of an extra regularisation parameter. In the related robust CS paper Pham and Venkatesh (2012), 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 chapter, 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 $\ell_1$ loss instead of the general $\ell_2$ loss. Indeed, the $\ell_1$-norm loss function we used in this work, which is also an open question discussed in Shi et al. (2011), is also known in the robust statistics literature to be optimal for noise modelled as a Cauchy distribution.

The chapter is organised as follows. We first derive our Multi-pose face recognition via sparse representation (MSRC) based on the ADMM framework in Section 5.1. Then, we provide extensive experiments on CMU-PIE, Yale B and Multi-PIE data sets in Section 5.2. Section 5.3 provided a summary of this chapter.
Figure 3.1: Four sparsity models for the coefficient matrix $X$. The column represents the poses/images in the gallery as grouped by subjects, the row represents different images in the unknown set $Y$. Each column denotes a sparse representation vector and each square denotes a coefficient. 

- Independent sparsity (as in SRC): all coefficients are selected independently based on $\ell_1$ regularisation; 
- Joint sparsity (as in JSRC): only few gallery images/poses are selected simultaneously by all test images; 
- Joint dynamic sparsity (as in JDSRC): sparsity in terms of sorted active groups, active coefficients are multiples of the number of test images; 
- Mixed Sparsity: a well trade-off balance of both $a)$ and $b)$ to adaptively select the suitable class-level and atom-level sparsity.
3.1 Multi-view mixed norm robust sparse representation

In this section, we present a novel method to overcome the issues with JD-SRC, we denote this method as mixed-norm sparse representation classification (MSRC). We first proceed with some necessary notations. Consider a gallery image set $A$, which contains $c$ classes. Each class $A_i$ has $N_i$ face images that may be captured with different poses. Suppose that $A = [A^1, \ldots, A^c]$, and $A^i = [a^i_1, \ldots, a^i_{N_i}] \in \mathbb{R}^{d \times N_i}$, where $d$ is the dimensionality of the images. Denote a test set of face images as $Y = [y_1, \ldots, y_M]$ where each $y_i \in \mathbb{R}^d$. These $M$ images may also be captured with different poses, but from the same person. The sparse representation coefficient matrix can be denoted as $X = [x_1, \ldots, x_M]$ with respect to $A$.

3.1.1 Model formulation

To start with, we recall from Shi et al. (2011) that as the $\ell_2$ norm used for residuals in SRC may not lead to a robustness solution, a $\ell_1$ norm should be applied on the residuals:

$$\hat{x} = \arg\min_{x \in \mathbb{R}^N} ||y - Ax||_1.$$  \hspace{1cm} (3.1)

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 optimisation.

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.
3.1 Multi-view mixed norm robust sparse representation

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 3.1(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 3.1(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 \( \ell_1 \) norm to solve them in the same time. The final decision is not based on any individual factor, it is an overall view (Figure 3.1(d)). We note that the difference between the sparsity patterns (c) and (d) as shown in Figure 3.1 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 coefficient in each column. Only the top active groups are finally chosen through global optimisation. 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
3.1 Multi-view mixed norm robust sparse representation

and stable formulation for sparse representation:

$$\hat{x} = \arg \min_{x \in \mathbb{R}^N} \|y - Ax\|_1 + \lambda \|x\|_1.$$  \hspace{1cm} (3.2)

Here, the regularisation parameter $\lambda$ specifies the desired sparsity. Clearly, the robust sparse formulation is even more general than (3.1), because (3.1) is a special case when one sets $\lambda = 0$. Thus, solving this formulation allows one to obtain a solution for (3.1) easily.

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

$$\hat{x}_1 = \arg \min_{x_1 \in \mathbb{R}^N} \{\|y_1 - A x_1\|_1 + \lambda \|x_1\|_1\},$$ \hspace{1cm} (3.3)

$$\vdots$$

$$\hat{x}_T = \arg \min_{x_T \in \mathbb{R}^N} \{\|y_T - A x_T\|_1 + \lambda \|x_T\|_1\}.$$ \hspace{1cm} (3.4)

We collect the variables in matrix quantities

$$X = [x_1, \ldots, x_T]$$ \hspace{1cm} (3.5)

$$Y = [y_1, \ldots, y_T].$$ \hspace{1cm} (3.6)

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

$$\hat{X} = \arg \min_X \|Y - AX\|_1 + \lambda \|X\|_1.$$ \hspace{1cm} (3.7)

Here, the $\ell_1$ norm for matrices is defined as $\|X\|_1 = \sum_{i,j} |X_{ij}|$. Since the $\ell_1$ norm of matrix can be rewrite as sum of $\ell_1$ norm of vectors in this matrix, the solution of (3.7) is equivalent to (3.3) and (3.4). The $\ell_1$ 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 at the 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 $X$ represents one view of a subject, and each
3.1 Multi-view mixed norm robust sparse representation

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 (3.7). An \( \ell_2,1 \) 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

\[
\hat{\mathbf{X}} = \arg \min_{\mathbf{X}} \| \mathbf{Y} - \mathbf{A} \mathbf{X} \|_1 + \lambda R(\mathbf{X}) \tag{3.8}
\]

where the mixed norm is defined as

\[
R(\mathbf{X}) = \gamma \| \mathbf{X} \|_1 + (1 - \gamma) \| \mathbf{X} \|_{2,1} \tag{3.9}
\]

Here, the block regulariser \( \ell_2,1 \) norm is defined as the sum of the \( \ell_2 \) norm of all rows of a matrix. It is known from statistics that such \( \ell_2 \) norm promotes dense solutions. The parameter \( \gamma \) 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 \( \gamma \) between 0 and 1, the formulation automatically adapts to the underlying statistics. In addition, since this mixed-norm regularisation is composed of \( \ell_1 \) and \( \ell_2,1 \) norms, we can show the upper bound for both prediction error and regularisation error as follows

\[
\| \mathbf{A}(\hat{\mathbf{X}} - \mathbf{X}^0) \|^2 / n + \lambda R(\hat{\mathbf{X}} - \mathbf{X}^0) \leq 4\lambda^2 \left( \frac{\gamma s_1}{\phi_1^2} + \frac{(1 - \gamma)s_2}{\phi_{2,1}^2} \right), \tag{3.10}
\]

where \( \hat{\mathbf{X}} \) is the solution obtained by the mixed norm and the \( \mathbf{X}^0 \) is the “true solution” or “ground truth”. The \( \phi_1 \) and \( \phi_{2,1} \) are some positive constants. The \( s_1 \) and \( s_{2,1} \) are the numbers of non-zero entries of \( \mathbf{X}^0 \) (see more details in Appendix A). We term this proposed formulation 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 exploit the shared information across the images in the query set by the \( \ell_{2,1} \) norm on \( \mathbf{X} \). 2). By introducing
the $\ell_1$ norm on $X$, 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 $\ell_1$ norm in (3.9). 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 highly 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 ground
3.1 Multi-view mixed norm robust sparse representation

truth of subject class A. Figure 3.2 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 recognised also has widespread poses. In SRC, images in the test set tend to select the nearest gallery images as their representatives in a rough proximity sense. Thus, the spatial distribution in Figure 3.2 aids in the explanation of the coefficients obtained by SRC, JSRC, and MSRC respectively as shown in Figure 3.3. 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 $U_1$ and $U_2$, both SRC and JSRC have a difficulty in distinguishing between class A and B, which is evident through the values of their coefficients (Figure (3.3.a) and (3.3.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 (3.3.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 $A_1$ 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 $A$.

### 3.1.2 Algorithm

We now discuss a solution for the formulation above. We follow the ADMM framework in the convex optimisation literature Boyd *et al.* (2011). By utilising the ADMM framework, we show that our algorithm is computationally efficient. Note that this problem is convex in $X$ 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

$$\min_{X, V, Z, T} \|V\|_1 + \alpha\|Z\|_1 + \beta\|T\|_{2,1}$$

subject to

$$V = AX - Y$$

$$Z = X$$

$$T = X.$$  \hspace{1cm} (3.11)

Note that an additional variable $T$ is introduced to the single-task case to effectively decouple the block regularisation. Thus, we can consider the augmented Lagrangian

$$\mathcal{L} = \|V\|_1 + \alpha\|Z\|_1 + \text{tr}[W_1^T(AX - V - Y)]$$

$$+ \frac{\eta_1}{2}\|AX - V - Y\|^2_2 + \beta\|T\|_{2,1}$$

$$+ \text{tr}[W_2^T(X - Z)] + \frac{\eta_2}{2}\|X - Z\|^2_F$$

$$+ \text{tr}[W_3^T(X - T)] + \frac{\eta_3}{2}\|X - T\|^2_F.$$  \hspace{1cm} (3.12)
3.1 Multi-view mixed norm robust sparse representation

Figure 3.3: Coefficients of SRC outlined in Wright et al. (2009), JSRC outlined in Tropp et al. (2006) and MSRC for Figure 3.2. 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.
3.1 Multi-view mixed norm robust sparse representation

Here, we omit the arguments \((X, V, Z, T, W_1, W_2, W_3)\) of the Lagrangian for notational simplicity and denotes \(\text{tr} \left[ \bullet \right] \) for the trace of a matrix. As with ADMM, we scale dual variables \(U_i = W_i/\eta_i, i = 1, 2, 3\), to obtain a simpler form

\[
\mathcal{L} = \|V\|_1 + \alpha\|Z\|_1 + \beta\|T\|_{2,1} + \frac{\eta_1}{2}\|AX - V - Y + U^k_1\|_F^2 + \frac{\eta_2}{2}\|X - Z + U^k_2\|_F^2 + \frac{\eta_3}{2}\|X - T + U^k_3\|_F^2 + \text{const}, \tag{3.13}
\]

where the constant is independent of the primal variables \(X, V, Z\).

Again, the updates for the variables are easily computed under the ADMM principle. For \(X\), we find the update from

\[
X^{k+1} = \arg\min_X \frac{\eta_1}{2}\|AX - V^k - Y + U^k_1\|_2^2 + \frac{\eta_2}{2}\|X - Z^k + U^k_2\|_2^2 + \frac{\eta_3}{2}\|X - T^k + U^k_3\|_2^2, \tag{3.14}
\]

which yields the exact solution

\[
X^{k+1} = H^{-1}q, \tag{3.15}
\]

where \(H^{-1} = (\eta_1 A^T A + (\eta_2 + \eta_3)I)^{-1}\) can be computed once, and the update term is

\[
q = \eta_1 A^T (V^k + Y - U^k_1) + \eta_2 (Z^k - U^k_2) + \eta_3 (T^k - U^k_3). \tag{3.16}
\]

As can be seen in (3.15), the update step of \(X\) is computationally expensive. Here, the matrix under inversion has dimensions \(N \times N\) where \(A \in \mathbb{R}^{d \times N}\). 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 \(H\) is a positive definite matrix then it admits
3.1 Multi-view mixed norm robust sparse representation

the factorisation $H = LL^T$ and thus $H^{-1}q$ can be efficiently computed by solving $LX_1 = q$ first, then $L^TX = X_1$, which can be written as $X = L^T \setminus (L \setminus q)$.

For variable $V$, the update step solves

$$V^{k+1} = \arg \min_V \|V\|_1 + \eta_1 \frac{1}{2} \|Q^k - V\|_2^2,$$  \hspace{1cm} (3.17)

where $Q^k = AX^{k+1} - Y + U_1^k$. Likewise, for $Z$ the update step solves

$$Z^{k+1} = \arg \min_Z \alpha \|Z\|_1 + \eta_2 \frac{1}{2} \|P^k - Z\|_2^2,$$  \hspace{1cm} (3.18)

where $P^k = X^{k+1} + U_2^k$.

As $\| \cdot \|_1$ is absolute value, the first terms in both (3.17) and (3.18) 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 $V$ and $Z$ are defined as follows

$$V^{k+1} = S_{1/\eta_1}(Q^k),$$  \hspace{1cm} (3.19)

$$Z^{k+1} = S_{\alpha/\eta_2}(X^{k+1} + U_2^k),$$  \hspace{1cm} (3.20)

where this soft-thresholding shrinkage operator is defined as

$$S_\tau(X) = \{ t : t_i = \text{sign}(X_{i,j}) \max(|x_{i,j}| - \tau, 0) \}. \hspace{1cm} (3.21)$$

For the last primal variable $T$, the update steps are only slightly different

$$T^{k+1} = \arg \min_T \frac{\eta_3}{2} \|X^{k+1} + U_3^k - T\|_F^2 + \beta \|T\|_{2,1}.$$  \hspace{1cm} (3.22)

To solve this problem, suppose that $t_i$ and $l_i$ are the $i$th row vectors of $T$ and $X^{k+1} + U_3^k$ respectively, then we decompose the problem as

$$T^{k+1} = \arg \min_{t_i,l_i} \sum_i \frac{\eta_3}{2} \|l_i - t_i\|_2^2 + \beta \|t_i\|_2.$$  \hspace{1cm} (3.23)

Thus, we can find each row of $T$ separately by exploiting the following result
3.1 Multi-view mixed norm robust sparse representation

Figure 3.4: Illustration of Lemma 3.1.1. The circle (or sphere in high dimension space) which centred at \( l \) with radius \( R \) is the set of all feasible points for \( t \). When we constraint \( t \) with minimises \( \ell_2 \) norm, we will have the only solution \( t_{\text{min}} \).

**Lemma 3.1.1.** The solution of the optimisation problem

\[
\min_t \frac{\eta_3}{2} \| l - t \|_2^2 + \beta \| t \|_2
\]

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

This result can be easily proved by a geometrical argument as shown in Figure 3.4. Indeed, suppose that \( t^* \) is the solution of the problem then we consider all feasible \( t \) such that \( \| l - t \|_2 = \| l - t^* \|_2 = R \). Then, it is observed that the set of those feasible points is the sphere centred at \( l \) with radius \( R \). Among all those feasible points, the solution must be the one that minimises \( \| t \|_2 \), 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 \( t = \kappa l \) with \( 1 \geq \kappa \geq 0 \). Then straightforward manipulations easily lead to the result.

Finally, the updates for dual variables are

\[
U_1^{k+1} = U_1^k + A X^{k+1} - V^{k+1} - Y, \quad (3.24)
\]
\[
U_2^{k+1} = U_2^k + X^{k+1} - Z^{k+1}, \quad (3.25)
\]
\[
U_3^{k+1} = U_3^k + X^{k+1} - T^{k+1}. \quad (3.26)
\]
3.1 Multi-view mixed norm robust sparse representation

Stopping criterion and convergence

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

$$
\begin{align*}
\min_{x} & \quad f(x) + g(z) \\
\text{s.t.} & \quad Ax + Bz = b.
\end{align*}
$$

(3.27)

where both $f(x)$ and $g(z)$ are convex functions. However, there are three primal variables in our problem (3.11). 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

$$
\begin{align*}
\min_{X, Z, T} & \quad f(X) + g(Z) + h(T) \\
f(X) & = \|Y - AX\|_1, \\
g(Z) & = \alpha\|Z\|_1, \\
h(T) & = \beta\|T\|_{2,1}, \\
\text{s.t.} & \quad X - Z = 0 \\
& \quad X - T = 0.
\end{align*}
$$

(3.28)

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

$$
k(Z') = g(Z) + h(T). 
$$

(3.29)

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

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

(3.30)
or equivalently $C_x X + C_{z'} Z' = 0$ where $C_x = [I; I]$ and $C_{z'} = -I$. Thus, the proposed formulation can be expressed in the same form as the original ADMM as follows

$$
\begin{align*}
\min_{X, Z'} & \quad f(X) + k(Z') \\
\text{s.t} & \quad C_x X + C_{z'} Z' = 0
\end{align*}
$$

and thus it inherits all desirable properties of ADMM.

According to Boyd et al. (2011), the reasonable termination criterion for the proposed method are when the Frobenius norms of the residual vectors for primal and dual are sufficiently small,

$$
\begin{align*}
S_1^k &= \|\eta_1 (V^{k+1} - V^k)\|_F \leq \varepsilon_{S_1}, \\
S_2^k &= \|\eta_2 (Z^{k+1} - Z^k)\|_F \leq \varepsilon_{S_2}, \\
S_3^k &= \|\eta_3 (T^{k+1} - T^k)\|_F \leq \varepsilon_{S_3}, \\
R_1^k &= \|AX^k - Y - V^k\|_F \leq \varepsilon_{R_1}, \\
R_2^k &= \|X^k - Z^k\|_F \leq \varepsilon_{R_2}, \\
R_3^k &= \|X^k - T^k\|_F \leq \varepsilon_{R_3}.
\end{align*}
$$

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

$$
\begin{align*}
\varepsilon_{S_1} &= \sqrt{n} \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \|\eta_1 U_1^k\|_F, \\
\varepsilon_{S_2} &= \sqrt{n} \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \|\eta_2 U_2^k\|_F, \\
\varepsilon_{S_3} &= \sqrt{n} \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \|\eta_3 U_3^k\|_F, \\
\varepsilon_{R_1} &= \sqrt{n} \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \max\{\|AX^k - Y\|_F, \| - V^k\|_F\}, \\
\varepsilon_{R_2} &= \sqrt{n} \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \max\{\|X^k\|_F, \| - Z^k\|_F\}, \\
\varepsilon_{R_3} &= \sqrt{n} \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \max\{\|X^k\|_F, \| - T^k\|_F\}.
\end{align*}
$$

where $\varepsilon_{\text{abs}} > 0$ is for absolute tolerance and $\varepsilon_{\text{rel}} > 0$ is for the relative tolerance. The notation $n$ indicates number of faces in the gallery set. The relative tolerance

48
3.1 Multi-view mixed norm robust sparse representation

Figure 3.5: Convergence of proposed method. The objective function is converged after 30 iterations.

$\varepsilon^{rel}$ might be chosen from $10^{-3}$ or $10^{-4}$ based on application practice Boyd et al. (2011). In this paper, we choose $10^{-4}$ for both absolute and relative tolerances.

Although the ADMM framework can be slow to converge to high accuracy, when we setup the proper stopping criterion, the proposed ADMM-based method can converge to modest accuracy within a few tens of iterations (see Figure 3.5). This behaviour 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 multiple views.
3.1 Multi-view mixed norm robust sparse representation

Figure 3.6: An example of the pose variations in CMU-PIE face database Sim et al. (2002). 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.

