DATA DRIVEN FACIAL IMAGE SYNTHESIS FROM POOR QUALITY LOW RESOLUTION IMAGE

LOKE YUAN REN

(B.Eng.(Hons.),NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2013

Acknowledgements

I would like to express my sincere thanks to my former thesis supervisor, Prof. Surendra Ranganath, for his constant guidance and support on my first half of project. I also would like to thank him for his valuable efforts on improving my technical writings and presentation skills.

I am very grateful to Dr. Tan Ping for his invaluable advice and patience throughout the course of my research. I am very indebted A/ Prof. Ashraf Ali Bin Mohamed Kassim and Prof. Y.V. Venkatesh for his motivation and inspiration on my second half of the project.

A big thank to everyone who worked in the Vision and Image Processing Laboratory and is working in the new laboratory, Vision and Machine Learning Laboratory. I would like to thank my friends and colleagues for unreservedly sharing your knowledge with me.

I am thankful to our Laboratory Technogist, Mr. Francis Hoon Keng Chuan for his technical assistance with softwares and hardwares.

Lastly but not the least, I am extremely grateful to my parents for loving me and believing in me.

Contents

Summary **vi**

List of Tables **x**

List of Figures **xi**

1 Introduction **1**

1.1 Background 1

1.1.1 Image Superresolution 2

1.1.2 Facial Image Superresolution 4

1.1.3 Image Inpainting 5

1.1.4 Facial Structure Recovery from Motion 6

1.2 Motivation 6

1.3 Applications 8

1.3.1 Gaming 8

1.3.2 Multimedia, Entertainment, Computer Graphics . . . 9

1.3.3 Social Networking 9

1.3.4 Face Recognition 10

1.4 Thesis Contributions 11

1.4.1 Facial Structure from Motion 11

1.4.2 Facial Image Inpainting 11
1.4.3 Image Superresolution on Generic Face 12
1.4.4 Image Retrieval of Facial Expression and Pose Estimation 13
1.4.5 Image Superresolution on Specific Face 14
1.5 Thesis Organization 15

2 Literature Review **16**
2.1 Structure Recovery 16
2.1.1 Shape From Silhouette 17
2.1.2 Structure from Motion 18
2.2 Facial Action Image Retrieval 22
2.2.1 Feature-based Tracking Approach 23
2.2.2 Model-based Tracking Approach 24
2.2.3 Facial Feature Detection and Representation 25
2.3 Image Inpainting 26
2.4 Image Superresolution 29
2.4.1 Nonlinear Interpolation Approach 31
2.4.2 Frequency Domain Approach 32
2.4.3 Regularization Approach 33
2.4.4 Learning Approach 34

3 Batch Algorithm for Facial Structure From Motion with Additional Shape Constraints **38**
3.1 Introduction 38
3.2 Overview of Factorization Algorithm for Non-rigid SFM . . . 40
3.3 Batch Algorithm Using Matrix Partitioning 42

3.4 Non-linear Shape Constraint Optimization 44
3.5 Summary of The Proposed Approach 45

4 Facial Image Inpainting with a Learned Guidance Vector Field 46

4.1 PCA Based Image Inpainting 47
4.2 Guidance Vector Field Image Inpainting 48
4.3 Iterative Learning of Guidance Vector Field for Image Inpainting 49
4.4 Patch Selection for Learning PCA Model 50

5 Image Superresolution on Generic Face 53

5.1 Edge Model 53
5.2 Backprojected Error Correction 55
5.3 POCS Algorithm 57

6 Image Superresolution on Specific Face 60

6.1 Intelligent Image Selection 62
6.1.1 Pose Discrimination 62
6.1.2 Shape Analysis 64
6.1.3 Color Constancy 67
6.1.4 Texture Analysis 69
6.1.5 Similarity Measurement for Image Retrieval System . 71
6.2 Image Alignment 72
6.3 MRF Model for Face Hallucination 73
6.3.1 Color Constraint 74
6.3.2 Edge Constraint 75

7 Experiments and Results **76**

7.1 Experiments on Structure From Motion 76

7.1.1 Quantitative Evaluation on Synthetic Data 76

7.1.2 Qualitative Evaluation on Facial Expressions 78

7.2 Experiments on Image Inpainting 81

7.3 Experiments on Face Hallucination 87

7.3.1 Generic Faces 87

7.3.2 Specific Faces 96

8 Conclusion **109**

8.1 Future Work 112

Bibliography **112**

Summary

In this thesis, we are interested in the three issues on facial images. There are facial structure from motion, facial image inpainting and face hallucination. Currently, commercial 3D modeling systems can only generate realistic 3D structures by using highly calibrated and expensive camera systems. The cost and limitation of the system make it very hard for general applications. Hence, reconstructing 3D structure using low-cost cameras has become a popular research topic recently. Moreover, facial images are commonly seen in our photo album. 3D facial structure is one of the most important and interesting applications in practice. We are not only interested at the recovery of the facial structure, but also the facial image enhancement. Since the input is captured from low cost and low resolution camera, the resolution and quality of the input usually are poor. Handling problems arising from noise and low resolution is a challenging aspect of this thesis. Besides that, missing data and occlusion problems are also considered. Reconstructing 3D objects with good visual quality involves two aspects, namely, finding its geometric structure accurately, and its high resolution texture map.

Manually retouching the corrupted images is time-consuming and tedious. Therefore, the objective of image inpainting is to automatically re-

store the missing information in the corrupted images. Face hallucination which aims to estimate a high resolution facial image from low resolution facial image(s), is a well-known inverse problem of long standing interest in face analysis and image processing. Even though today camera sensors are able to acquire high resolution images, faces in the images may be still in low resolution if the images are captured from a long distance. We are interested in generating a visually-pleasing and aesthetically attractive facial image from a corrupted low resolution image with a set of high resolution training images. Moreover, the unique structure of faces makes the problems challenging.

A non-rigid structure from motion algorithm is introduced. Our work improves the existing non-rigid factorization method by using a batch algorithm and applying a set of non-linear shape constraints. Our experiments have shown that the performance of the proposed algorithm is better than the prior work, qualitatively and quantitatively.

An image inpainting technique on facial images is introduced. In prior work, image inpainting was mainly applied on generic images with small corrupted regions. Generally, the missing region is filled by propagating the boundary information inward. Hence, the results are usually blur and unstructured. Our work handles the unique structure of face and retouching large portion of missing region. We propose a facial image inpainting with a learned guidance vector field (GVF). Assuming that the facial images are aligned, the GVF can be learned from the training data set based on a principal component analysis model. The GVF is formulated by a Poisson equation and solved by the Gauss-Seidel solver. Thus, our inpainting results are seamless and the structure of face is preserved.

High resolution images with high quality textural details are always desirable as inputs for many computer vision and pattern recognition problems, e.g. in algorithms for structure from motion, object recognition, motion analysis, image understanding etc. A low resolution image can be interpreted as an image which has been obtained by blurring a high resolution image and then downsampling it, causing loss of high frequency details. Thus, the objective of image superresolution aims to recover the missing high frequency details of the low resolution image. In the thesis, two image superresolution methods on facial images are proposed. One is applied on generic face and the other one is on specific face.

In image superresolution on generic face, the superresolution image we seek must satisfy a generic face model constraint - the high frequency information added must be consistent with the model - and the data constraint which ensures that the superresolution image is consistent with the given image. An iterative projection onto convex sets (POCS) algorithm is proposed to optimize these two constraints.