**Recognition and classification**

Once the sparse representation matrix $X$ is found for all views $Y$, the classification is delivered by computing the fitness of the query set with respect to the sparse solution. By following Zhang et al. (2012), there is only one decision made simultaneously on the class label for the whole query set based on $X$ by combining the residuals for each image in the query set. Denote $A_k$ is the subset of the gallery images corresponding faces of class $k$, and $X^k$ is the corresponding coefficient subset for all query images. The fitness for class $k$ is represented by the residual matrix $F_k = Y - A^k X^k$

$$\text{class}(Y) = \arg\min_k \|F_k\|_F^2. \quad (3.44)$$

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

50
3.2 Experiments

In this section, we present extensive experiments on CMU-PIE Sim et al. (2002), Yale B Georgiades et al. (2001) and Multi-PIE Gross et al. (2010) face databases. Examples of CMU-PIE images are shown in Figure 3.6. 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 prepossessing is performed before comparison, such as aligning and cropping face images, histogram equalisation 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 Shi et al. (2011), we also include the original SRC in Wright et al. (2009) and JSRC in Tropp et al. (2006). In addition, some popular base line face recognition techniques are also evaluated, including principal component analysis (PCA) and linear discriminate analysis (LDA) Belhumeur et al. (1997). Since SRC is originally used for single task, we follow Zhang et al. (2012) to use majority voting for SRC classification step and do the same for JSRC. 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 regularisation parameter $\lambda$ 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 difference is that MSRC has both the regularisation parameter $\lambda$ and the mixed norm parameter $\gamma$ (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 $\gamma$ in controlling the pose variation and our intensive numerical studies, we suggest that the computational increase in cross validation due to $\gamma$ might be reduced by a preliminary estimation of the pose variation. We
3.2 Experiments

Figure 3.7: Recognition rate under different number of views with dimension $d=64$ for CMU-PIE

notice that when there is large pose variation in the test set, a larger $\gamma$ is preferred and vice versa. This suggests that if we use a reliable pose detection method such as Murphy Chutorian and Trivedi (2009), we may have a good estimate of the pose variation and hence a fixed $\gamma$ can be set without sacrificing an increase in computation due to cross validation.

3.2.1 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 Zhang et al. (2012) for CMU-PIE data sets. Images in the training set are selected based on a pose subset $[0^\circ, \pm 22.5^\circ, \pm 45^\circ, \pm 67.5^\circ, \pm 90^\circ]$. 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.
3.2 Experiments

Figure 3.8: Recognition rate under different number of views with dimension $d=64$ for Yale B

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 also apply these settings to Yale B data sets. This makes our experiments more realistic and challenging.

In Figure 3.7 and 3.8 we compare the classification accuracy of the proposed method with others on both CMU-PIE and Yale B. 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
3.2 Experiments

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 Zhang et al. (2012), 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 Zhang et al. (2012). But, a closer inspection reveals that the authors in Zhang et al. (2012) used greedy algorithms for solving the sparse problems, which is known to be inferior to the convex optimisation 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 specialised ADMM.

3.2.2 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 Wright et al. (2009). Following Zhang et al. (2012), 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 $d$
3.2 Experiments

Figure 3.9: Recognition rate under different dimensionality with number of views $M=5$ for CMU-PIE

$M = [8, 16, 32, 64, 128]$. The comparison results are shown in Figure 3.9 and 3.10. 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.
3.2 Experiments

Figure 3.10: Recognition rate under different dimensionality with number of views $M=4$ for Yale B
3.2 Experiments

3.2.3 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 3.1 for CMU-PIE and 3.2 for Yale B. As shown in Table 3.1, 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 3.1. 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 3.2 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.
3.2 Experiments

<table>
<thead>
<tr>
<th>view(s)</th>
<th>Unseen</th>
<th>Mixed</th>
<th>Same</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>12.59%</td>
<td>17.91%</td>
<td>16.67%</td>
</tr>
<tr>
<td>LDA</td>
<td>24.40%</td>
<td>42.26%</td>
<td>41.29%</td>
</tr>
<tr>
<td>SRC</td>
<td>55.37%</td>
<td>84.78%</td>
<td>93.28%</td>
</tr>
<tr>
<td>JSRC</td>
<td>58.06%</td>
<td>88.66%</td>
<td>95.52%</td>
</tr>
<tr>
<td>JDSRC</td>
<td>48.51%</td>
<td>83.58%</td>
<td>92.99%</td>
</tr>
<tr>
<td>MSRC</td>
<td>73.88%</td>
<td>95.82%</td>
<td>98.21%</td>
</tr>
</tbody>
</table>

Table 3.1: Face recognition against unseen pose for CMU-PIE

<table>
<thead>
<tr>
<th>view(s)</th>
<th>Unseen</th>
<th>Mixed</th>
<th>Same</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>36.20%</td>
<td>36.10%</td>
<td>35.50%</td>
</tr>
<tr>
<td>LDA</td>
<td>52.70%</td>
<td>56.90%</td>
<td>52.70%</td>
</tr>
<tr>
<td>SRC</td>
<td>84.50%</td>
<td>88.50%</td>
<td>89.50%</td>
</tr>
<tr>
<td>JSRC</td>
<td>86.00%</td>
<td>89.00%</td>
<td>89.00%</td>
</tr>
<tr>
<td>JDSRC</td>
<td>84.00%</td>
<td>89.00%</td>
<td>89.00%</td>
</tr>
<tr>
<td>MSRC</td>
<td>90.00%</td>
<td>93.00%</td>
<td>94.50%</td>
</tr>
</tbody>
</table>

Table 3.2: Face recognition against unseen pose for Yale B

3.2.4 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. We followed the same setting in Zhang et al. (2012). It is organised as follows. Face images of all 13 poses from Figure 3.6 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: \([0^\circ, \pm 22.5^\circ], [0^\circ, \pm 45^\circ], [0^\circ, \pm 67.5^\circ] \) and \([0^\circ, \pm 90^\circ]\). There are 3 images (one image for each pose) in each group. The testing sets are generated by randomly selecting from
3.2 Experiments

these 4 pose groups. As can be seen in Table 3.3, 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 $[0^\circ, \pm 22.5^\circ]$ group with 8% improvement compared to its best competitor.

<table>
<thead>
<tr>
<th></th>
<th>22.5°</th>
<th>45°</th>
<th>67.5°</th>
<th>90°</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>24.43%</td>
<td>24.68%</td>
<td>24.38%</td>
<td>23.43%</td>
</tr>
<tr>
<td>LDA</td>
<td>59.75%</td>
<td>60.00%</td>
<td>58.11%</td>
<td>54.13%</td>
</tr>
<tr>
<td>SRC</td>
<td>82.24%</td>
<td>80.30%</td>
<td>80.15%</td>
<td>69.25%</td>
</tr>
<tr>
<td>JSRC</td>
<td>84.48%</td>
<td>84.63%</td>
<td>83.28%</td>
<td>74.18%</td>
</tr>
<tr>
<td>JDSRC</td>
<td>87.16%</td>
<td>84.63%</td>
<td>87.16%</td>
<td>78.21%</td>
</tr>
<tr>
<td>MSRC</td>
<td><strong>95.52%</strong></td>
<td><strong>94.33%</strong></td>
<td><strong>93.73%</strong></td>
<td><strong>88.96%</strong></td>
</tr>
</tbody>
</table>

Table 3.3: Face recognition against large pose variations

3.2.5 Face recognition with a large number of subjects

To examine how the compared methods scale with a large number of subjects, we use the Multi-PIE data set to perform two sets of experiments. To make our experiments more realistic and challenging, we mix all images from 4 different seasons altogether for 337 subjects with 102 females and 235 males. Images in each training set are selected based on following poses $[0^\circ, \pm 30^\circ, \pm 60^\circ, \pm 90^\circ]$. Three face images are selected for each subject with each pose in the training.
3.2 Experiments

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 from all 337 subjects and ten sets are created. These pairs of data set correspond to 64 (34 females and 34 males), 136 (68 females and 68 males), 204 (102 females and 102 males), 272 (all randoms) subjects. All subjects were randomly chosen from 337 subjects. Since we need to generate 10 different random datasets to obtain fair experiment results, we left at least 65 subjects for random selection purpose.

The Figure 3.11 and 3.12 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 Figure 3.11, 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 data sets because it has an advantage of a dynamic atom selection and fast convergence.

3.2.6 Computational efficiency

In this section, we demonstrate how the specialised 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 3.13 and 3.14, 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
3.2 Experiments

Figure 3.11: Face recognition against different scales. Number of subjects increases from 68 to 272 with different number of views $M=5$. 
3.2 Experiments

Figure 3.12: Face recognition against different scales. Number of subjects increases from 68 to 272 with different number of views $M=7$. 
3.2 Experiments

Figure 3.13: Computational comparison with dimension $d=64$ for CMU-PIE. The time axis is shown in log scale.

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 3.13 and 3.14, 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. Due to MSRC solving the problems in a matrix form at once, its algorithm could converge.
3.2 Experiments

Figure 3.14: Computational comparison with dimension $d=64$ for Yale B. The time axis is shown in log scale.

quicker than normal SRC which needs to solve problem individually (with some overheads). In addition, the completion time of the MSRC remains almost unchanged when the number of views increases in the experiment on CMU-PIE in log scale, the trend of the increase of the MSRC was less significant than other methods. Overall, the proposed MSRC achieves adequate performance for both CMU-PIE and Yale B, and its computation complexity is less sensitive to the increase of number of views.
In this chapter, we have proposed a mixed-norm sparse representation classification, which has been demonstrated to outperform rivals. 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 are brought. The proposed mixed-norm model achieves the optimal solution for both group-wise and element-wise. It allows us to extract useful shared information when a large pose variation presents in the query set. 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 the numerical algorithm to solve the proposed formulation. And we show the stopping criterion and convergence for the derived algorithm. We also have extensively studied and compared our MSRC with other methods on the CMU-PIE and Yale B data sets. The results indeed show the superior performance of the proposed methods under a different number of views, various dimensionality, view differences, and computational time and scalability.

In the next chapter, we will investigate the classification stage in general SRC framework. We will further exploit the measuring metric for SRC based methods.
Chapter 4

Optimal metric selection for sparse representation based classification

In this chapter, we further expand our existing supervised multi-view face recognition framework. Another limitation of sparse representation based classification will be addressed. As we discussed in Chapter 2, many researches on sparse representation have investigated on how to design a better fitting model to capture the insights of real world data. The first work in this area is Wright et al. (2009) which originally introduce the SRC to face recognition in dealing with extreme variations on lighting and occlusions. As described in Section 2.4.1, the SRC algorithm consists of two parts: the sparse representation via lasso-type $\ell_1$-minimisation and the nearest subspace classification using $\ell_2$ norm. Essentially, the algorithm represents a given face image as a sparse linear combination of other faces in the data set and determines which group of images corresponding to different individuals would give the best fit that determines classification.

Later on, many extensions of SRC have been discussed in the literature such
as Majumdar and Ward (2009); Qiu et al. (2010); Yang and Chu (2010). One of these works extends SRC in the same way as group lasso which proposed by Yuan and Lin (2006) extends the lasso from Majumdar and Ward (2009). The main argument in Majumdar and Ward (2009) is that the sparse solution in SRC does not favour grouping of correlated samples when lasso regularisation is used. Several modifications Wang et al. (2010); Yuan and Yan (2010); Chao et al. (2011) of the group sparse classification (GSC) Majumdar and Ward (2009) are also proposed. We noted that SRC was not directly compared with advanced techniques in the face recognition literature. Subsequent work has examined the performance of SRC relative to other face recognition techniques. One of which is Shi et al. (2011) which demonstrated that sparse representation is not essential for classification at all, and that non-sparse representation performed equally well if not even better.

Given the lack of in-depth studies on the role of sparse representation in classification, there is a research question that we will address in this chapter:

- What cause SRC’s limitation as pointed out by recent studies and if it is possible to improve?

In general sparse representation problem, $\ell_2$ norm represents the fitness of each class by measuring the residual vectors obtained in classification stage. We observe that the key issue of such classification problem lies in the choice of the metric norm of these residual vectors. We will demonstrate that the main limitation of SRC is the non-optimal choice of the $\ell_2$ norm for nearest subspace classification, which does not match with data statistics as these residual values may be considerably non-Gaussian. Based on this assumption, we propose an explicit but effective solution using the $\ell_p$ norm and explain theoretically and numerically why such metric norm would be able to suppress outliers and thus can significantly improve classification performance comparable to the state-of-art algorithms on some challenging data sets.
4.1 Group sparse classification

The chapter is organised as follows. In Section 4.1, we briefly review the group sparse classification. Then, we investigate the limitation of SRC in Section 4.2. Finally, extensive experiments on CMU-PIE, Yale B and Multi-PIE are presented in Section 4.3. This chapter is summarised in Section 4.3.

4.1 Group sparse classification

The was originally proposed by Majumdar and Ward (2009) for face recognition. It assumes all face images from same subject are lie closely in a local linear subspace and the linear spaces for other people are away from. Therefore, GSC treats all face images from one subject as a group. The solution should fulfil two conditions: dense within group and sparse between groups. These property allows it best represents the between and within class difference.

Under the sparsity-induced classification paradigm, the first step is to express a given face image \( a \) as a sparse linear combination of other training images. Assume we have a set of images \( A = [a_1, \ldots, a_n] \), so the approximation of the linear combination is given by \( \hat{x} = \sum_{i=1}^{n} x_i a_i = Ax \) where \( x = [x_1, \ldots, x_n] \) is the coefficient vector. As we discussed previously, the coefficient vector of SRC is sought to be sparse by solving

\[
x = \arg \min_x ||a - Ax||_2^2 + \lambda ||x||_1.
\]  

(4.1)

Now suppose further that the training images are naturally divided into \( g \) groups, our face image set is changed to \( A = [A_1, A_2, \ldots, A_g] \), where each \( A_i \) represents a set of face images which belongs to subject \( i \). In multi-view face recognition, such grouping can be based on the face pose information. Denote \( x_1, x_2, \ldots, x_g \) as the subsets of the vector coefficient \( x \) corresponding to those groups. Then GSC seeks a group sparse solution via

\[
x = \arg \min_x ||a - Ax||_2^2 + \lambda \sum_{i=1}^{g} ||x_i||_2.
\]  

(4.2)
4.2 Optimal metric selection

We note that there are greedy formulations for both SRC and GSC, but the Lagrangian formulation described above can be solved with convex optimisation algorithms that provide better numerical accuracy and stability. As can be seen, the difference between SRC and GSC is essentially the choice of the regularisation term. Group sparse implies sparse, but not the converse.

The other part, which we believe more important, is the nearest subspace classification. This is accomplished by computing the fitness of each individual subspace with respect to the sparse solution. Denote as \( A^k \) the subset of the training images corresponding faces of class \( k \), and \( x^k \) the corresponding coefficient subset. The fitness for class \( k \) is represented by the residual vector \( r_k = a - A^k x^k \). In previous works, the score for such a fitness is computed via the \( \ell_2 \) norm

\[
d_k = \|r_k\|_2,
\]  

(4.3)

and the class with a minimum score is selected.

4.2 Optimal metric selection

In Equation (4.3), the score for the residual vector is calculated by \( \ell_2 \) norm and this has been the practice without any scrutiny. In statistics, it is known that the \( \ell_2 \) norm is optimal in the maximum likelihood sense when the residual values are approximately Gaussian. However, if there are outliers in the residual values or if the empirical distribution of the residual values has heavy tails and departs considerably from Gaussian, such \( \ell_2 \) metric norm would be poor because it can be easily influenced by these bad outliers.

Our numerical investigation of the residual vectors from SRC reveals that it is actually the case here. In other words, we found the original SRC algorithm makes wrong decision for top candidates (with minimal scores) due to the presence of large residual values. *Where do these large values come from?* We note in
4.2 Optimal metric selection

the formulation (4.2), we optimise Gaussian-like criteria for residual from all classes. But in Equation (4.3) we use the solution of all classes to compute the residual of each individual class. While the fitting term in Equation (4.2) promotes Gaussian-like residual values for all classes, there is no such warranty in Equation (4.3).

We observe that a large number of incorrect classification decisions are made by SRC and GSC when the top two candidates are very similar overall but the true candidate is tempered with some large residual values. To illustrate, we show such a scenario in Figure 4.1. We note that the residual values corresponding to the true candidates have some large values but otherwise will fit better than the incorrectly picked candidates by SRC.

The above discussion necessitates a better scheme computing the score \(d_k\) of the residual vectors \(r_k\) other than the \(\ell_2\) norm. Such a scheme must be able to detect and suppress outliers in the residual values so as to being more robust. To achieve this, one might follow robust statistics to design better score or optimise general score functions. In this work, we propose a much simple strategy by using the \(\ell_p\) norm, which is only controlled by one parameter, the order \(p\) of the norm.

Let us discuss why such a metric norm is useful in achieving the goal, especially when \(p < 1\). Consider an oversimplified illustration in Figure 4.2 where \(r_1 = OA\) and \(r_2 = OB\). Here, dimension 2 is where the outliers are present. Clearly, \(\|r_1\|_2 < \|r_2\|_2\) as \(A\) lies in a smaller \(\ell_2\) ball. Suppose that dimension 1 determines the fitness then we would like to select \(r_2\). This is possible for some small \(p\) such that \(B\) lies on a “smaller” \(\ell_p\) ball. Effectively, the \(\ell_p\) ball has suppressed the outliers in dimension 2, and thus \(\|r_1\|_p > \|r_2\|_p\). We note that making \(p\) smaller suppresses large values and effectively amplifies small values. However, making \(p\) too small may suppress too many mid values and hence would reduce the performance. In practice, an optimal \(p\) might be chosen to match with the data statistics, which we will demonstrate in the experiments.
4.2 Optimal metric selection

Figure 4.1: Residual values of top candidates in SRC.
4.2 Optimal metric selection

Figure 4.2: Why does the $\ell_p$ norm suppress outliers?
4.3 Experiments

In this section, extensive experiments have been done on the widely-used CMU-PIE Sim *et al.* (2002), Yale B Georghiades *et al.* (2001) and Multi-PIE Gross *et al.* (2010) face databases. All images are cropped and normalised to $32 \times 32$ pixels with eyes and mouth properly aligned. PCA is then applied to the centralised data to achieve group orthogonal, which improves group lasso’s numerical property. Experiments are splits into two categories: standard experiment to demonstrate the advantage of $\ell_p$ norm and improved version of our MSRC which introduced in previous chapter by combining with utilised $\ell_p$ norm.