In image superresolution on specific face, we assume that the facial images belong to the same person but vary in pose and expression. An image retrieval system for pose and expression is proposed. Given a low resolution image as query image, a set of high resolution images with similar pose and expression are retrieved based on the pose, shape and texture. The selected high resolution images are used as the candidates for the image superresolution. Next, a Markov random field (MRF) model based on proposed color and edge constraints are used to find the optimum solution of the super-resolution image. High texture details of superresolution images which are four to eight times larger than the original low resolution images

are generated by the proposed approach.

In summary, we contribute an extended work on facial structure from motion by using a batch algorithm. We also proposed a learning-based image inpainting approach to restore the corrupted facial image based on a PCA model. We also overcome the face hallucination problem on low resolution facial images by a set of high resolution images. Our approaches can be used to handle low resolution images captured under dim environment. We also resolve the blurred, noise problem in our face hallucination approaches. In addition, our image retrieval system overcomes the pose and expression issue in facial images and significantly improves the performance of our face hallucination approach.

List of Tables

2.1 The hierarchy of transformations and their corresponding invariant properties 20

7.1 Mean squared error results on our approach, PCA-based inpainting and Poisson inpainting [69] 86

7.2 Sharpness measurement results on bilinear interpolation, neighborhood embedding method [21], fast image upsampling method [77], sparse representation method [99] and the proposed superresolution method 93

7.3 Comparison results on the proposed similarity measurement, Equation 6.7 with the conventional approaches, χ^2 distance and $L2$ norm. 97

List of Figures

1.1 Frames captured from CCTV. Faces in the low quality image are not recognizable. 3

2.1 (a) Original image, (b) corrupted image, (c) inpainted image by solving Navier-Stokes equation [20], (d) inpainted image by exemplar-based inpainting algorithm [25]. 28

2.2 Block diagram of image formation model from the real scene to low resolution images(Left) and block diagram of super-resolution image reconstruction from the multiple low resolution images (Right) 30

2.3 Shechtman's space-time regularization. The figure shows four 3x3x3 LR input patches and a 3x3x3 HR output patch. It illustrates that the pixel value in the HR output frame can be obtained by taking the mean of the input patches in respect of t (first pixel), x (fifth and sixth pixels) and y (fourth and seventh pixels). The weight matrix in the regularization equation is used to avoid smoothing (averaging) over spatial and temporal edges. When the temporal regularization term is given heavier weight, it implies that the pixel values in respect of t are similar (no motion are taken in that region or there are no temporal edges in that region). Thus, smoothing in temporal domain will not lose the detail information. Similarly to the spatial regularization term. . . 35

4.1 Illustration of missing region, $\mathcal{M}$, and bounding band, $\mathcal{B}$, in ROI. 48

5.1 A 2D geometric illustration of the relationships between the vectors. 56

6.1 Samples of input images captured under different lighting conditions and camera models 61

6.2 The flowchart of the high resolution candidate image retrieval system for image superresolution 63

6.3 A high resolution training image with 38 marked feature point. 64

6.4 Some high resolution training images with different poses. . 64

6.5 (a) Illustration of 3D body rotations about three orthogonal axes. These rotations are referred to as yaw, pitch and roll. (b) and (c) illustrate the ratio R_1 are affected by pose change. When face turn to right, the distance from the center of left eye to the nose tip (red line) is longer than the center of right eye to the nose tip (blue line). 65
6.6 High resolution training images with same pose but different expressions. 66
6.7 Fitting ellipse on the regions of interest (eyes and mouth). . 67
6.8 Images capture from different camera and under different lighting conditions. (a) Image captured by HD camera (b) Image captured by mobile phone camera (c) Image captured by video camera (indoor environment) (d) Image captured by video camera (outdoor environment) 67
6.9 Illustration of the color correction by histogram equalization 68
6.10 (a) A color image, (b) Canny edge of image (a), (c) Gradient magnitude of image (a), (d) Gradient angle of image (a). (e)the corresponding bin number of the gradient angle along the edges. (f) the gradient magnitude along the edges. (g)-(i)are the grids for $l = 0$ to $l = 2$ pyramid levels in the conventional PHOG. (j) is the grid for level $l = 1$ in the proposed EPHOG. Only the four cells at the center of the images are computed. (k)-(n) are the corresponding HOG descriptors of (g)-(j). 71

7.1 Relative errors of the three different appoaches of the factorization algorithms on rigid synthetic data under different levels of Gaussian white noise. (a, b and c). 78
7.2 Relative errors of the three different appoaches of the factorization algorithms on non-rigid data under different levels of Gaussian white noise (a, b and c). 79
7.3 (a) Happy expression image with rotation about $x = -10°$, $y = 20°$ and $z = 10°$ (b) Neutral expression image with rotation about $x = 0°$, $y = 0°$ and $z = 0°$ (c) Sad expression image with rotation about $x = -10°$, $y = -20°$ and $z = -10°$ (d) Surprise expression image with rotation about $x = -10°$, $y = 20°$ and $z = 10°$. 80
7.4 (a) Ground truth of 3D happy expression (b) Ground truth of 3D neutral expression (c) Ground truth of 3D sad expression (d) Ground truth of 3D surprise expression. 80
7.5 (a) Reconstructed 3D happy expression (b) Reconstructed 3D neutral expression (c) Reconstructed 3D sad expression (d) Reconstructed 3D surprise expression under 0% Gaussian white noise. 81
7.6 (a) Reconstructed 3D happy expression (b) Reconstructed 3D neutral expression (c) Reconstructed 3D sad expression (d) Reconstructed 3D surprise expression under 5% Gaussian white noise. 81

7.7 (a) Reconstructed 3D happy expression (b) Reconstructed 3D neutral expression (c) Reconstructed 3D sad expression (d) Reconstructed 3D surprise expression under 10% Gaussian white noise. 82
7.8 Image inpainting when a subject is in the training database or not: (a) the corrupted images, (b) inpainted images when the subject is not in the training database, (c) inpainted images when the subject is in the training database and (d) original images. 83
7.9 Comparison with other image inpainting approaches: (a) the corrupted images, (b) Poisson image inpainting, (c) inpainted images by iterative learning GVF (Algorithm 1) and (d) the original images. 84
7.10 Comparison with other image inpainting approaches: (a) the corrupted images, (b) inpainted images by PCA-based approach, (c) inpainted images by iterative learning GVF (Algorithm 1) and (d) the original images. 85
7.11 Samples of training images: (a) the original image obtained from an image sequence, (b) the corresponding warped image. 87
7.12 Image inpainting by patch-based learning model: (a) the original images, (b) corrupted images, (c) inpainted images when the test image is not in the training set and (d) inpainted images when the test image is in the training set. . . 88

7.13 Five high resolution training images obtained from Yale Face Database B are shown in the first row and corresponding simulated low resolution images are shown in the second row. The third row images are the bilinearly interpolated images. 89

7.14 The first four principal components of $\mathcal{H}$ (1^{st} row), $\tilde{\mathcal{L}}$(2^{nd} row) and $\mathcal{E}$ (3^{rd} row) and the matrix of $< \mathbf{h}_i, \tilde{\mathbf{l}}_j >$, i, $j = 1, \ldots, 4$. 89

7.15 Superresolution on images with different levels of magnification: (a) The original high resolution images, (b) SR images for $\times 4$ magnification, (c) bilinearly interpolated images for $\times 4$ magnification, (d) $\times 4$ magnification by pixel replication (e) SR images for $\times 2$ magnification and (f) Bilinearly interpolated images for $\times 2$ magnification. 91

7.16 SR experiments when subject is in the training database or not, with different methods. (a) The original high resolution images, (b) SR images by our approach (test image is included in the training dataset), (c) SR images by our approach (test image is not included in the training dataset), (d) SR images by conventional PCA reconstruction method when the test image is in the training dataset and (e) not included in the training dataset. 92

7.17 Superresolution (1^{st} row), bilinear (2^{nd} row), neighborhood embedding method [21](3^{rd} row), fast image upsampling method [77] (4^{th} row), and sparse representation method [99] (5^{th} row) on images with different blurs and noise: (a) Blurry image from a 5×5 averaging mask, (b) Blurry image from a motion blur filter, (c) 20db noisy image (d) Image with combined effects of (a), (b) and (c). 94

7.18 Image inpainting and superresolution: (a) The original high resolution image, (b) The damaged low resolution image, (c) The interpolated inpainted image and (d) The super-resolved inpainted image. 95

7.19 High resolution training images with different expressions and poses extracted from each dataset 96

7.20 (1^{st} row)A low resolution query image (2^{nd} row)The region of interest (left to right: left eye, right eye, mouth) with bicubic interpolation for image selection and the corresponding high resolution images retrieved from the training dataset are shown in the next each column 98

7.21 Image superresolution on a specific person with different expressions. (a)Input images (b) Bicubic interpolated results on input images (c)Bicubic interpolated results with color correction (d) Superresolution images for $\times 4$ magnification . 100

7.22 (a)Input image (b)Bicubic interpolation (c)Superresolution image by randomly selected high resolution patches (d) Superresolution image by randomly selected the patches from the aligned images (e)Superresolution image by selected patches from the aligned images with similar pose and expression for ×8 magnification 102
7.23 (a) Bicubic interpolation (b)The VISTA approach [34] (c)Neighborhood embedding method [21] (d)Fast image upsampling method [77] (e) Sparse representation method [99] (f)Our method for ×4 magnification 104
7.24 (a) Bicubic interpolation (b)The VISTA approach [34] (c)Neighborhood embedding method [21] (d)Fast image upsampling method[77] (e) Sparse representation method [99] (f)Our method for ×8 magnification 105
7.25 (a) Bicubic interpolation (b)The VISTA approach [34] (c)Neighborhood embedding method [21] (d)Fast image upsampling method[77] (e) Sparse representation method [99] (f)Our method for ×8 magnification 106
7.26 (a) Input image with pepper noise and underexposure (b) Bicubic interpolation (c)Neighborhood embedding method [21] (d) Sparse representation method [99] (e) Our method for ×2 magnification 108

Chapter 1

Introduction

1.1 Background

In digital photography era, most people have a lot of personal photos which recorded their memorable events in their personal computer. Since we are not professional photographers, some of them are not desirable. These low quality images usually are taken by low resolution cameras such as webcams, inexpensive pocket cameras, mobile phones etc. In certain circumstance, the images are taken by cameras from a long distance. The subjects in these images usually are unclear due to their low resolution and poor quality of the lenses and camera sensors. Therefore, simple image interpolation approaches which enlarge the size of the images are not able to resolve the problem here. The textural details need to be enhanced but we only have a single image with unique pose and expression. However, there are a lot of other high quality images in our photo album or we can easily capture high resolution personal photos later. Although the expressions and poses in these images are not exactly same as the low quality image to be enhanced, these high resolution image still can be used as examples

to improve the low quality images. This problem is also known as face hallucination, and it is discussed in Section 1.1.1.

In addition, the photos may be corrupted during data transmission and lossy image compression. The face may be occluded by other subject. Under these circumstances, the low resolution photos consist missing information which make the problem even more challenging. To overcome the problem, the corrupted regions need to be inpainted first. The image inpainting is investigated in Section 1.1.3.

Moreover, face is a non-rigid 3D object. Expression and pose variants make the shape and appearance of the face in the 2D image very different. Therefore, facial image registration is required to align the input image and traning images especially when the expression and pose are different.

1.1.1 Image Superresolution

High resolution images with high quality textural details are always desirable as inputs for many computer vision and pattern recognition problems, e.g. in algorithms for structure from motion, object recognition, motion analysis, image understanding etc. However, due to physical limitations and cost considerations in many applications, this type of high quality input data is not available for downstream processing. Thus, image enhancement methods like superresolution play a useful role to improve the performance of high-level computer vision problems such as recognizing people in a wideangle scene by CCTV, as for example in Figure 1.1.

Image superresolution which aims to estimate a high resolution image from low resolution image(s), is a well-known inverse problem of long standing interest in image processing. It is similar to image restoration

Figure 1.1: Frames captured from CCTV. Faces in the low quality image are not recognizable.

which aims to recover a good quality image from degraded images that are blurred and noisy. However, the objective of image superresolution technique is not only to recover good quality enhanced images, but also to increase size of the images. Thus, image superresolution is very challenging image processing problem.

During the image acquisition process, various blurring effects, noises and a large sampling interval contribute to loss of image sharpness. Let $\mathbf{h}$ denote as a m dimension of a desired high resolution image vector, $\mathbf{l}$ denote as a n dimension of a low resolution image vector where $n << m$. The $\mathbf{l}$ can be represented as

$$\mathbf{l} = \mathbf{A}\mathbf{h} + \mathbf{n} \tag{1.1}$$

where $\mathbf{A}$ is a $n \times m$ image acquisition process matrix involved optical blurring, motion blurring, image warping, down-sampling process, etc and $\mathbf{n}$ is a noise vector. Image superresolution techniques seek to solve this inverse problem and recover the missing high frequency or detail information in $\mathbf{h}$. Merely interpolating the given image to a large size will only produce a blurry image because $\mathbf{A}$ involved various blurring effects. The problem of

simultaneously estimating the missing high frequencies details, denoising and deblurring is very difficult in image processing because many unknown and highly correlated variables are to be resolved. In addition, we usually only have a single image of the subject with the same pose and expression. The information of an image is very limited. Some prior knowledge and additional constraints are required to resolve the problem. Thus, single image superresolution is more challenging than multiple image superresolution.

1.1.2 Facial Image Superresolution

As mentioned in Section 1.1, photos are sometimes captured by a camera from the long distance or the wide fields of view. Even though the camera is a high resolution camera, the faces in the photos can be low resolution and blurred if the person is located far away from the camera. Thus, superresolved at the face region is essential in practice.

Facial image superresolution is also known as face hallucination. It was first appeared in [5] by Baker et. al. It is a special case of image superresolution. The objective of face hallucination is synthesizing a high resolution face image from the low resolution image with a large collection of high resolution face images. The output image requires to preserve the basic structure of the low resolution image and its low frequency components. In addition, high frequency components are learned from the collection of high resolution images in database.

However, face hallucination is still open to debate on its application of face recognition because the synthesized high resolution face image relies on the training images. Only its low frequency signals are from the input source. The high frequency information synthesized from high resolution

image may affect the recognition accuracy. Nevertheless, it still can be applied on image visualization, image enhancement, image restoration, image synthesis etc. which do not require the accuracy of the face image. In image synthesis, generating visually pleasing image is more important than its accuracy.