4.3.1 $\ell_p$ norm with Group Sparse Classification

Using random sampling, we create training, testing, and validation (for selecting optimal parameters) sets. For GSC, we use pose label information available from the data sets to create groups. We measure the performance over 10 random splits and report the average. For group lasso, we use the advanced ADMM implementation, which is available from [http://www.stanford.edu/~boyd/papers/admm/](http://www.stanford.edu/~boyd/papers/admm/). Comprehensive experiments are delivered in four different perspectives: different $p$ value for $\ell_p$ norm, utilised $\ell_p$ norm under different dimensions, utilised $\ell_p$ norm with various number of groups and comparison with the state-of-the-art methods. All of these experiments are operated on CMU-PIE and Yale B data sets.

Performance dependence on $\ell_p$ norm.

To demonstrate how the $\ell_p$ norm influences GSC’s classification performance, we construct training and testing sets with 2 and 10 images per pose from each subject across all pose variations and vary the $\ell_p$ norm in the range between 0.1
4.3 Experiments

Figure 4.3: Recognition rates with different $\ell_p$ norm
4.3 Experiments

and 3. The average classification performance is shown in Figure 4.3. As can be seen, the GSC with $\ell_p$ norm metric in classification has the highest recognition rate at 95.8%, when $p=0.5$ in CMU-PIE and it achieves the highest recognition rate at $p=0.3$ in Yale B. Whereas, the $\ell_2$ metric cannot reach a satisfactory rate in both CMU-PIE and Yale B. We note particularly that these plots are interesting as they clearly support our claim in the previous sections. These curves show two things: 1. classification with different norm metric can provide various recognition rates; 2. $\ell_2$ norm metric in classification of GSC cannot achieve a fair recognition rate. Thus, these observations confirms our claim that using utilised $\ell_p$ norm metric to compute the residuals can improve the performance of GSC.

Performance dependence on dimensions.

We next investigate how classification depends on the data’s PCA dimension. Both the $\ell_p$ and $\ell_2$ norm metric in GSC are tested. The $p$ value for utilised $\ell_p$ norm is selected by optimising over the validation set. Moreover, the $p$ value ranged from 0.6 to 0.8 could achieve the best performance for most cases.

Figure 4.4 shows classification performance as the dimension is varied. On both CMU-PIE and Yale B, the performance of $\ell_p$ norm is always superior to the $\ell_2$ norm metric under any feature dimensions. When the feature length is low, the utilised $\ell_p$ norm can improve about 9% in Yale B. Once the dimension is above 300, the recognition rates of utilised $\ell_p$ norm are 2% (Yale B) or 6% (CMU-PIE) higher than those of normal $\ell_2$ norm. The highest recognition rates are achieved by $\ell_p$ norm with all 1024 features in both CMU-PIE and Yale B. We also notice that GSC is sensitive to feature dimensions when feature length is less than 300. However, if the feature length above 300, GSC is robust under various dimensions. In addition, the experiments also show that the $\ell_p$ norm can always improve the recognition rates in GSC classification under all dimensions. When the number of feature dimensions is beyond certain point, the $\ell_p$ norm metric can increase
4.3 Experiments

Figure 4.4: Dimension reduction based on PCA
Figure 4.5: With various number of groups the performance significantly.
4.3 Experiments

Performance dependence on group number.

To do this, we merge some groups together based on similarity of pose variations to show the effect of group information. The groups are merged by setting pose labels to the same. Both GSC with utilised $\ell_p$ norm and normal GSC are tested in this section. The resulting curve for CMU-PIE in Figure 4.5 clearly shows that the recognition rates are slightly different among different number of groups. This means group information cannot provide classification advantages. However, GSC with $\ell_p$ norm metric can consistently improve the recognition performance with any number of groups. The findings in this experiment also are consistent with the previous experiment.

Comparison with other state-of-art algorithms.

We next compare GSC with the $\ell_p$ norm metric and other state-of-the-art algorithms in face recognition. The variations of PCA, LDA are considered and also with another popular subspace method Locality preserving projections (LPP). Table 4.1 lists the performance of some advanced methods and our proposed GSC. When we use the $\ell_p$ norm metric in GSC classification, the performance is dramatically improved and achieves the highest recognition rate at 95.17% in CMU-PIE and 94.03% in Yale B.

4.3.2 Improved MSRC with $\ell_p$ norm

To examine how the utilised norm improves our MSRC, we use the Multi-PIE data set to perform three sets of experiment. Face images are mixed from 4 different seasons of all 337 subjects in Multi-PIE to make our experiments more realistic and challenging. We follow the experiment settings in Section 3.2.5. Images in each training set are selected based on following poses $[0^\circ, \pm 30^\circ, \pm 60^\circ, \pm 90^\circ]$. In each pose of each subject, three various images are chosen in the training.
4.3 Experiments

Table 4.1: Classification performance comparison for $\ell_p$ norm

<table>
<thead>
<tr>
<th>Method</th>
<th>CMU PIE</th>
<th>Yale B</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA Turk and Pentland (1991)</td>
<td>55.17 ± 0.78</td>
<td>57.03 ± 1.98</td>
</tr>
<tr>
<td>LPP Niyogi (2004)</td>
<td>89.29 ± 0.63</td>
<td>91.59 ± 1.03</td>
</tr>
<tr>
<td>OLPP Cai et al. (2006)</td>
<td>85.81 ± 0.77</td>
<td>92.64 ± 0.58</td>
</tr>
<tr>
<td>Regularised LDA Cai et al. (2008)</td>
<td>94.88 ± 0.28</td>
<td>93.75 ± 0.85</td>
</tr>
<tr>
<td>Smooth LDA Cai et al. (2007a)</td>
<td>94.47 ± 0.24</td>
<td>89.90 ± 1.55</td>
</tr>
<tr>
<td>SRC Wright et al. (2009)</td>
<td>89.25 ± 0.41</td>
<td>93.19 ± 0.22</td>
</tr>
<tr>
<td>GSC with utilised p</td>
<td><strong>95.17 ± 0.43</strong></td>
<td><strong>94.03 ± 1.30</strong></td>
</tr>
</tbody>
</table>

The plot of Figure 4.6 shows that the performance of MSRC with the standard $\ell_2$ norm and $\ell_p$ norm. As we can see, due to the robustness of our MSRC, both two methods perform quite stable during the increase of the number of subjects. If we compare with Figure 4.7, we notice that both of them are benefited from the increasing in the number of views per pose per subject. However, the utilised $\ell_p$ norm consistently achieved better performance than standard $\ell_2$ norm on all perspectives. Although the performance of $\ell_p$ norm is increased when the number of images raises from 5 views to 7 views, the increments drop slightly. This is because when the number of images per pose per subject increase ($M$ from 5 to 7), better representations are obtained by MSRC. The standard $\ell_2$ norm
4.4 Summary

In this chapter, we have demonstrated that a suitable metric norm is important factor for improving sparse representation based classification. The investigation on the distribution of the residual values for face images shows that there are classification could take the advantage of a more Gaussian-like residual which is led by the better representation. This follows our assumption that the $\ell_p$ norm could provide a more robust solution.

4.4 Summary

Figure 4.6: Multi-View ($M = 5$) with various number of subjects
4.4 Summary

Figure 4.7: Multi-View ($M = 7$) with various number of subjects

heavy tails which make the distribution depart from Gaussian. This leads an unfortunate result for standard $\ell_2$ norm classification. To overcome this issue, we implement an utilised $\ell_p$ norm classification for sparse representation classification problems. This utilised $\ell_p$ norm allows us to suppress outliers and achieve better performance. We extensively studied the utilised $\ell_p$ norm combined with Group representation classification and our mixed-norm representation classification on some traditional data sets. The results show superior performance under various $\ell_p$, dimensionality, scalability, and state-of-the-arts. All the results supported our assumption.
4.4 Summary

By combining the MSRC which we proposed in Chapter (3), we now have a fully functioning and effective multi-view face recognition framework for supervised scenarios. It has shown outstanding performance against other state-of-the-arts. Therefore, we would like to investigate the other part of multi-view face recognition problems. From next chapter, we will discuss the unsupervised face recognition problem.
Chapter 5

Mixed-norm regularised face image clustering

Recognising a human face based on given gallery set has been required by many real-world applications. However, there are more applications that the gallery set may not be available at the time. In forensic investigation, investigators need to analyse the videos which captured by the widely deployed surveillance systems during a critically short time after an incident occurred. However, with the growth of the deployment of surveillance systems, it is infeasible to manually analyse these tens of thousands of frames without pre processing. In this pre processing stage, a clustering method is required to group human faces in these frames into many distinct subjects. The investigators can analyse these clusters to identify the target subject. Under this scenario, multiple unknown human subjects are captured in a video sequence, and their faces are shown in each frame. The goal of this clustering method is labelling each face in every frame with respect to these unknown human subjects.

As we discussed in Chapter 2, face images which were extracted from frames can be represented as high-dimensional vectors. However, these high-dimensional
data usually lie in a low-dimensional structure. Outstanding research is carried out to solve this kind of problems, which is called Sparse Subspace Clustering (Elhamifar and Vidal 2013). SSC is based on the sparse representation principle. It is a powerful principle to deal with the high-dimensional data. By treating each pixel from the images as a feature dimension, it represents each given face image as a sparse linear combination of all other face images. As we know that the face images for one particular subject would have similar features due to the nature facts, then his/her faces should be represented by some similar sparse linear combinations. And these sparse linear combinations are naturally close to each other. In sparse subspace clustering, the sparseness for linear combinations are achieved by using an $\ell_1$-norm regularisation (see Section 2.4.4 for details). Once the sparse linear combinations are found by this $\ell_1$ norm for all images, an affinity graph can be obtained for these combinations. In the end, a spectral clustering is used to cluster the face images into distinct subjects.

Our goal in this chapter is to find a better regularisation on the coefficient matrix in the SSC formulation. In our view, the unstructured sparse solution promoted by the $\ell_1$-norm regularisation sought by SSC is motivated from a theoretical modelling of data distributed randomly into subspaces, known as $\ell_1$ subspace detection property Soltanolkotabi et al. (2012) or self-expressiveness property Wang and Xu (2013). However, it is also noted from Soltanolkotabi et al. (2012) that this property may not hold. In fact, real data are more likely not to follow this ideal modelling perfectly. Since face clustering solutions treat each pixel of a human face as a feature point, the accuracies of face detection methods are crucial. According to Zhang and Zhang (2010), although the face detections have been increasingly used by a large amount of applications in real world, it remains as a challenging task. Due to the variations of face pose and lighting condition, face detection methods can not extract ideal face images from video frames. Thus, it is our view that a better regularisation is needed for more flexibility in modelling the interaction between data points. Here, we argue that in real data, there could
be a group of strong points that are frequently shared between subspaces in the representation. This could be due to a small margin between subspaces among these points. However, the general sparseness of the representation of SSC still partially applies.

Motivated by the success of the mixed norm in the classification problem described in Chapter 3, we introduce a mixed-norm regularisation, which is a convex combination of the $\ell_1$ norm as used in SSC and the block norm $\ell_{2,1}$ which is well known in statistics for its ability to capture group structure Friedman et al. (2010); Huang and Zhang (2010). We note that a similar idea using a combination of the $\ell_1$ and $\ell_{2,1}$ norms was briefly suggested in Elhamifar (2012). However, it is unclear how to solve that formulation efficiently, how to relate the combination to the relative geometry between the clusters, and there is still a lack of comprehensive evaluation on vision problems to provide a sound understanding of this regularisation. This is what we aim to achieve in this chapter. In addition, we also propose an extended formulation that caters for different types of data corruptions. We then derive a provably convergent algorithm based on the alternating direction method of multipliers (ADMM) framework Boyd et al. (2011). We demonstrate the effectiveness of the proposed algorithm on well-known face benchmark data sets Yale B and multi-PIE. We show that the proposed algorithm provides further improvement to the latest version of SSC Elhamifar and Vidal (2013) which currently achieves state-of-the-arts clustering performance on these data sets. This suggests that exploiting group sparse constraints in addition to the regular sparse constraint is crucial for face clustering.

This chapter is organised as follows. The mixed-norm regularised face image clustering framework is described in Section 5.1, including the algorithm derivation based on ADMM framework. Section 5.2 presents the experimental results on the benchmark data sets: Yale B and Multi-PIE. Finally, a summary of the chapter is provided in Section 5.3.
5.1 Mixed-norm sparse subspace clustering

In order to present our mixed-norm regularised sparse subspace clustering approach, we first proceed with some necessary notations. Consider a set of $N$ face images in a $D \times N$ data matrix $A = [a_1, a_2, \ldots, a_N]$, where $a_i \in \mathbb{R}^D$ denotes one of the face images which were captured by video cameras. Since these faces are detected and cropped from video frames, they belong to one of $k$ subjects or they can form into $k$ clusters. When there is an affinity matrix, spectral clustering can perfectly perform the clustering according to this affinity matrix. Thus, the core problem of this challenging face clustering is how to generate an accurate affinity matrix. Recall Section 2.4.4, SSC Elhamifar and Vidal (2013) achieves outstanding performance by recovering a sparse coefficient matrix to create this affinity matrix. It seeks the sparse coefficient matrix $C$ by minimising the following,

$$
\min_C \quad ||C||_1 + \lambda||A - AC||_F^2,
\text{s.t.} \quad \text{diag}(C) = 0.
$$

Here, the regularisation parameter $\lambda$ specifies the desired sparsity and need to be positive. The $\ell_1$ norm is known to promote sparse solutions.

When the data is corrupted, SSC starts with the following modelling

$$
A = AC + E, \quad \text{diag}(C) = 0
$$

where the matrix variable $E$ accounts for the corrupted observations in the data. SSC then recovers $(C, E)$ by solving

$$
\min_{C,E} \quad ||C||_1 + \lambda||E||_1,
\text{s.t.} \quad A = AC + E, \quad \text{diag}(C) = 0.
$$

Since the face detector may not be able to 100% detect the faces in each frame, the data may be incomplete or missing. We can pose this as a special case of the corruption problem. We can fill the missing entries with random numbers.
5.1 Mixed-norm sparse subspace clustering

SSC uses a different approach to fill up the missing entries in which the missing locations of the data point \( \mathbf{a}_i \) and the corresponding rows in \( \mathbf{A} \) are removed. Then, SSC solves (2.17) using the modified \((\mathbf{a}_i^*, \mathbf{A}^*)\). Then, the sparse solution \( \mathbf{c}_i^* \) is used to recover incomplete \( \mathbf{a}_i \) as \( \mathbf{a}_i^{\text{complete}} = \mathbf{A} \mathbf{c}_i^* \) and also for clustering.

5.1.1 Model formulation

Now, we extend the SSC formulation with mixed-norm regularisation which proposed in previous chapter

\[
\min \quad \lambda \| \mathbf{C} \|_2^\gamma + \lambda_s \| \mathbf{E} \|_1,
\]

s.t. \( \mathbf{A} = \mathbf{A} \mathbf{C} + \mathbf{E}, \ \text{diag}(\mathbf{C}) = \mathbf{0} \). \hspace{1cm} (5.2)

Here, the term \( \mathbf{E} \) models the sparse corruption which is the same as in the basic SSC formulation, and that \( \| \mathbf{E} \|_1 \) promotes sparse solution as desired. The major difference to SSC is the regularisation term

\[
\| \mathbf{C} \|_2^\gamma = \gamma \| \mathbf{C} \|_1 + (1 - \gamma) \| \mathbf{C} \|_{2,1},
\]

which is a convex combination of both the \( \ell_1 \) norm \( \| \mathbf{C} \|_1 \) (defined as sum of absolute elements) and \( \ell_{2,1} \) block norm \( \| \mathbf{C} \|_{2,1} \) (defined as sum of the \( \ell_2 \) norm of the rows). Whilst the \( \ell_1 \) norm promotes sparsity at the individual level, the block norm \( \ell_{2,1} \) promotes group sparsity Friedman et al. (2010); Huang and Zhang (2010). In other words, it prefers few dense rows. Our key observation is that real data tend to be neither sparse nor group sparse perfectly. Thus, combining these two desirable properties in the mixed norm gives us the flexibility in modelling the actual properties of the underlying data. Here, the parameter \( \gamma \in [0, 1] \) controls the trade-off between sparseness at the individual level and group level.

The left subplot of Figure 5.1 shows that when \( \gamma = 0 \) the \( \ell_{2,1} \) norm extracts the common pattern effectively: we can clearly see the two disjoint diagonal blocks in the affinity matrix. However, there is still little noise on the off-diagonal entries.
due to the property of $\ell_2$ norm that promotes dense rows. On the other hand, the right subplot of Figure 5.1 shows that when $\gamma = 1$ the $\ell_1$ norm is able to suppress the off-diagonal noise, the coefficients in the main diagonal blocks are too sparse and this implies poor graph connectivity for each cluster. When we combine these two norms, we will have an affinity matrix as shown in the middle of Figure 5.1. It has a denser blocks than that obtained by the $\ell_1$ norm and less noise on the off-diagonal entries than that produced by the $\ell_{2,1}$ norm. We also notice that even if the data are not sorted, this regularisation will still have this property. The reason is the $\ell_2$ norm steps for the $\ell_{2,1}$ are solved individually for each entry in $C$.

Since the mixed norm is a convex combination of the two norms, this regularisation is also convex. We shall demonstrate subsequently that this simple extension can provide a significant improvement over the SSC formulation, which only seeks sparse solution for $C$. In other words, the parameter $\gamma$ allows us to control the clustering performance better with the specific characteristics of the underlying data. Though there is likely an optimal $\gamma$ for each individual problem, a suboptimal $\gamma$ may be used for a wide range of problems, which we shall demonstrate subsequently. This means a satisfactory performance can be achieved without much fine tuning.