1.1.3 Image Inpainting

Image inpainting was initially used to restore deteriorated artworks. Manually inpainting the artworks by retouchers is a time-consuming work. Retouchers need to fully understand the history background of the artworks and then carefully fill the missing or damaged parts of the artworks.

Today, these damaged artworks or corrupted old photos captured by film cameras can be scanned into digital format. We can apply digital image inpainting techniques on their digital version first. It not only saves retouchers' time, but also avoid mistakes to further damaged the original artworks. Thus, digital image inpainting is a non-trivial and challenging problem in image restoration.

Besides inpainting damaged artworks and old photos, digital image inpainting can also apply on corrupted images due to data transmission errors, removing objects, watermarks, subtitles and logos in images etc. Moveover, image superresolution can be considered as a special case of image inpainting because image superresolution is recovering pixel values between two sampling pixels in the low resolution image. Both of image super-resolution and image inpainting are trying to restore back the lost information in the input image.

1.1.4 Facial Structure Recovery from Motion

The structure of targets such as the face and human body may change with time. It is very difficult to recognize these non-rigid targets easily from 2D images because the targets' pose and structure are changing simultaneously. Thus, recovery of non-rigid structure has become a new focus of research in structure from motion over the past decade. In this case, we need to extract the target's pose as well as its deformation information from the image sequence. Moreover, the representation of the non-rigid structure is also another challenging problem. The facial structure recovery can give us a better understanding on the pose and expression of the face.

1.2 Motivation

The current state-of-the-art of structure recovery can only generate realistic 3D structures by using highly calibrated and expensive camera systems. These systems require not only high resolution cameras and special lighting condition, but also sophisticated calibration procedures. Moreover, the computational cost of these systems is also very high. Sometimes the target has to remain stationary for a few seconds during the acquisition process. The target may even need to have special markers or special phosphorescent powder on its surface to obtain more accurate results. Thus, current systems are still impractical for general purpose use, especially in outdoor environments. Hence, generating realistic 3D appearance of targets from low cost and low resolution video cameras could be of great practical importance. There is a high demand for simple 3D modeling systems which use low cost cameras in the billion-dollar gaming, entertainment and social

networking market because such systems can give users a realistic visual experience while they are playing games, watching movies or communicating with other users. Since each video frame can only provide limited information of the target's texture and 3D information, the challenge is to fuse information from each frame to generate a high quality 3D structure of the target. Both texture enhancement and recovery of the 3D geometry are ill-posed problems, as there may exist multiple solutions that satisfy the constraints from 2D images. Thus, additional prior knowledge is required to overcome these problems.

Low resolution images, in practice, are always low quality images. Various blur effects such as motion blur, optical blur etc are involved in these low resolution images because they are captured by low-cost cameras which do not have good lens to handle the focusing issue and hand shaking problem. In addition, the low-end camera always gives poor color tone because it uses single image sensor with demosaicing algorithm [31] to interpolate the full true RGB color image. Sensor noise and other random noisy signals appeared during image acquisition also make the images poor to visualize. Hence, from corrupted, blurred, noisy and low resolution image to high resolution image, it is an undeniably challenging topic in image processing. In this thesis, we are interested in using learning based approach to achieve image inpainting and image superresolution on low resolution facial images. The facial images are very common and useful images but they consist of unique structures eg. eye, mouth and nose.

For facial image inpainting, we believe that the missing information in images with unique structure can be learned from the uncorrupted dataset. Inpainting facial image is a very challenging work because we are very

sensitive to the changes on the face. In addition, the learned information also needs to be integrated with the boundary of the corrupted region, so that the result is seamless.

For face hallucination, the high resolution image is the ultimate target we are interested in. The blur kernels and noise are not important to recover them exactly. Thus, learning based face hallucination approach which directly estimates the high resolution image from the given low resolution input is a more practical solution on handling the single image superresolution problem. In addition, using all the images in the database for training not only increase the computational time, but also give poor results. If the training images which are similar to the input image are retrieved from a large collection of data first, then the learning process will be improved significantly.

1.3 Applications

Generating realistic high visual quality 3D appearance of objects from image sequences has many possible applications in multimedia, gaming, entertainment, object recognition and even archaeology. In the following, some applications are briefly described.

1.3.1 Gaming

Massive Multiplayer Online Role Playing Games (MMORPG) are now very popular. In MMORPG, players will create avatars to represent them, and personalize them. Our work can let players create an avatar which looks very much like them. The players may just need to stand in front of

a webcam to let the system create avatars which have the attributes of the players. This will definitely make the game more fun and interesting. Similar ideas can also apply to other offline games such as *The Sims*, a strategic life simulation computer game. Players can simulate the desired lifestyle with an avatar which looks like them.

1.3.2 Multimedia, Entertainment, Computer Graphics

The 3D reconstruction and animation of 3D structure can allow people in the audience to play roles in movies. Here people just need to stand at a particular location before entering the cinema theater. The system will reconstruct the 3D structure of these people and make them available for roles in the movie. Mitsui and Toshiba [1] showed a similar idea in Expo 2005, Aichi, Japan.

In movie production, there are many special effects which use 3D reconstruction and computer vision. Reconstructing the 3D structure from some portraits such as Leonardo da Vinci's Mona Lisa, showing a 3D newscaster on newspaper, animating some desired behavior using animal models are some applications.

1.3.3 Social Networking

Nowadays, social networking is an important communication channel with friends in our daily life. It brings people all over the world closer together. Teleconference, for example, can save our time and travel costs by meetings or presentations over the internet. However, the video teleconference is still very limited due to data transmission and quality of inexpensive webcam.

Low quality videos make the facial pose and expression of the users unclear. These body languages are important information to help users to give better presentation during teleconference. Generating high visual video from low cost webcam is always desired especially the facial videos.

1.3.4 Face Recognition

3D face reconstruction and 3D face morphing can improve the performance of face recognition systems. Numerous papers have shown that 3D face recognition algorithms perform much better than 2D. It is mainly because the same face can look very different from different view points. Hence, it is difficult to properly represent a face by its 2D views. In [60], Matthews et al. showed that a 2D face model can generate model instances that do not exist in the real 3D world. Using the 2D model for face recognition also requires more parameters than the 3D model, which can lead to longer computational time to estimate them. Moreover, parameter estimation for face fitting is a non-linear optimization problem and it is difficult to obtain the optimal solution for a large number of parameters. Thus, using more parameters implies that the robustness and accuracy of the system would be poorer.

3D face morphing is also greatly useful in face recognition, as it may be required to deform the input image to a particular expression to fit the face data in the database. This will increase the accuracy of face recognition compared to direct fitting without deformation.

In short, the 3D face reconstruction and the 3D face morphing can handle intra-class variations due to pose and face expression changes, respectively.

1.4 Thesis Contributions

The objective of this thesis is the development of robust methods for facial structure from motion, face hallucination and facial image inpainting from corrupted low resolution images. The key contributions of the thesis are summarized in the following sections.

1.4.1 Facial Structure from Motion

Based on Jing Xiao et al.'s [97] closed-form solution for non-rigid SFM with rotation and basis constraints , we proposed an extended work which uses non-linear shape constraints to improve the existing closed-form solution under noisy conditions. The proposed batch algorithm partitions the measurement matrix and recovers a 3D structure from each partition separately. Then a non-linear optimization algorithm that seeks to preserve the shape by maintaining the length and angle between pairs of feature points is applied to estimate a refined 3D structure using all the solutions from the partitions. Qualitative and quantitative evaluations showed that the new algorithm gives robust and accurate results compared to the original factorization method for both rigid and non-rigid structures.

1.4.2 Facial Image Inpainting

In this part of thesis, we seek to inpaint face images containing missing regions using training images and the undamaged region of the given image. For generating a realistic and visually pleasing face from a corrupted facial image, we need to satisfy two criteria. The first is to retain the structure of the face faithfully. Faces have unique structure with regions such as eyes,

nose and mouth, and if these unique structures are missing, there is no way to find a similar patch directly from the undamaged region of the image. Thus, we need to rely on the help of training images to learn the structure, and use the learned model to provide the structure of the missing region.

The second criterion is the smoothness between the recovered region and its surroundings. To have a realistic and visually pleasing effect, smoothness or continuity of the textural details is as important as the structure of the image. Thus, we seek for the solution to satisfy a set of Poisson equations with Dirichlet boundary conditions. These two criteria can jointly retain the structure of the image as well as maintain continuity between the inpainted region and its surroundings.

Our approach is based on the iterative projection onto convex sets (POCS) algorithm. POCS can iteratively reconstruct a signal by incorporating multiple convex constraints. It was initially proposed by Papoulis [67] and Gerchberg [36] for signal extrapolation and is also known as the Gerchberg-Papoulis algorithm. It has since been applied in various image restoration and image enhancement problems. Stark and Oskoui [82] applied POCS with data constraints and prior knowledge for superresolution of images. Tekalp et al. [66] also applied POCS to restore out-of-focus blurry images. Bandwidth constraints, spatial support contraints, consistency, positivity, etc. are typical examples of the convex constraints used in POCS approaches.

1.4.3 Image Superresolution on Generic Face

The following two contributions of the thesis are the study of facial image superresolution. Here, we use image superresolution on generic face of

single facial modality. In the next section, the use of image superresolution on the same person with multiple facial modality (pose and expression) is briefly discussed.

A set of high resolution face images with same pose and expression is used for estimating the missing details in the given low resolution input image with the same modality. The estimated high frequency details are added back to the input image to increase its sharpness. However, if the given low resolution image is not a member of the training set, as will be the case in practice, it must be ensured that the superresolution image produced after the details are added is consistent with the given image. In other words, we seek to estimate a super-resolved image which is constrained to have missing details consistent with the training data set, and is also consistent with the given input image.

Since the pose and expression of the persons are same here, the face images can be simply aligned by affine image alignment. Inspired by the work in image inpainting, we proposed another POCS method for extrapolating the missing high frequency components in the input low resolution image. Since the input image may not be in the training set, an image consistency constraint is imposed in the POCS method to preserve the low frequency information in the input image.

1.4.4 Image Retrieval of Facial Expression and Pose Estimation

Since face is a non-rigid structure, changes in facial expression and pose will significantly affect the final results if they are not handled appropriately. Compared to the texture and the ellipse shape of open eyes and

the texture and the shape of eyelids in closed eye images, the texture and structure features between these two sets of images suppose to be different. However, the state-of-the-art features like Gabor features, Scale-invariant feature transform (SIFT) [56] features etc., are not easy to be extracted from low quality images. Here, we need to retrieve high resolution images with same facial expression and pose as the low resolution input image. Thus, we proposed a fast and robust image retrieval method with combined the structural and textural information of the low resolution input image. The shape ratios are used to represent the shape feature of the facial expression. The Pyramid of Histograms of Orientation Gradients (PHoG) features [12] are used to represent the texture of the expression on low quality images. PHoG is a SIFT-inspired feature descriptor and it has been widely used in human detection in surveillance system [37]. Experimentally, PHoG features extracted from low resolution images are well-performed on similar expression retrieval in high resolution image set.

1.4.5 Image Superresolution on Specific Face

Since most people are not professional photographers, some photos are blur, with poor lighting conditions or low resolution. To enhance these photo especially the face region which we are most interested in, image superresolution on specific face plays a role here. It is noted that facial pose and facial expression of these poor quality images vary significantly but the subject is known. Since the work in the previous section is not able to handle the large changes in pose and expression, we proposed another local based data-driven image superresolution approach on this issue.

To enhance the quality of the face region on same person, we can learn

it from high resolution training images with the same person's face. On this problem, we assumed that the subject is known and high resolution images of the same subject with different expressions and poses are provided. A Markov random field (MRF) model based on a proposed color and edge constraints are used to find the optimum solution of the superresolution image. High textural detail superresolution images which are four to eight times larger than the original low resolution images are generated by the proposed approach.

1.5 Thesis Organization

The rest of the thesis is organized as follows: Chapter 2 briefly provides an overview of the related work in structure recovery, image inpainting and image superresolution. Some inspired works are discussed. In Chapter 3, a batch algorithm for facial structure from motion with additional shape constraints is introduced. In Chapter 4, facial image inpainting technique is studied. A new approach which combines PDE-based and learning based approach on image inpainting is proposed. The details of the approach are discussed in this chapter. Our works on facial image superresolution are investigated in Chapter 5 and Chapter 6. An approach of the facial image superresolution on generic face is discussed in Chapter 5. The facial image superresolution on specific face is investigated in Chapter 6. All experiments and results are showed in Chapter 7. The thesis conclusion and future work are discussed in Chapter 8.

Chapter 2

Literature Review

In this chapter, some related work which have made significant contributions to our research are briefly discussed. In the first session, the prior work on structure recovery is studied. In the second section, the study of facial action image retrieval is investigated. We focused work on facial expression analysis and recognition because there is limited work on facial image retrieval. In the third session, several different approaches of image inpainting are discussed. In the last section, the prior work on image super-resolution is discussed.

2.1 Structure Recovery

Obtaining 3D models from an image sequence is a very challenging task in computer vision because both, the structure of the target and camera motion are not known. It is an ill-posed problem, which requires additional constraints and assumptions to overcome the difficulty. Rigid structures, piecewise continuous surfaces and weak perspective camera model are some assumptions usually used for structure recovery. Two current state-of-the-

art structure recovery methods, including Shape From Silhouette (SFS) [23],[48] and Structure From Motion (SFM) are discussed in the following subsection. The SFM algorithm include self-calibration [62] and factorization algorithm[90],[85].

2.1.1 Shape From Silhouette

Shape From Silhouette (SFS) is a 3D shape reconstruction method from silhouettes or contour images of targets. It was first proposed by Baumgart [8] in 1974 to estimate the 3D shapes of a baby doll and a toy horse from four silhouette images acquired from different viewpoints. The SFS is mainly used to reconstruct static objects under known camera geometry. Basically, the SFS finds the tightest possible bound of the Visual Hull, which is the intersection of the visual cones formed by the silhouettes and camera centers. The more distinct the silhouette images are, the better the reconstruction of the 3D target shape by the SFS method.

Recently, different extensions of SFS have been proposed. In [23], Cheung et al. extend the traditional SFS methods to estimate the 3D shapes of targets with unknown motion, such as articulated object, by combining all of the silhouette images of the targets over time. Space Carving, proposed by Kutulakos et al. [48], is one of the important works on SFS. Space carving reconstructs the structure by iteratively removing portions of an initial set of voxels based on N given images until it converges to the Visual Hull. Although space carving allows the input cameras to be at arbitrary positions, the cameras' geometry has to be known for reconstruction.