5.1.2 Algorithm

Next, we discuss a solution for the proposed model above. We derive our algorithm under the ADMM framework. Note that this problem is convex in $C$ and hence there exists a global minimum.

First, we group the $\ell_1$-norm terms together by introducing $T = [C; E]$ and $P = [A, \theta I]$ where $\theta = \frac{\lambda}{\gamma \lambda}$. Then, it can be shown that the problem is equivalent
5.1 Mixed-norm sparse subspace clustering

Figure 5.1: An illustration of affinity matrix $\mathbf{C}$ based on mixed norm for a two-clusters scenario. For convenient, the data are sorted depending on their class labels. Since data points from the same clusters are highly correlated with each other, they may have common patterns. The $\ell_{2,1}$ norm (when $\gamma = 0$) could extract the shared information from the data points, and the $\ell_1$ norm (when $\gamma = 1$) allows the model to focus on the closest data points. When we merge these two together, this model will select a group of similar data points which are likely to come from the same cluster.

\begin{equation}
\min \quad \lambda_1 \|\mathbf{C}\|_1 + \lambda_2 \|\mathbf{C}\|_{2,1} + \lambda_e \|\mathbf{E}\|_1,
\text{s.t.} \quad \mathbf{A} = \mathbf{AC} + \mathbf{E}, \quad \text{diag} (\mathbf{C}) = 0,
\end{equation}

or, equivalently,

\begin{equation}
\min \quad \lambda_1 \|\mathbf{T}\|_1 + \lambda_2 \|\mathbf{C}\|_{2,1},
\text{s.t.} \quad \mathbf{A} = \mathbf{PT}, \quad \text{diag} (\mathbf{C}) = 0,
\end{equation}

where $\lambda_1 = \gamma \lambda$ and $\lambda_2 = (1 - \gamma)\lambda$.

Now, the problem (5.6) can be decoupled by introducing auxiliary variables $\mathbf{X}$ and $\mathbf{V}$ for $\mathbf{T}$ so that we can solve for $\mathbf{X}, \mathbf{V}$ and $\mathbf{C}$ separately and easily whilst respecting the equality constraint. Here, $\mathbf{V}_1$ and $\mathbf{V}_2$ are the top and bottom blocks of $\mathbf{V}$ that match with the blocks $\mathbf{T}_1$ and $\mathbf{T}_2$ of $\mathbf{T}$ that correspond to $\mathbf{C}$ and $\mathbf{E}$. We need to note that because of $\mathbf{V}$ taking the place of $\mathbf{C}$ in $\lambda_2 \| \mathbf{C} \|_{2,1}$.
only the top part of $V$ is considered when optimising $\|V\|_{2,1}$.

$$
\begin{align*}
\min & \quad \lambda_1 \|T\|_1 + \lambda_2 \|V\|_{2,1}, \\
\text{s.t.} & \quad A = PX, X = T, X = V, \\
& \quad \text{diag}(T_1) = 0, \text{diag}(V_1) = 0.
\end{align*}
$$

(5.6)

With this introduction, we write the augmented Lagrangian as follows

$$
\begin{align*}
\mathcal{L} &= \lambda_1 \|T\|_1 + \lambda_2 \|V\|_{2,1} + \frac{\mu_1}{2} \|A - PX\|_F^2 \\
&\quad + \langle L_1, A - PX \rangle + \frac{\mu_2}{2} \|X - T\|_F^2 \\
&\quad + \langle L_2, X - T \rangle + \frac{\mu_3}{2} \|X - V\|_F^2 + \langle L_3, X - V \rangle, \\
\text{s.t.} & \quad \text{diag}(T_1) = 0, \text{diag}(V_1) = 0.
\end{align*}
$$

(5.7)

where $\langle \bullet, \bullet \rangle$ denotes inner product of two matrices; $L_1, L_2, L_3$ are simply the Lagrangian multipliers for the three constraints they correspond to. For clarity, we note the main variables of the Lagrangian are $X, V$ and $T$. In the ADMM framework, we typically initialise all unknown variables to zero and sequentially update these variables and the Lagrangian multipliers until some convergence criteria, in terms of the primal and dual residuals, are met. In each update, we minimise the Lagrangian by only varying one variable of interest. The ADMM update at iteration $k + 1$ can be derived as follows:

- Updating $X$: it can be easily shown that this reduces to solving a minimisation of a quadratic function in terms of the matrix variable $X$, which has the explicit solution

$$
X^{k+1} = Q^{-1}(\mu_1 P^T A + P^T L_1^k + \mu_2 T^k - L_2^k + \mu_3 V^k - L_3^k),
$$

(5.8)

where $Q = \mu_1 P^T P + (\mu_2 + \mu_3) I$ is a fixed matrix and can be inverted only once in advance. If the dimension of $Q$ is large, it is possible to reduce the computational cost of this update by applying Cholesky factorization on $Q$. It is known from linear algebra that if $Q$ is a positive definite matrix then
5.1 Mixed-norm sparse subspace clustering

it admits the factorisation $Q = LL^T$ and thus $x = Q^{-1}q$ can be efficiently computed by solving $Lx_1 = q$ first, then $L^Tx = x_1$, which can be written as $x = L^T \setminus (L \setminus q)$.

• Updating $T$: it is straightforward to group relevant terms of the Lagrangian and show that the update for $T$ is found from solving

$$
T^{k+1} = \arg \min_T \lambda_1 \left( \frac{\mu_2}{2} \| X^{k+1} - T^k \|_F^2 \\
+ \| T^k \|_1 + \langle L^k_2, X^{k+1} - T^k \rangle \right),
$$

subject to $\text{diag}(T_1) = 0$. This problem is in fact element-wise and the solution is well-known in the ADMM literature as

$$
T^{k+1} = \begin{bmatrix} M_1 - \text{diag}(M_1) \\
M_2 
\end{bmatrix},
$$

where $M_1$ is the top square block of $M$, and the soft-thresholding shrinkage operator is defined element-wise as:

$$
S_\tau^1(a) = (1 - \tau/|a|)_+ a.
$$

• Updating $V$: similarly, we also need to solve the following problem when updating $V$

$$
V^{k+1} = \arg \min_V \left( \frac{\mu_3}{2} \| X^{k+1} - V^{k} \|_F^2 \\
+ \lambda_2 \| V^k \|_{2,1} + \langle L^k_3, X^{k+1} - V^k \rangle \right),
$$

subject to $\text{diag}(V_1) = 0$. Once again, this problem is also well-known in the ADMM literature and the updating for $V$ is defined as follows

$$
V^{k+1} = \begin{bmatrix} N_1 - \text{diag}(N_1) \\
N_2 
\end{bmatrix},
$$

$N^{k+1} = S_{\lambda_2/\mu_3}^2( X^{k+1} + L^k_3/\mu_3),$

91
where $\mathbf{N}_1$ is the top square block of $\mathbf{N}$, and the soft-thresholding shrinkage operator $S^2_\tau$ in the vector form is defined \textit{row-wise} as

$$S^2_\tau (\mathbf{a}) = (1 - \tau / \| \mathbf{a} \|) \mathbf{a}.$$ \hfill (5.14)

- Updating Lagrangian multipliers: we follow the standard procedure in ADMM theory Boyd \textit{et al.} (2011) as follows:

$$L_{1}^{k+1} = L_{1}^{k} + \mu_1 (A - PX_{k+1}),$$
$$L_{2}^{k+1} = L_{2}^{k} + X_{k+1} - T_{k+1},$$
$$L_{3}^{k+1} = L_{3}^{k} + X_{k+1} - V_{k+1}. \tag{5.15}$$

\textbf{Stopping criterion:} iterations are terminated when a maximum number of iteration is reached, or all residuals are sufficiently small. In (5.7), there are three primal residuals:

$$||A - PX_{k+1}||_\infty \leq \epsilon,$$
$$||X_{k+1} - T_{k+1}||_\infty \leq \epsilon,$$
$$||X_{k+1} - V_{k+1}||_\infty \leq \epsilon. \tag{5.16}$$

When these residuals are approaching to 0, it indicates that the current solution fulfills the constraints in (5.7). Moreover, we also have three dual residuals:

$$||X_{k+1} - X_k||_\infty \leq \epsilon,$$
$$||T_{k+1} - T_k||_\infty \leq \epsilon,$$
$$||V_{k+1} - V_k||_\infty \leq \epsilon. \tag{5.17}$$

When these dual residuals are small enough, we can guarantee that an optimal solution (at least a local minimal) is found. In this work, we set maxIter = 50 and $\epsilon = 2 \times 10^{-4}$, which we find an adequate balance between accuracy and speed.
5.1 Mixed-norm sparse subspace clustering

**Final spectral clustering:**

Once the coefficient matrix $C$ is obtained, the next step is to do final clustering. This step involves constructing a balanced affinity graph:

$$\bar{C} = (C + C^T)/2,$$  \hspace{1cm} (5.18)

followed by computing the Laplacian of $\bar{C}$ as

$$L_C = I - D^{-1/2} \bar{C} D^{-1/2},$$  \hspace{1cm} (5.19)

where $I$ is an identity matrix of appropriate dimension. $D$ is a diagonal matrix where $D_{ii} = \sum_{j=1}^{N} \bar{c}_{ij}$. Then we use the smallest eigenvalues of $L_C$ to estimate number of subspaces and the corresponding data points are clustered using $k$-means Kanungo *et al.* (2002) with the respective eigenvectors as starting points Elhamifar and Vidal (2013). Figure 5.2 shows the convergence of proposed algorithm.

**Convergence Analysis:**

According to the ADMM theory Boyd *et al.* (2011), for any given convex optimisation of the following form

$$\text{minimise} \quad f(x) + g(z)$$

subject to \quad $Ex + Fz = c$ \hspace{1cm} (5.20)

where $x \in \mathbb{R}^n, z \in \mathbb{R}^n, E \in \mathbb{R}^{p \times n}, F \in \mathbb{R}^{p \times n},$ and $c \in \mathbb{R}^p$, the ADMM update steps will converge. We show that the convergence property still holds when we extend to more than two variables, which is the case in the proposed algorithm. First, we rewrite the constraints in the formulation (5.7) as follows

$$\begin{bmatrix} 0 \\ I \end{bmatrix} T + \begin{bmatrix} 0 \\ 0 \end{bmatrix} V + \begin{bmatrix} P \\ -I \end{bmatrix} X = \begin{bmatrix} A \\ 0 \end{bmatrix}. \hspace{1cm} (5.21)$$
5.1 Mixed-norm sparse subspace clustering

Figure 5.2: Convergence of proposed method. The objective function is converged after 100 iterations.

Denote $E = [0, I, 0]^T$, $G = [0, 0, I]^T$, $H = [P, I, I]^T$ and $C = [A, 0, 0]^T$. It is easily seen that the above equation can be written in the following form:

$$ET + GV +HX = C.$$  \hfill (5.22)

Then, we can reduce the number of variables by introducing $F = [G \ H]$ and $U = [V \ X]^T$, so that Equation (5.23) becomes

$$ET + FU = C.$$  \hfill (5.23)

Clearly, this equation is expressed in terms of only two variables. Now, we also express the objective function (5.7) in terms of these two variables so that we can use the ADMM theory. Indeed, the objective function is

$$\text{minimise} \quad f(T) + g(U),$$  \hfill (5.24)
Algorithm 5.1 Solving Eq. (5.8) for mixed-norm sparse subspace clustering.

**Input:** $P = [X, \theta I]$, and parameter $\lambda$.

**Output:** Representation matrix $T$.

1. $T, V, A$ and $L_i \leftarrow 0$
2. repeat
3. Fixing $T$ and $V$ to calculate $A$ by Eq.(5.9).
4. Fixing $A$ and $V$ to calculate $T$ by Eq.(5.10) and Eq.(5.11).
5. Fixing $T$ and $A$ to calculate $V$ by Eq.(5.13) and Eq.(5.14).
6. Update Lagrangian multiplier via Eq. (5.16).
7. until All residuals are sufficiently small in Eq.(5.17) and Eq.(5.18)
8. return $T$

where $f(T) = \lambda_1 \|T\|_1$, and $g(U) = \lambda_2 \|V\|_{2,1} + 0 \times X$. Due to the properties of $\ell_1$ and $\ell_{2,1}$ norm, both $f(T)$ and $g(U)$ are convex. Thus, if we let $x = T$ and $z = U$, the convergence of proposed algorithm follows directly from Boyd *et al.* (2011).

### 5.2 Experiments

We compare the proposed method with SSC Elhamifar and Vidal (2013) and other related methods discussed in Chapter 2 on two well-known data sets Yale B and CMU-PIE for face clustering. We denote our method as the mixed-norm SSC.

#### 5.2.1 Single-view clustering

Face clustering is also a very challenging subspace segmentation problem. As posed in Elhamifar and Vidal (2013), the collection of face images for varying illumination and at a fixed pose from multiple people can be approximately mod-
5.2 Experiments

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LSA</th>
<th>SCC</th>
<th>LRR</th>
<th>LRSC</th>
<th>SSC</th>
<th>mixed-norm SSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.78</td>
</tr>
<tr>
<td>Mean</td>
<td>32.80</td>
<td>16.62</td>
<td>2.54</td>
<td>5.32</td>
<td>1.91</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>47.66</td>
<td>7.82</td>
<td>0.78</td>
<td>4.69</td>
<td>1.56</td>
<td>0.00</td>
</tr>
<tr>
<td>3 Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.42</td>
</tr>
<tr>
<td>Mean</td>
<td>52.29</td>
<td>38.16</td>
<td>4.21</td>
<td>8.47</td>
<td>2.78</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>50.00</td>
<td>39.06</td>
<td>2.60</td>
<td>7.81</td>
<td>2.08</td>
<td>0.52</td>
</tr>
<tr>
<td>5 Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.67</td>
</tr>
<tr>
<td>Mean</td>
<td>58.02</td>
<td>58.90</td>
<td>6.90</td>
<td>12.24</td>
<td>4.34</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>56.87</td>
<td>59.38</td>
<td>5.63</td>
<td>11.25</td>
<td>4.06</td>
<td>1.25</td>
</tr>
<tr>
<td>8 Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.64</td>
</tr>
<tr>
<td>Mean</td>
<td>59.19</td>
<td>66.11</td>
<td>14.34</td>
<td>23.72</td>
<td>5.51</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>58.59</td>
<td>64.65</td>
<td>10.06</td>
<td>28.03</td>
<td>5.37</td>
<td>1.95</td>
</tr>
<tr>
<td>10 Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.21</td>
</tr>
<tr>
<td>Mean</td>
<td>60.42</td>
<td>73.02</td>
<td>22.92</td>
<td>30.36</td>
<td>5.94</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>57.50</td>
<td>75.78</td>
<td>23.59</td>
<td>28.75</td>
<td>7.19</td>
<td>2.03</td>
</tr>
</tbody>
</table>

Table 5.1: Clustering error (%) of different algorithms on the Yale B data set without pre-processing the data.

elled as a union of 9D subspaces. This implies that subspace clustering is highly applicable. We provide a comparative evaluation of mixed-norm SSC against state-of-art benchmark methods on Yale B data set. We divided baseline methods into two groups: 1) sparse and low rank solutions and 2) other spectral methods. The sparse and low-rank baseline methods are SSC Elhamifar and Vidal (2013), Low Rank Representation (LRR) Liu et al. (2013a), Low Rank Subspace Clustering (LRSC) Favaro et al. (2011), Spectral Curvature Clustering (SCC) Chen and Lerman (2009) and Local Subspace Affinity (LSA) Yan and Pollefeys (2006). We set $\gamma = 0.8$, whereas $\lambda$ is tuned to optimal.
5.2 Experiments

This data set is created from the original Yale B data set used in face recognition. Here, for face clustering the images from 38 subjects are downsampled to 48 × 42 pixels. The face clustering data is then constructed from taking a certain number of subjects (2, 3, 5, 8, 10) from these 38 individuals. Clustering is ideal if images of each individual belong to a subspace cluster. According to these settings, the misclassification errors are reported in both mean and median of one hundred randomly generated data sets. The results will reflect the overall performance and the statistical stability.

We compare our method against SSC on this problem. As can be seen in Table 5.1, when the number of subjects is 2, the misclassification rate of the proposed method is reduced by a margin of nearly 59% and 85% with respect to the closest best methods SSC and LRSC and the median reaches 0%. This performance suggests that the segmentation is nearly perfect. When the number of subjects increases from 2 to 5, the similar improvement in performance is also noticeable. The proposed method reduces the misclassification rate of SSC (the second best) by about half. Although the improvement become slight smaller, when the number of subjects raises to 8 and 10, mixed-norm SSC still shows significant advantages on stability. The medians of the clustering misclassification rates are only one third of SSC.

5.2.2 Multi-view clustering

In multi-view scenarios, one object may be present under different views, especially on correspondent cameras. Face images are captured in pairs (or groups) for each subject on different cameras by each frame. Although we know these images are belong to the same subjects, we still need to cluster them with images from other frames for this subject. As we mentioned in the previous section, subspace clustering is suitable for fixed pose. We cannot directly cluster multi-view face images. Since the face images in each view lie in a union of 9D subspaces, we can
5.2 Experiments

![Clustering Accuracy Graph](image)

**Figure 5.3:** Mixed-norm SSC with majority voting

label each face by standard face clustering algorithms first, and then use a majority voting to make the final decision. Due to the lack of papers on multi-view, multi-view experiments are conducted by extending single-view state-of-the-art methods. Since the SSC achieves the second best performance in the previous experiment, we take it as the baseline.

Experiment data set is constructed based on Multi-PIE. A system of 15 cameras is used to take images in Multi-PIE. Thirteen cameras are placed at the head height and spaced at 15° intervals. This data set is naturally suitable for our problem, and it allows us to simulate the correspondent face images clustering in different angles. To evaluate the proposed method, we follow the Elhamifar and
5.2 Experiments

Multi-views: $[-30^\circ, 0^\circ, +30^\circ]$  

Figure 5.4: Mixed-norm SSC with majority voting

Vidal (2013)'s setting to create the experiment datasets with different number of subjects as (2, 3, 5, 8, 10). All images are cropped and resampled to $40 \times 40$ pixels for computational convenience. In order to measure the performance about multi-view data, four groups of multi-view settings are chosen: $[\pm 15^\circ, 0^\circ]$, $[\pm 30^\circ, 0^\circ]$, $[\pm 45^\circ, 0^\circ]$ and $[\pm 60^\circ, 0^\circ]$. We use three images for each view. According to these settings, results are reported in clustering accuracy mean of one hundred randomly generated data sets.