Although silhouette images can be easily obtained by low-level computer vision methods, the performance of the SFS method would be limited by

the quality of information from the silhouette images. The reconstruction results depend on the number of images used and their resolution.

2.1.2 Structure from Motion

Given an image sequence captured by an uncalibrated camera undergoing unknown motion, recovering the structure of the targets in the image sequence and the camera motion is known as Structure From Motion (SFM). It was first introduced by Longuet-Higgins [54] to reconstruct a scene from two views. SFM assumes that the target correspondences are known but the camera motion is not known. On the other hand, the SFS assumes that the correspondences are unknown but the camera motion is known. In practice, finding correspondences from the image sequence is much easier than knowing the camera motion over the frames. In the following, two important SFM methods, self-calibration and factorization algorithm, are discussed. The main difference between them is that self-calibration finds the camera parameters first and then reconstructs the structure while the factorization algorithm solves for the camera motion and the structure geometry simultaneously.

Self-Calibration

Self-calibration is the process of finding camera parameters from multiple uncalibrated images. The camera parameters consist of the intrinsic parameters and the extrinsic parameters. The intrinsic parameters include the focal length, the location of the image center, the effective pixel size in the horizontal and vertical direction and radial distortion coefficient. The extrinsic parameters are the translation vector and the rotation matrix

which specify the transformation between the camera and the world reference coordinate. Mathematically, they provide a transformation to map the 2D image coordinates to 3D world coordinates. Thus, once the camera parameters are found, the structure of the target can be recovered from the images.

Self-calibration was first introduced by Faugeras et al. in [62]. They used absolute conic and Kruppa equations to calibrate the camera. In projective geometry, the absolute conic is a conic on the plane at infinity, and also an invariant under the similarity transformation. Thus, it is an important tool to recover the intrinsic parameters in many self-calibration methods. Triggs [92] also introduced the absolute dual quadric, the dual of the absolute conic to autocalibrate the camera. In [92], a nonlinear optimization method was applied to recover the absolute quadric and conic simultaneously.

The stratified approach is another well-known self-calibration method which was proposed by Pollefeys [73]. The stratified approach made use of a new constraint, called the modulus constraint, for self-calibration. The approach started from projective calibration, followed by affine calibration, next metric calibration and finally Euclidean calibration. The concept of stratified approach uses the different transformations and their corresponding invariant properties to recover structures. The hierarchy of transformations is shown in Table 2.1.

In most works, the focus is on calibrating constant, but unknown camera parameters. Recently, variable camera parameters are receiving attention from some research groups because the auto-focus and zoom functions in commercial cameras violate the assumption of constant parameters [42],

Transformation	DOF	Matrix	Invariant Properties
Projective	15	$\begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix}$	cross ratio
Affine	12	$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$	relative distance along direction, parallelism, plane at infinity
Metric	7	$\begin{pmatrix} sr_{11} & sr_{12} & sr_{13} & t_x \\ sr_{21} & sr_{22} & sr_{23} & t_y \\ sr_{31} & sr_{32} & sr_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}$	relative distances, angles, absolute conic
Euclidean	6	$\begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}$	absolute distance, volume

Table 2.1: The hierarchy of transformations and their corresponding invariant properties

[55].

Factorization Algorithm

The factorization algorithm was first introduced by Tomasi and Kanade [90] to reconstruct 3D rigid structure from a single-view image sequence captured under arbitrary motion. It has been widely applied to the SFM problem over the past two decades. Basically, the factorization algorithm for the SFM decomposes the image feature tracks (*measurement matrix*) into motion of the camera and the 3D shape matrix via Singular Value Decomposition (SVD) and the rank theorem. However, the problem is ill-posed, as linear transformations of the solutions also yield valid motions and shape bases. Therefore, it is not possible to recover structure from the image sequence without introducing some prior knowledge. Additional

constraints such as orthogonality of the rotation matrix are required to recover the structure.

Generally, the orthographic camera model is chosen in the factorization algorithm because it is a good approximation to the perspective camera model when the reconstructed target is far from the camera and the depth variation within the target is relatively small. [85] and [72] also proposed extended factorization algorithms for perspective and paraperspective models, respectively.

Recently, recovery of different kinds of structures such as multiple linearly moving objects [39], articulated objects [98], model-based non-rigid objects [16], [14], [91], [97] have been reported. Model-based non-rigid object recovery is attractive because many interesting non-rigid objects in nature such as human faces can be represented by models. Reconstructing 3D human faces is very useful in face recognition. Compared to 2D face images, 3D faces are invariant to pose changes. The pose changes significantly affect the performance of face recognition algorithms. Therefore, we can use non-rigid factorization to decompose the pose and deformation of the non-rigid structure from an image sequence.

To model the deformation of these non-rigid objects, the weighted combination of basis shapes has been used in non-rigid SFM [16]. With this model, Jing Xiao et al. [97] showed a closed-form solution for the non-rigid SFM problem with rotation constraints and basis constraints. The solution is exact only when the data is noise free. The method does not work satisfactorily with noisy data [15].

Missing data and occlusion are two common problems in structure recovery. If the data is corrupted or feature points are occluded in some

frames, finding SFM by factorization algorithms becomes problematic because the correspondences in particular frames are unknown, making the measurement matrix incomplete [17], [80], [28]. In [90], Tomasi et al. use an interpolation technique to fill in the missing data based on the available matrix elements. Filling in missing data using the available data is known as *Imputation* in statistics. Later, Jacobs [28] proposed a closed-form solution to overcome the missing data problem. The idea of Jacobs' algorithm is to build up an orthogonal subspace from subsets of the matrix's columns. The algorithm fills in the missing elements by finding the closest affine space represented by the corresponding column in the matrix. However, these two methods work poorly under noise or seriously incomplete data. Shum et al. [80], Henrik et al. [2] and Buchanan et al [17] proposed alternation algorithms to solve the factorization problem with missing data. Basically, in these iterative algorithms, one of the decomposition matrices is taken as known, and the other is solved for directly.

2.2 Facial Action Image Retrieval

Facial Action Coding System (FACS) was first proposed by Ekman and Friesen [30]. It is mainly used for facial behaviour analysis. In [30], the facial behaviour is decomposed into 46 action units(AUs) which anatomically related to facial muscles. It has been widely inspired work on facial expression detection, tracking and recognition. In this thesis, we are interested at facial expression retrieval. To the best of our knowledge, very little work has been published on facial expression image retrieval. The objective of image retrieval is to find the similar facial action with the input image from a large collection of images. However, most of the FACS are

developed to recognize and analyze the given facial expressions like happy, sad, surprised, fear, angry etc. In general, FACS involves facial feature detection, feature tracking and pattern recognition. The related work is discussed in the following subsection.

2.2.1 Feature-based Tracking Approach

Beside feature detector, feature tracking is another mean to detect and match salient features over image frames. These salient features should be independent to image scale, rotation, illumination changes and viewpoint changes. Besides that, their neighborhood should provide sufficient information for matching algorithms.

For matching, since the position, orientation and calibration of the cameras may not be known, the correspondences are usually obtained based on spatial and frequency domain information in the 2D images. Generally, the feature matching algorithms compute some measurement error functions or similarity functions. In [101], Zhang et. al. matched the detected Harris corners using correlation windows. In Schmid et al. [76], a rotationally invariant descriptor of the local image region was used for feature matching.

To improve the matching algorithm, the motion of the feature points over the video sequence can be assumed to be piecewise smooth. Thus, instead of searching the entire image, the correspondence can be found based on its location in the previous image frame. In the following a few tracking algorithms are discussed:

- Kanade-Lucas-Tomasi Feature Tracker (KLT) [57], [89] was proposed by Lucas et al. It assumes that all points in the feature neighborhood

move with the same velocity. Thus, KLT can track distinct features with small motion very well. However, if the distinct features are too close, KLT will give incorrect feature matching.

- Kalman Filter is an efficient recursive filter that estimates the state of a dynamic system from incomplete and noisy measurements. The incomplete measurements can be due to occlusion or data corruption. In computer vision, it is used to predict the next position of the detected feature based on its position in the current frame. In [96], Greg et al. showed a number of tracking examples by Kalman filtering. Bar Shalom et al. [6] attempted a mixture representation by Extended Kalman Filter (EKF) for tracking nonlinear dynamic systems.

- Conditional Density Propagation (ConDensation) Algorithm was first proposed by M. Isard et al. [11]. The algorithm uses particle filtering to model and predict the next state of the feature. Their work showed that ConDensation algorithm performs better than the Kalman Filter and the EKF for non-linear dynamic systems because the ConDensation algorithm makes no assumptions about linearity or Gaussian probability density function (pdf); it iteratively predicts and measures the pdf of the feature over the video sequence.

2.2.2 Model-based Tracking Approach

Model-based approach is another method to improve feature detection. This approach assumes that the set of corresponding points changes under certain constraints. Active Appearance Models (AAM) [24], [95] is one of

the classical models which has been widely applied for face and medical imaging registration [13]. AAMs construct a target model from training images or 3D range scans. The model usually consists of a linear shape model and a linear appearance model. Then, the model is fit to the input images by varying the parameters. Model representation, fitting, computational cost and occlusion are four challenging problems in model-based approaches. In [95], an AAM head model is used to track a driver's head and the eye corners, eye region and head pose are robustly extracted for gaze estimation. In [59], the AAM is used for lip-reading. Koterba et al. [47] applied the improved AAM in [61] to fit the faces as well as to calibrate cameras.

2.2.3 Facial Feature Detection and Representation

Feature detection is an important preprocess in many computer vision applications because salient features are very useful not only in feature matching and tracking, but also in image registration, image understanding and object recognition. Usually, corners, vertices, edges [40], [27] and some salient patterns [56] are commonly chosen as the features because of their uniqueness in the gradient domains.

Harris corners [40] are the most widely used feature detector. Harris corner detector selects feature points that have large gradients in all directions at a predetermined scale. However, it would be very sensitive to changes in image scale. Lideberg [50] and Mikolajczyk et. al. [63] refined the Harris corner detector to a robust scale-invariant detector with determinant of the Hessian matrix and the Laplacian. In the later work, Lowe [56] proposed the Scale Invariant Feature Transform (SIFT) to fast

and robustly detect the interest points in the image by a Difference of Gaussian (DoG) filter. In [9], Bay proposed a scale and rotation invariant interest point detector, named Speeded Up Robust Features (SURF). The robustness and computation speed are improved compared to the previous work.

2.3 Image Inpainting

In image processing, image inpainting can be adopted to not only recover the damaged region, but can also be extended to texture synthesis [29], disocclusion [86], and super-resolution [70]. In [10], Bertalmio et al. also applied their inpainting approach for super-resolution. Liu et. al. [53] used an inpainting technique on an image to remove a fence which was occluding a target object.

Digital image inpainting techniques can be generally categorized into three approaches, PDE-based, exemplar-based and learning-based. PDE-based image inpainting propagates the image Laplacian in isophote directions from the exterior to fill in the missing region, and was introduced by Bertalmio et al .[70]. Later, Bertalmio et al. [10] extended this by applying the Navier-Stokes equation from fluid dynamics. Chan et al. [20] also applied a Euler-Lagrange method for image inpainting. Pérez et al. [69] solved the Poisson partial differential equation with Dirichlet boundary conditions and a given guidance vector field (GVF). All these smoothness constraint based image inpainting approaches work well on small regions for removing subtitles, watermarking, etc. because these approaches only rely on the smoothness constraint functions and boundary information. No other information about the missing region is used. Thus, undesirable blur

usually occurs when inpainting larger missing regions, as for example in Fig. 2.1 (c).

Exemplar-based image inpainting [41], [7], [29], [3], techniques are another group of approaches which synthesize the missing region from patches obtained from the uncorrupted regions of the image or a database. Basically, they search for optimum patches from the available ones to fill in the missing region such that the boundaries between neighboring patches in the inpainted image are smooth. These techniques are usually used for removing a bigger foreground object appearing over a background such as grass, sky, water etc. and inpainting the object-removed region to merge smoothly with existing background; this did not consider any structural information. The technique was extended by Criminisi et al. [25] and Sun et al. [83] to consider the structural information around the missing region by imposing constraints. This information helps to inpaint the missing region on more general background scenes. However, a unique structure such as a human face is difficult to learn directly from the uncorrupted region and manually imposing hard constraints is not trivial and perhaps impractical. An example of applying exemplar-based inpainting of [25], is shown in Fig. 2.1 (d), from which it can be seen that it is not always easy to seek similar patches from the available region to fill in the missing portion. Therefore, some kind of training process is required for inpainting specific object classes.

Learning-based image inpainting can be considered as a more advanced type of exemplar-based image inpainting. It does not directly find the best patch from the dataset to fill in the missing region. Rather, it learns the structure from the uncorrupted region and a training data set to synthesize

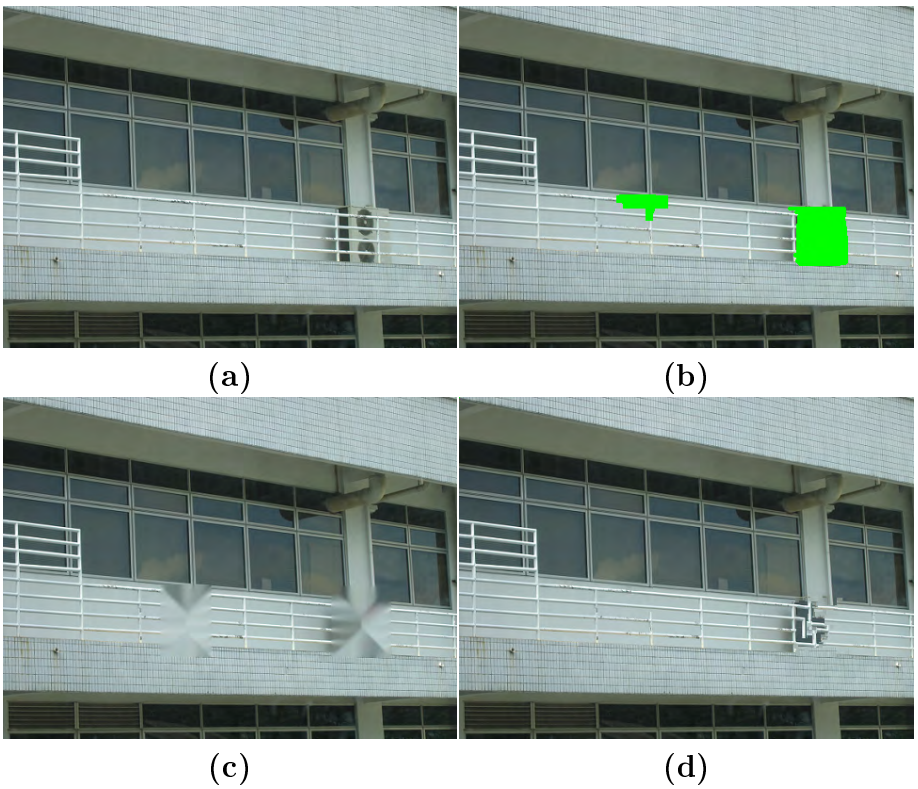


Figure 2.1: (a) Original image, (b) corrupted image, (c) inpainted image by solving Navier-Stokes equation [20], (d) inpainted image by exemplar-based inpainting algorithm [25].

a new patch for the missing region. The new patch may not be exactly the same as in the training data set. In [49], Levin et. al. estimated the missing region by using the image statistics of the training images. Roth and Black [74] used Markov random field (MRF) to learn the missing region from small image patches. Turaga and Chen [93] used a mixture of eigenspaces to recover corrupted video frames.