As can be seen in Figure 5.3, the SSC with majority voting achieves better performance when the number of subjects is 2. However, when the number of subjects increases, the performance of proposed Mixed-norm SSC increases significantly.
5.3 Summary

Especially, when there are ten subjects, the proposed method outperforms SSC with 2.92% extra. Due to the more accurate modelling, the mixed norm brings robust clustering results for any number of subjects.

When it comes to more complex problems, the variations of face pose increase dramatically in the multi-view settings, the proposed method still provides robust performance. As shown in Figure 5.4, 5.5 and 5.6, the performance of Multi-view SSC drops slightly when the pose variations increase from $\pm 15^\circ$ to $\pm 60^\circ$. Caused by the drawback of ideal modelling of SSC, its performance drops significantly, when the pose variations increase. We notice that when the pose variation reaches $\pm 60^\circ$ (Figure 5.6), the minimum difference between the proposed method and SSC is still as large as 2.4%.

In general, the accurate modelling of proposed method for the real world data causes robust clustering results on the multi-view tasks. It demonstrates outstanding performance in the problems that against with different level of pose variations.

5.3 Summary

This chapter presented a robust sparse subspace clustering method that utilises both group sparse representation and the local information between the data points to solve face clustering problems. The mixed-norm regularisation encourages the sparse representations for multiple objects to be consistent with their subspace assumptions. We claim that the mixed norm could better represent the real-world data than the original $\ell_1$ norm. We also discuss explicit formulation when the data is corrupt and incomplete with an efficient method based on ADMM to solve this formulation. And we also provide the convergence prove and stopping criterion of the proposed algorithm. We present extensive experiments in both standard view and multi-view against other state-of-the-arts on both Yale
5.3 Summary

B and Multi-PIE data sets. The outstanding performance of proposed methods validates our assumptions.

In the next chapter, we will investigate the limitation of existing sparse representation based methods. We can further extend our MSRC to achieve better performance on multi-view face recognition problems.

Figure 5.5: Mixed-norm SSC to majority voting
5.3 Summary

Multi-views: $[-60°, 0°, +60°]$

Clustering Accuracy (%)

Figure 5.6: Mixed-norm SSC to majority voting
Chapter 6

Multi-view face clustering for low frame-rate videos

In this chapter, we propose a new multi-view face clustering method based on sparse subspace representation. The sparse solution promoted by the $\ell_1$-norm regularisation sought by SSC is motivated by self-expressiveness property Wang and Xu (2013). The underlying assumption of $\ell_1$-norm regularisation is that the dictionary needs to be overcomplete Wright *et al.* (2009). As shown in Shi *et al.* (2011), the $\ell_1$-norm regularisation may not be necessary when the dictionary is not over-complete. In fact, the data from low frame rate cameras in the real world are more unlikely to follow this ideal modelling. Therefore, the $\ell_1$ regularisation is not necessary for the proposed method.

Instead of using the mixed norm to exploit the correlated information with majority voting, a more directly approach is proposed. It can efficiently use the complementary information obtained from other views directly. The subspace representation for each view is constructed first. At the same time, a global subspace representation is developed based on the complementary information. Then we build the affinity matrix via this global subspace representation.
6.1 Multi-view face clustering

A provably convergent algorithm for the proposed method is derived by using the alternating direction method of multipliers (ADMM) framework Boyd et al. (2011). We also show the power of the proposed method by using Multi-PIE face data set Gross et al. (2010). In summary, the contributions of this chapter are:

- A multi-view clustering framework by exploiting the complementary information;
- A computationally efficient algorithm for solving the proposed formulation;
- Extensive experiments on the Multi-PIE face data-set.

This chapter is organised as follows. We propose the multi-view subspace clustering with algorithm derivation in Section 6.1. Then, we present extensive experiments on Multi-PIE in Section 6.2. In the end, we provide a summary of this chapter in Section 6.3.

6.1 Multi-view face clustering

In multi-view scenarios, one object may be present under different views, especially on correspondent cameras. Face images are captured in pairs (or groups) for each subject on different cameras by each frame. Although we know these face images are belong to the same subjects, we still need to cluster them with images from other frames for this subject. In normal face clustering, we treat these different views as face poses. In general, SSC Elhamifar and Vidal (2013) can deal with any single pose at a time, but it cannot use all these poses at the same time. This is because the non-linear separability of face poses. Instead of using part of data, we would like to expand the revised SSC into a multi-view version.
6.1 Multi-view face clustering

Figure 6.1: An illustration of normalised affinity matrices based on different views for three clusters problem. Affinity matrix has denser diagonal blocks indicates a better spectral clustering performance. (left) affinity matrix of three clusters for one view. (right) affinity matrix of three clusters for another view. Both matrices only exploit the information from its own view, they are fail to form denser diagonal blocks.

6.1.1 Model formulation

Assume we have a data-set that contains $l$ views for $k$ subjects and is presented as

$$A = \{A_1, A_2, \cdots, A_l\},$$  \hspace{1cm} (6.1)

$$A_i = \{A_i^1, A_i^2, \cdots, A_i^k\},$$ \hspace{1cm} (6.2)
6.1 Multi-view face clustering

Figure 6.2: An illustration of normalised affinity matrix based on proposed model for three clusters problem. Affinity matrix has denser diagonal blocks indicates a better spectral clustering performance. Since the proposed model builds the connection between these two views, the coefficients from two views complete each other to construct a better affinity matrix.

where each \( A_i \in \mathbb{R}^{D \times N_i} \) and \( A \in \mathbb{R}^{D \times N} \). In order to obtain benefits from the multi-view data to clustering these images into \( k \) subjects, we propose the following formulation

\[
\min_{\sum_{i=1}^{l} \{||A_i - A_i C_i||_F^2 + \lambda R(T - C_i)\},} \\
\text{s.t. } \text{diag}(C_i) = 0,
\] (6.3)

where \( R \) is a loss function to measure the disagreement between the coefficient matrices \( T \) and \( C_i \) for each view. The parameter \( \lambda \) is a positive parameter which trades off the reconstruction error and the loss function \( R \). We also set \( \text{diag}(C_i) \) to \( 0 \) to prevent a trivial solution. As discussed at the beginning of this chapter, the \( \ell_1 \) regularisation term as used in the original SSC formulation is not necessary when dealing with low frame rate cameras. Therefore, we remove the \( \ell_1 \) regulariser on \( C \) to reduce the complexity without any degradation in performance. The loss function \( R \) is defined as follows

\[
R(T - C_i) = ||T - C_i||_F^2.
\] (6.4)
6.1 Multi-view face clustering

In this formulation, we maximise the correlation between the correspondent data under each view by using the loss function $R$. This allows us to find a unified latent structure $T$ based on the representation of each view. Since this $T$ is built on top of the coefficient matrices for each view, it can best represent the relationship between each cluster. In addition, a balance between the reconstruction error for individual view and overall is considered when $T$ is involved in finding $C_i$.

As shown in Figure 6.1 and 6.2, this unified latent structure $T$ can extract the complementary information from different views. There are three clusters in Figure 6.2. However, neither view 1 nor view 2 can form three dense diagonal blocks in its affinity matrix. We also notice that although they fail to form the blocks, some of the coefficients are complementary to each other, such as row 5 column 4 in subplots (a) and (b). When we introduce the correlation term, a balance between reconstruction error and complementary information among views is achieved. This is clearly shown in subplot (c), the coefficients appear in both subplots (a) and (b) are enhanced and coefficients appear only in one view are suppressed. Three dense diagonal blocks are clearly separable in proposed affinity matrix. In addition, both reconstruction and correlation term are solved simultaneously, this makes $T$ best to represent all views in both coefficients and reconstruction error.

6.1.2 Algorithm

Since both terms in the proposed formulation are the Frobenius norm, it looks like there is a “closed-form solution” for proposed formulation. However, due to the zero diagonal constraint on $C_i$, it cannot be solved analytically. Thus, we derive an optimisation algorithm to solve the proposed formulation of multi-view subspace representation based on the ADMM framework Boyd et al. (2011). By introducing an auxiliary variable $V_i$ for $C_i$, we can solve the problem in two
6.1 Multi-view face clustering

steps. And this allows us to fit the zero diagonal constraint in:

\[
\min \sum_{i=1}^{l} \{ ||A_i - A_i C_i||_F^2 + \lambda ||T - V_i||_F^2 \},
\]

s.t. \( \text{diag}(C_i) = 0, \text{diag}(V_i) = 0, \)

\[ C_i = V_i. \tag{6.5} \]

With this auxiliary variable, we write the augmented Lagrangian as follows

\[
\mathcal{L} = \sum_{i=1}^{l} \{ ||A_i - A_i C_i||_F^2 + \lambda ||T - V_i||_F^2
\]

\[ + \frac{\mu}{2} ||C_i - V_i||_F^2 + <L_i, C_i - V_i>, \]

s.t. \( \text{diag}(C_i) = 0, \text{diag}(V_i) = 0, \)

\[ \tag{6.6} \]

where \( <\bullet, \bullet> \) denotes the inner product of two matrices, and \( L_i \) is simply the Lagrangian multiplier. The main variables need to be solved are \( C_i, V_i \) and \( T \).

We initialise all these main variables and Lagrangian multiplier \( L_i \) as zero and update them in each step until some stopping criteria are met. In each update, only one variable is solved and all others are fixed. Since \( T \) across all views, we need to find \( C_i \) for each view before we calculate \( T \) in each step. Next, we will present the update rules for each \( C_i, V_i \) and \( L_i \).

- Updating \( C_i \): it can be easily shown the terms related to \( C_i \) form a quadratic function. There is a closed form solution

\[
C_i^{k+1} = Q^{-1}(A_i^T A_i + \mu V_i - L_i^k), \tag{6.7}
\]

where \( Q = (A_i^T A_i + \mu I) \). Since \( Q \) is a fixed matrix, its inverse can be calculated only once in advance. This could significantly improve the efficiency. If \( Q \) is too large, a Cholesky factorisation can be applied on \( Q \) to further reduce the computation cost. According to linear algebra, if \( Q \) is positive definite then it follows \( Q = PP^T \). For any \( a = Q^{-1}q \), we will have \( a = P^T \setminus (P \setminus q) \). Once \( C_i \) is solved, we subject it to \( \text{diag}(C_i) = 0 \). The solution for this is well-known in ADMM literature as

\[
C_i^{k+1} = C_i^k - \text{diag}(C_i^k). \tag{6.8}
\]
Algorithm 6.1 Solving Eq. (6.5) for multi-view subspace representation.

**Input:** Multi-view data-set \( A = \{A_1, A_2, \cdots, A_l\} \), and parameter \( \lambda \).

**Output:** Representation matrix \( T \).

1. \( T, C_i, V_i \) and \( L_i \leftarrow 0 \)
2. repeat
3. \( \text{for } i = 1 \text{ to } l \text{ do} \)
4. Fixing \( V_i \) to calculate \( C_i \) by Eq.(6.7) and Eq.(6.8).
5. Fixing \( C_i \) to calculate \( V_i \) by Eq.(6.9) and Eq.(6.10).
6. Update Lagrangian multiplier via Eq. (6.11).
7. \( \text{end for} \)
8. Solve \( T \) by Eq.(6.12)
9. until Converged as per (6.15) and (6.16)
10. return \( T \)

- Updating \( V_i \): it is another quadratic function when we collect all terms that related to \( V_i \). For solving \( V_i \), we have
  \[
  V_{i}^{k+1} = \frac{\lambda T_{i}^{k} + \mu C_{i}^{k}}{\lambda + \mu}.
  \] (6.9)

  Recall Eq. (6.8), the constraint for \( V_i \) is presented as
  \[
  V_{i}^{k+1} = V_{i}^{k} - \text{diag}(V_{i}^{k}).
  \] (6.10)

- Updating Lagrangian multipliers: we follow standard procedure of ADMM as follows:
  \[
  L_{i}^{k+1} = L_{i}^{k} + C_i - V_i.
  \] (6.11)

Now, we have all the variables \( C_i^{k+1}, V_i^{k+1} \) and \( L_i^{k+1} \) for each view except \( T \). By applying some basic linear algebra, we have

\[
T^{k+1} = \frac{1}{n} \sum_{i} V_i.
\] (6.12)
6.1 Multi-view face clustering

**Final spectral clustering:**

Once the coefficient matrix $C$ is obtained, the next step is to do final clustering. This step involves constructing a balanced affinity graph:

$$\tilde{C} = (C + C^T)/2,$$  \hspace{1cm} (6.13)

followed by computing the Laplacian of $\tilde{C}$ as

$$L_C = I - D^{-1/2} \tilde{C} D^{-1/2},$$  \hspace{1cm} (6.14)

where $I$ is an identity matrix of appropriate dimension. $D$ is a diagonal matrix where $D_{ii} = \sum_{j=1}^{N} \tilde{c}_{ij}$. Then we use the smallest eigenvalues of $L_C$ to estimate number of subspaces and the corresponding data points are clustered using $k$-means Kamungo *et al.* (2002) with the respective eigenvectors as starting points Elhamifar and Vidal (2013).

**Convergence:**

The iterative procedure for solving the problem are terminated when a maximum number of iterations is reached or both the primal and dual residuals are sufficiently small.

- Primal residual. When these residuals are sufficiently close to zero, it indicates that constraints are sufficiently met with the current solution:

$$||C^k_i - V^k_i||_{\infty} \leq \epsilon.$$  \hspace{1cm} (6.15)

- Dual residuals. When these residuals are small enough, an optimal solution is reached:

$$||C^{k+1}_i - C^k_i||_{\infty} \leq \epsilon,$$

$$||V^{k+1}_i - V^k_i||_{\infty} \leq \epsilon,$$

$$||T^{k+1} - T^k||_{\infty} \leq \epsilon.$$  \hspace{1cm} (6.16)
6.2 Experiments

In this section, the proposed method is evaluated and compared with SSC Elhamifar and Vidal (2013) with different multi-view settings on Multi-PIE Gross et al. (2010) data-set. Since images for each view points are captured at the same time by different camera, this data-set is naturally suitable for our problem. To evaluate the proposed method, we use the images from the best representative views between $-60^\circ$ and $+60^\circ$ to construct the experiment data-set. The views out of this range such as $\pm 75^\circ$ and $\pm 90^\circ$ etc were impractical for this thesis. All images are cropped and re-sampled to $40 \times 40$ pixels for computational convenience. An example of Multi-PIE images is shown in Figure (6.3).

Three multi-view settings for SSC are used as baselines: (1) **Best Single View**: SSC is applied on each view independently. Then the performance is reported by a single view which achieves the best clustering accuracy. (2) **Feature Concatenation**: We stack all feature vectors of each view into one, and apply the SSC directly onto it. (3) **Average of Affinity Matrices of Each View**: Constructing an average affinity matrix based on affinity matrices from each view and using this average affinity matrix for spectral clustering in SSC.
6.2 Experiments

6.2.1 Performance against different number of clusters

In this experiment, we split the data into 4 cases as follows:

- $[-15^\circ, 0, +15^\circ]$
- $[-30^\circ, 0, +30^\circ]$
- $[-45^\circ, 0, +45^\circ]$
- $[-60^\circ, 0, +60^\circ]$

The clusters in face clustering represent different people. Thus, data is constructed for variant number of clusters (2, 3, 5, 8, 10) from 337 people in each case. For different number of clusters, we randomly select 30 pairs of subjects to obtain an average performance. A total number of nine face images are used in each cluster, three for each view. The experiment results are shown in Figures (6.4), (6.5), (6.6), (6.7) respectively.

Due to the property of proposed method, it utilise the complementary information between different views. As can be seen in Fig (6.4), the proposed method consistently remains at 100% accurate across all numbers of clusters. It also outperforms other methods in Fig (6.5) and Fig (6.7). Although its performance is slightly worse than the STACK method on 2 and 3 clusters in Fig (6.6), it achieves better results than others on 5, 8 and 10 clusters later. We also notice that the best single view approach takes the second place most of the times. This seems to contradict to our assumption that the multi-view can provide more information.

We have to argue that the best single-view approach is used as a guideline in this paper. To find the best view, the ground-truth is required. However, the ground-truth is not available in any clustering problem. The purpose of using the best single view is to show the best potential performance among different views in theory. Moreover, both average affinity matrix and feature concatenation are naively combine the different views, they will be easily influenced by some bad
6.2 Experiments

Figure 6.4: Clustering accuracy (%) of different algorithms on $[-15^\circ, 0, +15^\circ]$. 

**Single-SSC** is best single view. **Stack-SSC** is feature concatenation. **AVG-SSC** is the average of affinity matrices of each view. **MSSC** is the proposed method.
Figure 6.5: Clustering accuracy (%) of different algorithms on $[-30^\circ, 0, +30^\circ]$.

**Single-SSC** is best single view. **Stack-SSC** is feature concatenation. **AVG-SSC** is the average of affinity matrices of each view. **MSSC** is the proposed method.
Figure 6.6: Clustering accuracy (%) of different algorithms on $[-45^\circ, 0, +45^\circ]$. **Single-SSC** is best single view. **Stack-SSC** is feature concatenation. **AVG-SSC** is the average of affinity matrices of each view. **MSSC** is the proposed method.
6.2 Experiments

Figure 6.7: Clustering accuracy (%) of different algorithms on $[-60^\circ, 0, +60^\circ]$. 

**Single-SSC** is best single view. **Stack-SSC** is feature concatenation. **AVG-SSC** is the average of affinity matrices of each view. **MSSC** is the proposed method.
6.2 Experiments

views. If the clustering perform extremely bad on one view, the final result will decrease. However, because of solving the global affinity matrix $T$ and reconstruction error at same time, this bad view influence is hard to affect proposed method.

6.2.2 Performance against different number of images per cluster

We next examine the clustering performance under different number of images per cluster. In order to achieve this goal, we use 3 and 5 face images for each view. In total, there are 9 and 15 images per cluster respectively. Since the goal of this experiment is to show the performance for different number of images per cluster, we simply consider 2-clusters scenarios. The test data-set is still split into 4 cases from $-60^\circ$ to $+60^\circ$ which are same as the previous experiment. Instead of providing the best single-view results in last section, we report the performance of each single view. **Front** represents face image with $0^\circ$. **View 1** represents face images from $-60^\circ$ to $-15^\circ$. **View 2** represents face images from $+15^\circ$ to $+60^\circ$. All results are reported in average on randomly generated 30 pairs of subjects.

The results in Table (6.1) clearly show that none of the single view can achieve better performance than the proposed method. We notice that there is a large performance decrease for **View 1** and **View 2** when number of images per cluster increases. Meanwhile, the clustering performance for **Front** is improved. This is caused by the underlying assumption of SSC, in which clusters need to be separable in a subspace sense. There are two types of noise in multi-view data-set: one is illumination, the other is slight pose variation. As we discussed in section 1, face images with illumination are linearly separable Basri and Jacobs (2003). Therefore, the illumination is not a big issue. Since the face contour of frontal images is insensitive to slight pose variation, a small pose variation won’t
6.2 Experiments

(a) 3 images per cluster per view

<table>
<thead>
<tr>
<th>View 1</th>
<th>Front</th>
<th>View 2</th>
<th>Stack</th>
<th>Avg</th>
<th>MSSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-15^\circ, 0^\circ, +15^\circ$</td>
<td>95.00</td>
<td>86.11</td>
<td>96.11</td>
<td>97.78</td>
<td>95.00</td>
</tr>
<tr>
<td>$-30^\circ, 0^\circ, +30^\circ$</td>
<td>97.22</td>
<td>86.11</td>
<td><strong>98.33</strong></td>
<td>97.78</td>
<td>96.67</td>
</tr>
<tr>
<td>$-45^\circ, 0^\circ, +45^\circ$</td>
<td>96.67</td>
<td>86.11</td>
<td>90.00</td>
<td><strong>98.33</strong></td>
<td>94.44</td>
</tr>
<tr>
<td>$-60^\circ, 0^\circ, +60^\circ$</td>
<td>93.89</td>
<td>86.11</td>
<td><strong>98.33</strong></td>
<td>93.33</td>
<td>96.11</td>
</tr>
</tbody>
</table>

(b) 5 images per cluster per view

<table>
<thead>
<tr>
<th>View 1</th>
<th>Front</th>
<th>View 2</th>
<th>Stack</th>
<th>Avg</th>
<th>MSSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-15^\circ, 0^\circ, +15^\circ$</td>
<td>93.67</td>
<td>92.33</td>
<td>92.67</td>
<td>94.33</td>
<td>96.00</td>
</tr>
<tr>
<td>$-30^\circ, 0^\circ, +30^\circ$</td>
<td>91.67</td>
<td>92.33</td>
<td>92.00</td>
<td>92.67</td>
<td>94.33</td>
</tr>
<tr>
<td>$-45^\circ, 0^\circ, +45^\circ$</td>
<td>92.00</td>
<td><strong>92.33</strong></td>
<td>87.33</td>
<td>89.67</td>
<td>93.00</td>
</tr>
<tr>
<td>$-60^\circ, 0^\circ, +60^\circ$</td>
<td>93.33</td>
<td>92.33</td>
<td>92.00</td>
<td>86.33</td>
<td>95.33</td>
</tr>
</tbody>
</table>

Table 6.1: Clustering performance (%) under different number of images per clusters.

affect the clustering performance of frontal image. However, the face contour of side view images is sensitive to slight pose variation, a small change in pose will provide a totally different face contour. Therefore, the performance of View 1 and View 2 will decrease when number of images per cluster decreases and Front will increase when number of images per cluster increases. Both feature concatenation and average affinity matrix approaches are also influenced by View 1 and View 2. They fail to compete with proposed method in most of the cases. As shown in Table (6.1), the proposed method does not suffer this problem. It improves about 4% half of the times. When the number of images per cluster increases, its performance is improved slightly except $-45^\circ, 0^\circ, +45^\circ$. Although the performance of proposed method is decreased in $-45^\circ, 0^\circ, +45^\circ$, it is still the best and same with Front. This indicates that the proposed method can take the benefit from Front when number of images per cluster increases.
6.3 Summary

This chapter proposed a multi-view clustering method by exploiting the complementary information between face poses. Instead of processing each individual view of the same subject who under correspondent cameras separately, we integrate the individual processes and seeking the global representation together. A unified latent structure is constructed directly based on subspace representations from each view simultaneously. The global affinity matrix obtained from this unified latent structure can deliver a better clustering performance. We also provide a computationally efficient algorithm for solving the proposed formulation based on ADMM. The experiments on Multi-PIE show that the proposed method achieves outstanding performance which validates our claims. Insights into how the method works have also been discussed.
Chapter 7

Conclusions and future directions

This thesis addresses the problem of multi-view face recognition. We investigated this problem in both supervised and unsupervised manners. In supervised face recognition, we improved sparse representation classification in two different stages: we proposed a new mixed-norm regularisation to better describe multi-view face, we also introduced an utilised $\ell_p$ norm for classification step. Experiments show that the proposed new framework outperformed other state-of-the-arts. In unsupervised face clustering, we extended our mixed norm to sparse subspace clustering, this led to an improvement in multi-view face clustering. Moreover, an multi-view face clustering that finds global affinity matrix based on subspace representation of individual view is proposed. A few conclusions can be draw throughout our research and they cloud inspire some future research directions. These will be discussed in this chapter.
7.1 Mixed-norm sparse representation classification

We can conclude from Chapter 3 that the poor face recognition results for multi-view face problem are usually received by using shared or local information only. Since the multi-view images for same subjects are highly correlated, it is necessary to use the shared information. However, this shared information may not be able to form proper representation when the pose variation is too large or missing. Thus, we presented a mixed-norm regularised sparse representation to deal this type of problems. We combined the advantages of both SRC and JSRC to introduce a trade-off between the $\ell_1$ norm and $\ell_2,1$ norm. Due to the property of this mixed norm, the proposed method decreases the influence of large pose variations and missing pose issues when gains the benefit from the correlation of input images. An optimal representation can be found when some face images with a certain degree of unseen pose variations. Besides, we also investigated the open problem in the robust sparse representation which is using $\ell_1$ norm on the loss function. This allows us to achieve a more robust solution.

To solve this formulation, we derived a provably convergent algorithm based on the powerful alternative directions method of multipliers framework. We also provided the stopping conditions and convergence proofs for our algorithm. We conducted the extensive experiments on CMU-PIE, Yale B and Multi-PIE databases for multi-view face recognition. These experiments compared the proposed method with other state-of-the-arts algorithms in different aspects. The proposed method significantly outperforms the performance of other methods when the number of poses changes from 1 to 7 on both CMU-PIE and Yale B. It also presented an outstanding performance in the experiments under different dimensions and pose differences. Our mixed-norm method also increases as more as 15% improvement from 58.06% (JSRC - the second best) to 73.88% when against
7.2 Utilised metric for sparse representation

with unseen pose. In the end, we demonstrated the recognition rate of proposed method in large scales. The results show that our mixed-norm method is insensitive to the change of the number of subjects. All these experiments validate our claims that the mixed norm can strike the balance between shared and local information to reach an optimal representation for multi-view face recognition.

Based on the above findings, a few future directions are worth addressing. The images that used in this work are gray-scale without any feature extraction. We foresee further improvement by using colour images to obtain more useful information. Also, there are many remarkable pieces of research done on feature extraction which can also be employed to improve the pre processing stage in this work. Moreover, LFW dataset which collected images from multiple data sources has drawn a great attention of face recognition research community. It would be necessary to demonstrate how MSRC performs on this challenging dataset. Last, the trade-off parameter used in the mixed norm is selected via the validation sets. We could use Gaussian distance or other metrics to measure the correlation between face images in the query set to estimate this parameter. This estimation could improve the generalisability of the mixed-norm method.

7.2 Utilised metric for sparse representation

Although the SRC is claimed to be robust in face recognition, it still has some limitations. We have shown in Chapter 4 that the key issue of such classification problem lies in the choice of the metric norm of the residual vectors, which represent the fitness of each class. We found that limitation of the SRC algorithm was caused by the suboptimal choice of the $\ell_2$ norm used in classification stage. Since the residual values of the real world data may not necessarily follow the Gaussian distribution, this theoretical modelling usually led to a suboptimal solution. To overcome this problem, we proposed an explicit solution by using $\ell_p$ norm to re-
7.2 Utilised metric for sparse representation

place the $\ell_2$ norm. And we also explained theoretically and numerically why such metric norm could suppress outliers and improve the classification performance.

Face recognition experiments are carried out in two different scenarios: single-view and multi-view. In single-view experiments, we demonstrated the changes in performance for various $\ell_p$. It clearly showed that the recognition rate peaked at 95.8% on CMU-PIE data set when we that the $p$ to 0.5. In addition, the proposed framework also achieves exceptional performance under different dimensions or group numbers when against with other state-of-the-arts. In multi-view experiments, we evaluate the power of proposed framework combining with our MSRC on Multi-PIE data set. We use as many as 272 subjects to simulate the multi-view problem. As seen in Chapter 4, MSRC with $\ell_p$ norm consistently surpassed the standard MSRC regardless number of views which demonstrated an excellent achievement.

There are two possible branches in future works. On one hand, we could further improve the classification step of SRC by using a cumulative calculation. Instead of selecting the minimum residual to define the label, a group of top candidates is chosen for label assignation. A ranking algorithm is suggested to use for finding a portion of the residuals of these candidates. Since the residuals are not followed the Gaussian distribution when there are outliers, a better solution could be achieved by a segment of the original residual. This should further improve the performance based on $\ell_p$-norm metric. On the other hand, we could adopt this $\ell_p$-norm metric on some subspace methods, such as PCA, LDA etc. In the final step of these subspace methods, they use the nearest neighbour in the Euclidian space which uses $\ell_2$ norm to measure the residual between two data points. If the residuals are not Gaussian distributed, we may obtain a better solution when using $\ell_p$ norm.
The unstructured sparse solution promoted by the SSC is motivated from a theoretical modelling. It assumes that the data is distributed randomly into subspaces which are known as subspace detection property or self-expressiveness. However, the face data in real world doesn’t follow this ideal modelling perfectly. In Chapter 5, we presented a new formulation to model the face data. The combination of $\ell_1$ norm as used in SSC and the block norm $\ell_{2/1}$ was introduced. This mixed norm can meet the requirements for capturing the group structure in the affinity matrix.

A convergent algorithm is derived based on ADMM framework along with its stopping criteria and convergence proofs. Since the original SSC was designed for single-view clustering, we extended both SSC and our multi-view SSC into multi-view clustering with a majority voting mechanism. We found that the performance of our algorithm surpassed other state-of-the-arts significantly when the number of subjects above 5 in single-view clustering. When it came to multi-view clustering, the proposed demonstrated the capabilities to handle with large pose variations. The most significant improvement compared with standard SSC was occurred in 10 subjects with $\pm 60^\circ$ pose variations.

Since the mixed norm which used in SSC is similar to Chapter 3, one of the future directions is exploiting the way of estimate the trade-off parameter. However, the model formulation of this mixed-norm SSC is entirely different with MSRC, so we can’t simply use the same approach to estimate that parameter. This trade-off balances the group representation and local information, thus, a local adjacency graph with k-nearest neighbours can be used to estimate this parameter.
7.4 Multi-view face clustering for low frame videos

In Chapter 6, we found that the underlying assumption of $\ell_1$ regularisation which used in SSC is not suitable for low frame rate cameras with multi-view face clustering task. First, this regularisation fails to exploit the complementary information among the views; second, this regularisation expects an over-complete dictionary which is not available for low frame rate cameras. We proposed a novel multi-view face clustering method by finding a unified latent structure. This latent structure can extract the complementary information from different views. Then a global affinity matrix could be obtained from this latent structure, which delivers a better clustering performance.

We derived this algorithm based on ADMM framework and provided the stopping criteria and convergence proofs. We evaluated the proposed method with three multi-view settings for SSC: Best Single View, Feature Concatenation and Average of Affinity Matrices. Due to a large number of subjects and pose variations, the experiments are conducted on Multi-PIE database. Benefit from the latent structure, the proposed method achieved extraordinary clustering accuracies in a different number of pose groups. Its performance was robust when the number of the clusters is increased. When the number of images per subjects was changed, the proposed method still outperformed all other methods. These extensive experiments validated our claims that the proposed formulation could model the multi-view face data from low frame rates cameras better.

One of the future directions that may be interesting is to introduce a regulariser on the global affinity matrix. We could add our mixed norm in Chapter 5 to the formulation of multi-view subspace clustering. We expect the additional force on forming a group structure in the global affinity matrix could further improve the performance. Another possible future direction is using a weighted matrix to refine the latent structure. A set of weightings can be learned from the
neighbouring information. This could reduce the influence of outliers.
Appendix A

Theoretical analysis of the mixed-norm regularisation formulations

This chapter studies the oracle properties of the proposed mixed-norm in this thesis with random design. The theory for lasso and group lasso are originally developed by Bühlmann and van de Geer (2011), we tailor the theory to our specific case – the mixed norm. In this appendix, we first introduce some notations and terminologies, and then we show the upperbounds of the prediction error and the regularisation error for the proposed mixed norm.

A.1 Notations and inequalities

In this section, we provide some notations and show some inequalities that will be used for the proofs presented subsequently.
A.1 Notations and inequalities

A.1.1 Support and complement sets

Assume we have a coefficient matrix $X$, and that $X$ is sparse, we can decompose it into two parts

$$S = \{i, j : X_{i,j} \neq 0\}, \quad S_c = \{i, j : X_{i,j} = 0\} \quad \text{(A.1)}$$

where $S$ is called the active set, and it contains all the indices of non-zero entries of $X$. $S_c$ is the complement of $S$, it contains all the indices of zero entries of $X$. Therefore, we can rewrite $||\bullet||_1$ as

$$||X||_1 = \sum_{i,j \in S} ||X_{i,j}||_1 + \sum_{i,j \notin S} ||X_{i,j}||_1 = ||X_S||_1 + ||X_{S_c}||_1, \quad \text{(A.2)}$$

and $||\bullet||_{2,1}$ as

$$||X||_{2,1} = \sum_{i,j \in S} ||X_{i,j}||_2 + \sum_{i,j \notin S} ||X_{i,j}||_2 = ||X_S||_{2,1} + ||X_{S_c}||_{2,1}. \quad \text{(A.3)}$$

Now, we can extend this concept to our mixed norm $R(\bullet) = \gamma ||\bullet||_1 + (1-\gamma) ||\bullet||_{2,1}$

$$R(X) = \gamma ||X||_1 + (1-\gamma) ||X||_{2,1}$$

$$= \gamma ||X_S||_1 + \gamma ||X_{S_c}||_1 + (1-\gamma) ||X_{S_1,1}||_{2,1} + (1-\gamma) ||X_{S_{c_1}}||_{2,1}. \quad \text{(A.4)}$$

After re-arranging the terms, we have

$$R(X, S_1, S_{2,1}) = \gamma ||X_S||_1 + (1-\gamma) ||X_{S_{2,1}}||_{2,1},$$

$$R(X, S_{1}, S_{c_1,1}) = \gamma ||X_{S_c}||_1 + (1-\gamma) ||X_{S_{c_2,1}}||_{2,1}. \quad \text{(A.5)}$$

Therefore, when have we $R(X) = \gamma ||X||_1 + (1-\gamma) ||X||_{2,1}$, the support and complement sets for it can be defined by

$$R(X) = R(X, S_1, S_{2,1}) + R(X, S_{c_1, S_{c_2,1}}) = R(X_S) + R(X_{S_c}). \quad \text{(A.6)}$$
A.1 Notations and inequalities

A.1.2 Inequalities

In this part, we would like to show

\[ R(X^0) - R(\tilde{X}) \leq R(X^0 - \tilde{X}), \quad (A.7) \]

where \( R(X^0) = \gamma ||X^0||_1 + (1 - \gamma) ||X^0||_{2,1} \).

According to the reverse triangle inequality

\[ | |a| - |b| | \leq |a - b|, \quad (A.8) \]

we can show the following inequality for \( ||\bullet||_1 \)

\[
||A||_1 - ||B||_1 = \sum_i \sum_j |A_{ij}| - \sum_i \sum_j |B_{ij}|
= \sum_i \sum_j | |A_{ij}| - |B_{ij}| |
\leq \sum_i \sum_j |A_{ij} - B_{ij}|
\leq ||A - B||_1,
\quad (A.9)
\]

and

\[
||A||_{2,1} - ||B||_{2,1} = \sum_j ||A_j||_2 - \sum_j ||B_j||_2
= \sum_i (||A_i||_2 - ||B_i||_2)
\leq \sum_i ||A_i - B_i||_2
\leq ||A - B||_{2,1}
\quad (A.10)
\]

for \( ||\bullet||_{2,1} \). Then, we can derive the following inequality for mixed norm

\[
R(X^0) - R(\tilde{X}) = (\gamma ||X^0||_1 + (1 - \gamma) ||X^0||_{2,1}) - (\gamma ||\tilde{X}||_1 + (1 - \gamma) ||\tilde{X}||_{2,1})
= \gamma ||X^0||_1 - \gamma ||\tilde{X}||_1 + (1 - \gamma) ||X^0||_{2,1} - (1 - \gamma) ||\tilde{X}||_{2,1}
\leq \gamma ||X^0 - \tilde{X}||_1 + (1 - \gamma) ||X^0 - \tilde{X}||_{2,1}
\leq R(X^0 - \tilde{X}), \text{ or } R(\tilde{X} - X^0).
\quad (A.11)
\]
A.1 Notations and inequalities

A.1.3 Inequalities for empirical processes

For any two matrix \( M \) and \( N \) where both \( M \) and \( N \in \mathbb{R}^{p,q} \), we have

\[
|MN| \leq \sum_{i,j}^{p,q} |M_{i,j}N_{i,j}| \leq \sum_{i,j}^{p,q} (\max_{1 \leq i,j \leq p,q} M_{i,j})|N_j|
\]

\[
= (\max_{1 \leq i,j \leq p,q} M_{i,j}) \sum_{i,j} |N_j|, \tag{A.12}
\]

Then, following inequality is obtained

\[
|MN| \leq (\max_{1 \leq i,j \leq p,q} M_{i,j})||N||_1. \tag{A.13}
\]

For any two matrix \( M \) and \( N \) where both \( M \) and \( N \in \mathbb{R}^{p,q} \), by applying Cauchy-Schwarz inequality, we have

\[
|MN| = \sum_{j}^{p} ||M_j||_2 ||N_j||_2 \leq \sum_{j}^{p} ||M_j||_2 ||N_j||_2, \tag{A.14}
\]

and it can be further converted to

\[
\sum_{j}^{p} ||M_j||_2 ||N_j||_2 \leq \sum_{j}^{p} (\max_{1 \leq j \leq p} ||M_j||_2)||N_j||_2
\]

\[
\leq (\max_{1 \leq j \leq p} ||M_j||_2) \sum_{j}^{p} ||N_j||_2. \tag{A.15}
\]

Then we have following inequality

\[
|MN| \leq (\max_{1 \leq j \leq p} ||M_j||_2)||N||_{2,1}. \tag{A.16}
\]

Now, we can use Equation A.13 and Equation A.16 to define the mixed norm \( R(N) \), where \( R(N) = \gamma||N||_1 + (1 - \gamma)||N||_{2,1} \)

\[
|MN| = r|MN| + (1 - \gamma)|MN| \leq r \sum_{i,j}^{p,q} |M_{i,j}N_{i,j}| + (1 - \gamma) \sum_{j}^{p} ||M_jN_j||_2
\]

\[
\leq r(\max_{1 \leq i,j \leq p,q} M_{i,j})||N||_1 + (1 - \gamma)(\max_{1 \leq j \leq p} ||M_j||_2)||N||_{2,1}. \tag{A.17}
\]
A.1 Notations and inequalities

Assume we have a constant $\lambda_0$, such that

$$\max_{1 \leq i, j \leq p, q} M_{i,j} \leq \lambda_0, \text{ and}$$

$$\max_{1 \leq j \leq p} ||M_j||_2 \leq \lambda_0 \quad (A.18)$$

Then, Equation A.17 becomes

$$|MN| \leq \gamma \lambda_0 ||N||_1 + (1 - \gamma) \lambda_0 ||N_j||_{2,1},$$

$$\leq \lambda_0 (\gamma ||N||_1 + (1 - \gamma) ||N||_{2,1}), \quad (A.19)$$

which is equivalent to

$$|MN| \leq \lambda_0 R(N). \quad (A.20)$$
A.2 Theory for the mixed-norm regularisation

The linear model described in Chapter 3 is in the following form:

\[ Y = AX + E \text{ or } E = Y - AX, \]  

(A.21)

where \( Y \) is the matrix of responses, \( A \) is the data matrix, and \( X \) is the matrix of measurement errors. To simplify in this section, we assume the \( A \) is fixed and \( E \) is \( \mathcal{N}(0, \sigma^2 I) \)-distributed. If we further assume there always exists some "true parameter value" \( X^0 \) which are the true solution of the linear model, we will have the following condition with respect to Equation 3.8:

\[
||Y - A\hat{X}||_2^2/n + \lambda R(\hat{X}) \leq ||Y - AX^0||_2^2/n + \lambda R(X^0),
\]  

(A.22)

where \( \hat{X} \) is the solution of the mixed norm and \( R(\bullet) = \gamma \|\bullet\|_1 + (1 - \gamma)\|\bullet\|_{2,1} \).

However, the \( \ell_1 \) loss function is not a quadratic function, we cannot analysis it directly. As an alternative, we show that the \( \ell_1 \) norm for any arbitrary matrix can be bounded by an Frobenius norm

\[
||B||_1^2 = \sum_{i,j;i\neq j} |B_{ij}I_{jj}|^2 
\leq \sum_{i,j} B_{ij}^2 \times \sum_{j} I_{jj} 
= M \sum_{i,j} B_{ij} 
= M ||B||_F^2.
\]  

(A.23)

Therefore, we can approximate the errors for Equation A.22 by solving

\[
||Y - A\hat{X}||_F^2/n + \lambda R(\hat{X}) \leq ||Y - AX^0||_F^2/n + \lambda R(X^0),
\]  

(A.24)

According to the basic inequality (see Section A.2.1), we can always cast the problem as follows

\[
||A(\hat{X} - X^0)||_F^2/n + \lambda R(\hat{X}) \leq 2E^T A(\hat{X} - X^0)/n + \lambda R(X^0).
\]  

(A.25)
A.2 Theory for the mixed-norm regularisation

Let \( M = 2E^T A/n \) and \( N = \hat{X} - X^0 \), we plug \( M \) and \( N \) to Equation A.20. Now, we convert the problem into

\[
||A(\hat{X} - X^0)||^2_F/n + \lambda R(\hat{X}) \leq \lambda_0 R(\hat{X} - X^0) + \lambda R(X^0).
\]  
\( \text{(A.26)} \)

Define \( C > 1 \) as a constant such that \( C\lambda_0 \leq \lambda \), Equation A.26 becomes

\[
C||A(\hat{X} - X^0)||^2_F/n + C\lambda R(\hat{X}) \leq \lambda R(\hat{X} - X^0) + C\lambda R(X^0).
\]  
\( \text{(A.27)} \)

Since \( S_c \) indicates the zero entries in \( X^0 \), the \( R(X^0_{Sc}) = 0 \), this gives us

\[
C||A(\hat{X} - X^0)||^2_F/n + C\lambda R(\hat{X}_{Sc}) + C\lambda R(X^0_{Sc}) \leq \lambda R(\hat{X}_S - X^0_{S}) + \lambda R(X^0_{Sc}) + C\lambda R(X^0_{Sc}).
\]  
\( \text{(A.28)} \)

We apply the triangle inequality on the left hand side to have

\[
C||A(\hat{X} - X^0)||^2_F/n + (C - 1)\lambda R(\hat{X}_{Sc}) \leq (C + 1)\lambda R(\hat{X}_S - X^0_{S})
\]  
\( \text{(A.29)} \)

After re-arranging the terms, we have

\[
C||A(\hat{X} - X^0)||^2_F/n + (C - 1)\lambda R(\hat{X}_{Sc}) \leq (C + 1)\lambda R(\hat{X}_S - X^0_{S}) \leq (C + 1)\lambda R(\hat{X}_S - X^0_{S}) + (C - 1)\lambda R(\hat{X}_S - X^0_{S})
\]  
\( \text{(A.30)} \)

Since \( R(\hat{X} - X^0) = R(\hat{X}_S - X^0_{S}) + R(\hat{X}_{Sc} - X^0_{Sc}) \) and \( X^0_{Sc} = 0 \), then the problem becomes

\[
C||A(\hat{X} - X^0)||^2_F/n + (C - 1)\lambda R(\hat{X}_{Sc}) \leq (C + 1)\lambda R(\hat{X}_S - X^0_{S}) + (C - 1)\lambda R(\hat{X}_S - X^0_{S})
\]  
\( \text{(A.31)} \)

We can merge the terms to have

\[
C||A(\hat{X} - X^0)||^2_F/n + (C - 1)\lambda R(\hat{X} - X^0) \leq 2C\lambda R(\hat{X}_S - X^0_{S})
\]  
\( \text{(A.32)} \)

Since \( R(X) = \gamma||X||_1 + (1 - \gamma)||X||_{2,1} \), we assume for all \( X \) satisfying \( (C - 1)R(X_{Sc}) \leq 2C R(X_S) \). Therefore, this mixed norm will fulfil the requirements.
of compatibility conditions (see Section A.2.2) for both $\ell_1$ (Equation A.55) and $\ell_{2,1}$ (Equation A.58) at the same time

$$2C\lambda R(X_S) = \gamma 2C\lambda \|X_S\|_1 + (1 - \gamma) 2C\lambda \|X_S\|_2,1$$

$$\leq \gamma \frac{2C\lambda \sqrt{s_1}\|AX\|_F}{\phi_1 \sqrt{n}} + (1 - \gamma) \frac{2C\lambda \sqrt{s_{2,1}}\|AX\|_F}{\phi_{2,1} \sqrt{n}}$$

$$\leq \gamma \left(4 \frac{C\lambda \sqrt{s_1}}{2\phi_1} \frac{\|AX\|_F}{\sqrt{n}} \right) + (1 - \gamma) \left(4 \frac{C\lambda \sqrt{s_{2,1}}}{2\phi_{2,1}} \frac{\|AX\|_F}{\sqrt{n}} \right)$$

$$\leq \gamma \left(\|AX\|_F^2/n + (1 - \gamma)\|AX\|_F^2/n + \frac{C^2\lambda^2 s_1}{\phi_1^2} + \frac{1 - \gamma}{\phi_{2,1}^2} + (1 - \gamma)\frac{C^2\lambda^2 s_{2,1}}{\phi_{2,1}^2} \right)$$

$$\leq \|AX\|_F^2/n + C^2\lambda^2 \left(\frac{\gamma s_1}{\phi_1^2} + \frac{1 - \gamma}{\phi_{2,1}^2} s_{2,1} \right). \quad \text{(A.33)}$$

If we plug $2C\lambda R(X_S)$ back to Equation A.32, it becomes

$$C\|A(\hat{X} - X^0)\|^2_F/n + (C - 1)\lambda R(\hat{X} - X^0)$$

$$\leq \|A(\hat{X} - X^0)\|^2_F/n + C^2\lambda^2 \left(\frac{\gamma s_1}{\phi_1^2} + \frac{1 - \gamma}{\phi_{2,1}^2} s_{2,1} \right), \quad \text{(A.34)}$$

and it can be simplified as

$$(C - 1)\|A(\hat{X} - X^0)\|^2_F/n + (C - 1)\lambda R(\hat{X} - X^0)$$

$$\leq C^2\lambda^2 \left(\frac{\gamma s_1}{\phi_1^2} + \frac{1 - \gamma}{\phi_{2,1}^2} s_{2,1} \right). \quad \text{(A.35)}$$

Now, we can see that this theorem presents two results. Firstly, it shows the bound

$$(C - 1)\|A(\hat{X} - X^0)\|^2_F/n \leq C^2\lambda^2 \left(\frac{\gamma s_1}{\phi_1^2} + \frac{1 - \gamma}{\phi_{2,1}^2} s_{2,1} \right) \quad \text{(A.36)}$$

for the prediction error. And secondly, it gives the bound

$$(C - 1)\lambda R(\hat{X} - X^0) \leq C^2\lambda^2 \left(\frac{\gamma s_1}{\phi_1^2} + \frac{1 - \gamma}{\phi_{2,1}^2} s_{2,1} \right) \quad \text{(A.37)}$$

for the mixed norm error. Therefore, the model in Chapter 3 can be bounded by

$$\|A(\hat{X} - X^0)\|^2_F/n + \lambda R(\hat{X} - X^0) \leq 4\lambda^2 \left(\frac{\gamma s_1}{\phi_1^2} + \frac{1 - \gamma}{\phi_{2,1}^2} s_{2,1} \right). \quad \text{(A.38)}$$

where we set $C = 2$.  

134
A.2 Theory for the mixed-norm regularisation

A.2.1 Basic inequality

First, we would like to show the basic inequality for general regression analysis with a regularisation term $R(\bullet)$. Assume we have following condition

$$||Y - AX||_F^2/n + \lambda R(X) \leq ||Y - AX\hat{X}\||_F^2/n + \lambda R(X^0). \tag{A.39}$$

This inequality is equivalent to

$$||A(\hat{X} - X^0)||_F^2/n + \lambda R(\hat{X}) \leq 2E^T A(\hat{X} - X^0)/n + \lambda R(X^0). \tag{A.40}$$

In order to converted this basic inequality in Equation A.40, we let $C = n\lambda R(\hat{X})$ and $D = n\lambda R(X^0)$ in Equation A.39. Now, we have

$$1/n(||Y - A\hat{X}||_F^2 + C) \leq 1/n(||Y - AX^0||_F^2 + D), \tag{A.41}$$

we can rewrite this equation as

$$||Y - A\hat{X}||_F^2 + ||Y - AX^0||_F^2 + C \leq 2||Y - AX^0||_F^2 + D. \tag{A.42}$$

By adding $-2(Y - AX^0)^T(Y - AX^0)$ to both sides of the above equations in order to group two quadratic terms into a single quadratic term, we have

$$||(Y - A\hat{X}) - (Y - AX^0)||_F^2 + C \tag{A.43}$$

on the left side, and it can be further simplified as

$$||A(\hat{X} - X^0)||_F^2 + C. \tag{A.44}$$

Meanwhile, the right hand side of Equation A.42 becomes

$$2||Y - AX^0||_F^2 - 2(Y - AX^0)^T(Y - A\hat{X}) + D. \tag{A.45}$$

Since $||Y - AX^0||_F^2$ can be rewritten as $(Y - AX^0)^T(Y - AX^0)$, the Equation A.42 can be converted to

$$||A(\hat{X} - X^0)||_F^2 + C \leq 2(Y - AX^0)^T(A\hat{X} - AX^0) + D. \tag{A.46}$$
This is equivalent to
\[ \|A(\hat{X} - X^0)\|_F^2 + C \leq 2E^T A(\hat{X} - X^0) + D. \] (A.47)

Then we plug C and D back with \(1/n\)
\[
1/n(||A(\hat{X} - X^0)||_F^2 + C) \leq 1/n(2E^T A(\hat{X} - X^0) + D) \tag{A.48}
\]
\[
1/n(||A(\hat{X} - X^0)||_F^2 + n\lambda R(\hat{X})) \leq 1/n(2E^T A(\hat{X} - X^0) + n\lambda R(X^0)) \tag{A.49}
\]
\[
||A(\hat{X} - X^0)||_F^2/n + \lambda R(\hat{X}) \leq 2E^T A(\hat{X} - X^0)/n + \lambda R(X^0). \tag{A.50}
\]

### A.2.2 Compatibility condition

**Compatibility condition for \(\ell_1\):** For a matrix \(X\) with support \(S\), we have the following inequality for the \(\ell_1\) norm by paying a price \(s\) (cardinality of \(S\))
\[
\|X_S\|_1 \leq \sqrt{s}||X_S||_F. \tag{A.51}
\]

By calling the scaled Gram matrix, we can relate \(||X_S||_F\) to \(||AX_S||_F\)
\[
\hat{\Sigma} = A^T A / n. \tag{A.52}
\]

Since
\[
||AX||_F^2/n = X^T \hat{\Sigma} X, \tag{A.53}
\]
if for some constant \(\phi > 0\), we have
\[
||X_S||_F^2 \leq \frac{||AX||_F^2}{n\phi^2}, \tag{A.54}
\]
which gives us
\[
||X_S||_1 \leq \sqrt{s}||X_S||_F \\
\leq \frac{\sqrt{s}||AX||_F}{\sqrt{n\phi}}, \tag{A.55}
\]
A.2 Theory for the mixed-norm regularisation

Compatibility condition for \( \ell_{2,1} \):

We can apply the similar approach in compatibility condition of \( \ell_1 \) for \( \ell_{2,1} \) term \( ||X_S||_{2,1} \). First, we show the following inequalities to relate the \( \ell_{2,1} \) norm with \( \ell_F \) norm

\[
||X_S||_{2,1} = \sum_i ||X^i_S||_2 = \begin{bmatrix} ||X^1_S||_2 \\ ||X^2_S||_2 \\ \vdots \\ ||X^i_S||_2 \end{bmatrix} \leq \sqrt{s} \begin{bmatrix} ||X^1_S||_2 \\ ||X^2_S||_2 \\ \vdots \\ ||X^i_S||_2 \end{bmatrix} = \sqrt{s}||X_S||_F, \quad (A.56)
\]

where \( s \) is the cardinality of support set \( S \).

Then, we introduce a constant \( \phi \) with Gram matrix \( \hat{\Sigma} = A^T A / n \) to \( \sqrt{s}||X||_F \), we have

\[
\sqrt{s}||X_S||_F \leq \frac{\sqrt{s}||AX||_F}{\sqrt{n}\phi}, \quad (A.57)
\]

Now, we can recall Equation A.56 to show the upperbound of \( ||X_S||_{2,1} \),

\[
||X_S||_{2,1} \leq \frac{\sqrt{s}||AX||_F}{\sqrt{n}\phi}. \quad (A.58)
\]
Bibliography


Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.
Index

\( \ell_p \) ball, 69
\( \ell_{2,1} \) norm, 38
ADMM, 27
affinity, 26
alternating direction, 28
CMU-PIE, 7
corruption problem, 26
cropped image, 8
dictionary, 8
dynamic, 8
dynamic active set, 24
Eigenfaces, 11
face clustering, 4
Fisherface, 12
gallery set, 8
group sparse classification, 67
group sparsity, 86
joint dynamic sparse representation classification, 23
joint sparse representation, 21
Laplacian, 26
linear discriminate analysis, 12
mixed norm, 6
multi view, 3
Multi-PIE, 7
multi-view face recognition, 4
nearest neighbour, 11
nearest subspace, 11
orthogonal matching pursuit, 24
poses, 3
principal component analysis, 11
query set, 8
residual vectors, 21
scaled dual variable, 28
self-expressiveness property, 83
shared information, 3
soft-thresholding shrinkage operator, 44
sparse representation classification, 20
sparse subspace clustering, 25
spectral clustering, 26
Yale B, 7
# AUTHORSHIP STATEMENT

## 1. Details of publication and executive author

<table>
<thead>
<tr>
<th>Title of Publication</th>
<th>Publication details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-view Subspace Clustering for Face Images</td>
<td>DICTA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name of executive author</th>
<th>School/institute/Division if based at Deakin; Organisation and address if non-Deakin</th>
<th>Email or phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xin Zhang</td>
<td>Centre for Pattern Recognition and Data Analytics</td>
<td><a href="mailto:contact@xinzheng.me">contact@xinzheng.me</a></td>
</tr>
</tbody>
</table>

## 2. Inclusion of publication in a thesis

<table>
<thead>
<tr>
<th>Is it intended to include this publication in a higher degree by research (HDR) thesis?</th>
<th>Yes</th>
<th>If Yes, please complete Section 3 If No, go straight to Section 4</th>
</tr>
</thead>
</table>

## 3. HDR thesis author's declaration

<table>
<thead>
<tr>
<th>Name of HDR thesis author if different from above. (If the same, write &quot;as above&quot;)</th>
<th>School/institute/Division if based at Deakin</th>
<th>Thesis title</th>
</tr>
</thead>
<tbody>
<tr>
<td>As above</td>
<td>Centre for Pattern Recognition and Data Analytics</td>
<td>Sparse representation for face images</td>
</tr>
</tbody>
</table>

If there are multiple authors, give a full description of HDR thesis author’s contribution to the publication (for example, how much did you contribute to the conception of the project, the design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)

Conception of the project, design of methodology, experimental protocol, analysis, drafting the manuscript.

I declare that the above is an accurate description of my contribution to this paper, and the contributions of other authors are as described below.

Signature and date: 12 - May - 2016

## 4. Description of all author contributions

<table>
<thead>
<tr>
<th>Name and affiliation of author</th>
<th>Contribution(s) (for example, conception of the project, design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duc-Son Pham</td>
<td>Conception of the project, revising and commenting on manuscript</td>
</tr>
<tr>
<td>Dinh Phung</td>
<td>Revising and commenting on manuscript</td>
</tr>
<tr>
<td>Wanquan Liu</td>
<td>Revising and commenting on manuscript</td>
</tr>
<tr>
<td>Svetla Venkatesh</td>
<td>Revising and commenting on manuscript</td>
</tr>
</tbody>
</table>
5. Author Declarations

I agree to be named as one of the authors of this work, and confirm:

i. that I have met the authorship criteria set out in the Deakin University Research Conduct Policy,

ii. that there are no other authors according to these criteria,

iii. that the description in Section 4 of my contribution(s) to this publication is accurate,

iv. that the data on which these findings are based are stored as set out in Section 7 below.

If this work is to form part of an HDR thesis as described in Sections 2 and 3, I further

v. consent to the incorporation of the publication into the candidate’s HDR thesis submitted to Deakin University and, if the higher degree is awarded, the subsequent publication of the thesis by the university (subject to relevant Copyright provisions).

<table>
<thead>
<tr>
<th>Name of author</th>
<th>Signature*</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duc-Son Pham</td>
<td>Signature Redacted by Library</td>
<td>12-May-2016</td>
</tr>
<tr>
<td>Dinh Phung</td>
<td>Signature Redacted by Library</td>
<td>12-05-16</td>
</tr>
<tr>
<td>Wanquan Liu</td>
<td>Signature Redacted by Library</td>
<td>12-05-2016</td>
</tr>
<tr>
<td>Svetla Verskatesh</td>
<td>Signature Redacted by Library</td>
<td>12-05-16</td>
</tr>
</tbody>
</table>

6. Other contributor declarations

I agree to be named as a non-author contributor to this work.

<table>
<thead>
<tr>
<th>Name and affiliation of contributor</th>
<th>Contribution</th>
<th>Signature* and date</th>
</tr>
</thead>
</table>

* If an author or contributor is unavailable or otherwise unable to sign the statement of authorship, the Head of Academic Unit may sign on their behalf, noting the reason for their unavailability, provided there is no evidence to suggest that the person would object to being named as author.

7. Data storage

The original data for this project are stored in the following locations. (The locations must be within an appropriate institutional setting. If the executive author is a Deakin staff member and data are stored outside Deakin University, permission for this must be given by the Head of Academic Unit within which the executive author is based.)

<table>
<thead>
<tr>
<th>Data format</th>
<th>Storage Location</th>
<th>Date lodged</th>
<th>Name of custodian if other than the executive author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital</td>
<td>N:\ResearchGroups\PraDa\</td>
<td>10/05/2016</td>
<td>Dinh Phung</td>
</tr>
</tbody>
</table>

This form must be retained by the executive author, within the school or institute in which they are based.

If the publication is to be included as part of an HDR thesis, a copy of this form must be included in the thesis with the publication.
**AUTHORSHIP STATEMENT**

1. Details of publication and executive author

<table>
<thead>
<tr>
<th>Title of Publication</th>
<th>Publication details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual Object Clustering via Mixed-Norm Regularization</td>
<td>WACV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name of executive author</th>
<th>School/Institute/Division if based at Deakin; Organisation and address if non-Deakin</th>
<th>Email or phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xin Zhang</td>
<td>Centre for Pattern Recognition and Data Analytics</td>
<td><a href="mailto:contact@xinzhang.me">contact@xinzhang.me</a></td>
</tr>
</tbody>
</table>

2. Inclusion of publication in a thesis

<table>
<thead>
<tr>
<th>Is it intended to include this publication in a higher degree by research (HDR) thesis?</th>
<th>Yes</th>
</tr>
</thead>
</table>

3. HDR thesis author’s declaration

<table>
<thead>
<tr>
<th>Name of HDR thesis author if different from above, (If the same, write “as above”)</th>
<th>School/Institute/Division if based at Deakin</th>
<th>Thesis title</th>
</tr>
</thead>
<tbody>
<tr>
<td>As above</td>
<td>Centre for Pattern Recognition and Data Analytics</td>
<td>Sparse representation for face images</td>
</tr>
</tbody>
</table>

If there are multiple authors, give a full description of HDR thesis author’s contribution to the publication (for example, how much did you contribute to the conception of the project, the design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)

Conception of the project, design of methodology, experimental protocol, analysis, drafting the manuscript

I declare that the above is an accurate description of my contribution to this paper, and the contributions of other authors are as described below.

| Signature and date | 12 - May - 2016 |

4. Description of all author contributions

<table>
<thead>
<tr>
<th>Name and affiliation of author</th>
<th>Contribution(s) (for example, conception of the project, design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duc-Son Pham</td>
<td>Conception of the project, revising and commenting on manuscript</td>
</tr>
<tr>
<td>Dinh Phung</td>
<td>Revising and commenting on manuscript</td>
</tr>
<tr>
<td>Wanquan Liu</td>
<td>Revising and commenting on manuscript</td>
</tr>
<tr>
<td>Budhadiiya Saha</td>
<td>Literature review and data collection</td>
</tr>
<tr>
<td>Svetla Venkatesh</td>
<td>Revising and commenting on manuscript</td>
</tr>
</tbody>
</table>
5. Author Declarations
I agree to be named as one of the authors of this work, and confirm:
i. that I have met the authorship criteria set out in the Deakin University Research Conduct Policy,
ii. that there are no other authors according to these criteria,
iii. that the description in Section 4 of my contribution(s) to this publication is accurate,
iv. that the data on which these findings are based are stored as set out in Section 7 below.

If this work is to form part of an HDR thesis as described in Sections 2 and 3, I further
v. consent to the incorporation of the publication into the candidate’s HDR thesis submitted to Deakin
University and, if the higher degree is awarded, the subsequent publication of the thesis by the
university (subject to relevant Copyright provisions).

<table>
<thead>
<tr>
<th>Name of author</th>
<th>Signature*</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dac-Son Pham</td>
<td>Signature Redacted by Library</td>
<td>12-May-2016</td>
</tr>
<tr>
<td>Dinh Phung</td>
<td>Signature Redacted by Library</td>
<td>12-05-16</td>
</tr>
<tr>
<td>Wanquan Liu</td>
<td>Signature Redacted by Library</td>
<td>12-05-2016</td>
</tr>
<tr>
<td>Sreetha Venkatesh</td>
<td>Signature Redacted by Library</td>
<td>12-05-16</td>
</tr>
<tr>
<td>Budhaditya Saha</td>
<td>Signature Redacted by Library</td>
<td>12-05-2016</td>
</tr>
</tbody>
</table>

6. Other contributor declarations
I agree to be named as a non-author contributor to this work.

<table>
<thead>
<tr>
<th>Name and affiliation of contributor</th>
<th>Contribution</th>
<th>Signature* and date</th>
</tr>
</thead>
</table>

* if an author or contributor is unavailable or otherwise unable to sign the statement of authorship, the Head of
Academic Unit may sign on their behalf, noting the reason for their unavailability, provided there is no evidence to
suggest that the person would object to being named as author

7. Data storage
The original data for this project are stored in the following locations. (The locations must be within an appropriate
institutional setting. If the executive author is a Deakin staff member and data are stored outside Deakin
University, permission for this must be given by the Head of Academic Unit within which the executive author is
based.)

<table>
<thead>
<tr>
<th>Data format</th>
<th>Storage Location</th>
<th>Date lodged</th>
<th>Name of custodian if other than the executive author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital</td>
<td>N:\ResearchGroups\PRaDA\</td>
<td>10/05/2016</td>
<td>Dinh Phung</td>
</tr>
</tbody>
</table>

This form must be retained by the executive author, within the school or institute in which they are based.
If the publication is to be included as part of an HDR thesis, a copy of this form must be included in the thesis
with the publication.
# AUTHORSHIP STATEMENT

## 1. Details of publication and executive author

<table>
<thead>
<tr>
<th>Title of Publication</th>
<th>Publication details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed-norm sparse representation for multi view face recognition</td>
<td>Pattern Recognition</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name of executive author</th>
<th>School/Institute/Division if based at Deakin; Organisation and address if non-Deakin</th>
<th>Email or phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xin Zhang</td>
<td>Centre for Pattern Recognition and Data Analytics</td>
<td><a href="mailto:contact@xinzhang.me">contact@xinzhang.me</a></td>
</tr>
</tbody>
</table>

## 2. Inclusion of publication in a thesis

<table>
<thead>
<tr>
<th>Is it intended to include this publication in a higher degree by research (HDR) thesis?</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>If Yes, please complete Section 3 If No, go straight to Section 4.</td>
</tr>
</tbody>
</table>

## 3. HDR thesis author's declaration

<table>
<thead>
<tr>
<th>Name of HDR thesis author if different from above. (If the same, write “as above”)</th>
<th>School/Institute/Division if based at Deakin</th>
<th>Thesis title</th>
</tr>
</thead>
<tbody>
<tr>
<td>As above</td>
<td>Centre for Pattern Recognition and Data Analytics</td>
<td>Sparse representation for face images</td>
</tr>
</tbody>
</table>

If there are multiple authors, give a full description of HDR thesis author's contribution to the publication (for example, how much did you contribute to the conception of the project, the design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)

I declare that the above is an accurate description of my contribution to this paper, and the contributions of other authors are as described below.

Signature and date: 12 May 2016

## 4. Description of all author contributions

<table>
<thead>
<tr>
<th>Name and affiliation of author</th>
<th>Contribution(s) (for example, conception of the project, design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duc-Son Pham</td>
<td>Conception of the project, revising and commenting on manuscript</td>
</tr>
<tr>
<td>Svetla Venkatesh</td>
<td>Revising and commenting on manuscript</td>
</tr>
<tr>
<td>Wanquan Liu</td>
<td>Revising and commenting on manuscript</td>
</tr>
<tr>
<td>Dinh Phung</td>
<td>Revising and commenting on manuscript</td>
</tr>
</tbody>
</table>
5. Author Declarations

I agree to be named as one of the authors of this work, and confirm:

i. that I have met the authorship criteria set out in the Deakin University Research Conduct Policy,

ii. that there are no other authors according to these criteria,

iii. that the description in Section 4 of my contribution(s) to this publication is accurate,

iv. that the data on which these findings are based are stored as set out in Section 7 below.

If this work is to form part of an HDR thesis as described in Sections 2 and 3, I further

v. consent to the incorporation of the publication into the candidate’s HDR thesis submitted to Deakin

University and, if the higher degree is awarded, the subsequent publication of the thesis by the

university (subject to relevant Copyright provisions).

<table>
<thead>
<tr>
<th>Name of author</th>
<th>Signature*</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duc-Son Pham</td>
<td>Signature Redacted by Library</td>
<td>12 - May - 2016</td>
</tr>
<tr>
<td>Svetla Venkatesh</td>
<td>Signature Redacted by Library</td>
<td>12 - 05 - 16</td>
</tr>
<tr>
<td>Wanquan Liu</td>
<td>Signature Redacted by Library</td>
<td>12 - 05 - 2016</td>
</tr>
<tr>
<td>Dinh Phung</td>
<td>Signature Redacted by Library</td>
<td>12 - 05 - 16</td>
</tr>
</tbody>
</table>

6. Other contributor declarations

I agree to be named as a non-author contributor to this work.

<table>
<thead>
<tr>
<th>Name and affiliation of contributor</th>
<th>Contribution</th>
<th>Signature* and date</th>
</tr>
</thead>
</table>

* If an author or contributor is unavailable or otherwise unable to sign the statement of authorship, the Head of Academic Unit may sign on their behalf; noting the reason for their unavailability, provided there is no evidence to suggest that the person would object to being named as author

7. Data storage

The original data for this project are stored in the following locations. (The locations must be within an appropriate institutional setting. If the executive author is a Deakin staff member and data are stored outside Deakin University, permission for this must be given by the Head of Academic Unit within which the executive author is based.)

<table>
<thead>
<tr>
<th>Data format</th>
<th>Storage Location</th>
<th>Date lodged</th>
<th>Name of custodian if other than the executive author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital</td>
<td>N:\ResearchGroups\PraDa\</td>
<td>10/05/2016</td>
<td>Dinh Phung</td>
</tr>
</tbody>
</table>

This form must be retained by the executive author, within the school or institute in which they are based.

If the publication is to be included as part of an HDR thesis, a copy of this form must be included in the thesis with the publication.
# AUTHORSHIP STATEMENT

## 1. Details of publication and executive author

<table>
<thead>
<tr>
<th>Title of Publication</th>
<th>Publication details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal metric selection for improved multi-pose face recognition with group information</td>
<td>ICPR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name of executive author</th>
<th>School/Institute/Division if based at Deakin; Organisation and address if non-Deakin</th>
<th>Email or phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xin Zhang</td>
<td>Centre for Pattern Recognition and Data Analytics</td>
<td><a href="mailto:contact@xinzhanglei.me">contact@xinzhanglei.me</a></td>
</tr>
</tbody>
</table>

## 2. Inclusion of publication in a thesis

<table>
<thead>
<tr>
<th>Is it intended to include this publication in a higher degree by research (HDR) thesis?</th>
<th>Yes</th>
<th>If Yes, please complete Section 3 If No, go straight to Section 4.</th>
</tr>
</thead>
</table>

## 3. HDR thesis author’s declaration

<table>
<thead>
<tr>
<th>Name of HDR thesis author if different from above. (If the same, write “as above”)</th>
<th>School/Institute/Division if based at Deakin</th>
<th>Thesis title</th>
</tr>
</thead>
<tbody>
<tr>
<td>As above</td>
<td>Centre for Pattern Recognition and Data Analytics</td>
<td>Sparse representation for face images</td>
</tr>
</tbody>
</table>

If there are multiple authors, give a full description of HDR thesis author’s contribution to the publication (for example, how much did you contribute to the conception of the project, the design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)

Conception of the project, design of methodology, experimental protocol, analysis, drafting the manuscript, revising the manuscript.

I declare that the above is an accurate description of my contribution to this paper, and the contributions of other authors are as described below.

Signature and date: Signature Redacted by Library

12 May 2016

## 4. Description of all author contributions

<table>
<thead>
<tr>
<th>Name and affiliation of author</th>
<th>Contribution(s) (for example, conception of the project, design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dac-Son Pham</td>
<td>Conception of the project, revising and commenting on manuscript</td>
</tr>
<tr>
<td>Wanquan Liu</td>
<td>Revising and commenting on manuscript</td>
</tr>
<tr>
<td>Svetla Venkatesh</td>
<td>Revising and commenting on manuscript</td>
</tr>
</tbody>
</table>
5. Author Declarations

I agree to be named as one of the authors of this work, and confirm:

i. that I have met the authorship criteria set out in the Deakin University Research Conduct Policy,

ii. that there are no other authors according to these criteria,

iii. that the description in Section 4 of my contribution(s) to this publication is accurate,

iv. that the data on which these findings are based are stored as set out in Section 7 below.

If this work is to form part of an HDR thesis as described in Sections 2 and 3, I further

v. consent to the incorporation of the publication into the candidate’s HDR thesis submitted to Deakin University and, if the higher degree is awarded, the subsequent publication of the thesis by the university (subject to relevant Copyright provisions).

<table>
<thead>
<tr>
<th>Name of author</th>
<th>Signature*</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duc-Son Pham</td>
<td>Signature Redacted by Library</td>
<td>12-May-2016</td>
</tr>
<tr>
<td>Wangqun Liu</td>
<td>Signature Redacted by Library</td>
<td>12-May-2016</td>
</tr>
<tr>
<td>Svetla Venkatesh</td>
<td>Signature Redacted by Library</td>
<td>12-05-2016</td>
</tr>
</tbody>
</table>

6. Other contributor declarations

I agree to be named as a non-author contributor to this work.

<table>
<thead>
<tr>
<th>Name and affiliation of contributor</th>
<th>Contribution</th>
<th>Signature* and date</th>
</tr>
</thead>
</table>

* If an author or contributor is unavailable or otherwise unable to sign the statement of authorship, the Head of Academic Unit may sign on their behalf, noting the reason for their unavailability, provided there is no evidence to suggest that the person would object to being named as author.

7. Data storage

The original data for this project are stored in the following locations. (The locations must be within an appropriate institutional setting. If the executive author is a Deakin staff member and data are stored outside Deakin University, permission for this must be given by the Head of Academic Unit within which the executive author is based.)

<table>
<thead>
<tr>
<th>Data format</th>
<th>Storage Location</th>
<th>Date lodged</th>
<th>Name of custodian if other than the executive author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital</td>
<td>N:\ResearchGroups\PRD\</td>
<td>10/05/2016</td>
<td>Dinh Phung</td>
</tr>
</tbody>
</table>

This form must be retained by the executive author, within the school or institute in which they are based.

If the publication is to be included as part of an HDR thesis, a copy of this form must be included in the thesis with the publication.