Since information is lost, it is not possible in general to recover it exactly. Thus, image inpainting is generally more concerned with recovering a realistic, visually pleasing reconstruction instead of exactly filling in the missing region. This is the main difference from other image restoration problems such as noise reduction which seek to recover the original infor-

mation as faithfully as possible from the noisy data.

2.4 Image Superresolution

Generally, image superresolution (SR) techniques can be categorized into two approaches, reconstruction-based SR and learning-based SR. Reconstruction-based SR approaches seek to recover a high-resolution image from multiple low resolution images, and reverse the effects of downsampling and blurring due to motion and optical lenses. The techniques in [45], [58] which align the low-resolution input images and then apply non-uniform interpolation onto a high resolution grid are examples of the reconstruction-based SR approaches. The difficulty of these approaches is registration using the low resolution images. Regularization methods using prior knowledge such as edge information [65] are also commonly used in reconstruction-based SR.

Tsai and Huang [75] first introduced SR image reconstruction. The idea of SR image reconstruction is to estimate a high resolution image from a set of low resolution images. In other words, SR seeks to recover and compensate for motion blur, camera blur effect, under-sampling and other distortions, by using multiple low resolution images. The block diagram of the image formation model and SR image reconstruction is shown in Figure 2.2. [68], [31] Obviously, the SR image reconstruction is also an ill-posed problem. Some prior knowledge or assumptions are required for overcoming the problem. Some approaches are briefly discussed here.

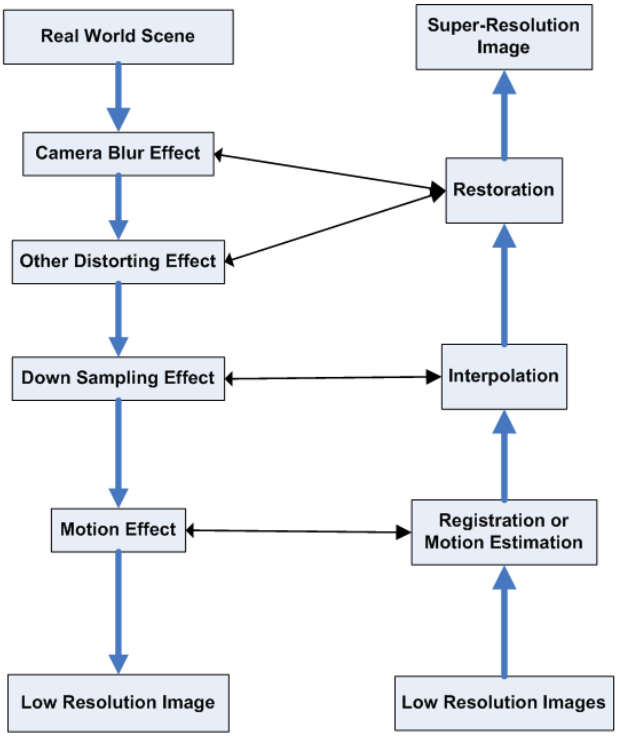


Figure 2.2: Block diagram of image formation model from the real scene to low resolution images(Left) and block diagram of super-resolution image reconstruction from the multiple low resolution images (Right)

2.4.1 Nonlinear Interpolation Approach

For interpolation approach, motion effect and distortion effect that were present during image formation are recovered or compensated independently. See Figure 2.2. Hence, the interpolation is the most intuitive approach for SR image reconstruction. Generally, the interpolation approach fits given sampled data with a continuous function under some smoothness assumptions to recover from the downsampling effect. Before applying the interpolation approach, the other effects are compensated. First, the relative motion among the set of low resolution images is estimated. Then, the low resolution images are registered based on their relative motion. Finally, an interpolation algorithm is applied to obtain a high resolution image. If the images are noisy, post-processing algorithms such as Wiener filtering are required to be applied on the interpolated HR image to remove the noise. The objective of the interpolation algorithm is to fill up missing samples in the high resolution image based on their neighborhood pixels. Linear, bilinear and cubic spline interpolators are the standard linear algorithms available in commercial image processing software. These algorithms are simple and fast but they cannot accurately estimate the high-frequency details and produce artifacts in the high resolution image. Thus, some researchers [94], [88] introduced non-linear interpolation algorithms to interpolate the missing data in the high resolution image. In [94], Stefan et al. first classify neighborhood pixels into three categories, constant (smooth regions), oriented (edges or directional patterns) and irregular (category between constant and oriented). Then, they performed adaptive interpolation based on the quadratic Volterra filter [87] and the classified neighborhood pixels. The advantages of this adaptive interpola-

tion algorithm are that the important details such as edges are preserved in the high resolution image and it can reduce artifacts due to aliasing or ringing.

These nonlinear interpolation approaches are not able to handle the distortions in the low resolution images. Its performance is limited on the simple distortion characteristics over all the low resolution images because the motion estimation, interpolation and noise reduction are applied independently [68]. Thus, the errors are potentially accumulated from motion estimation to noise reduction.

2.4.2 Frequency Domain Approach

This approach is based on the shifting property of the Fourier transform and the aliasing relationship between the continuous Fourier transform (CFT) and discrete Fourier transforms (DFT) of the undersampled frames to reconstruct the HR image from the low resolution images. In [75], it is assumed that the low resolution images are noise-free. Then, reconstruction of the HR image from the low resolution images is to determine the relationship between CFT and DFT which is an inverse (matrix) problem. Later, Kim et al. [81] extended this approach to blurred and noisy images by using a weighted recursive least squares formulation. The advantages of the frequency domain approach is that its well developed and it is convenient for parallel hardware implementation. However, it can only handle images with global translational motion and a simple noise model [68].

2.4.3 Regularization Approach

The SR image reconstruction can be written in matrix form as:

$$\mathbf{Ah} = \mathbf{l} \tag{2.1}$$

where $\mathbf{h}$ is the vectorized HR image values, $\mathbf{l}$ is the vectorized low resolution image values and $\mathbf{A}$ is a matrix that represents the relationship between $\mathbf{h}$ and $\mathbf{l}$. This is an underdetermined set of linear equations which can be solved by regularization. Typically, there are deterministic as well as stochastic approaches to solve this ill-posed problem. Constrained least squares (CLR) and maximum *a posteriori* (MAP) approaches are the popular regularization methods used in the deterministic approach and the stochastic approach, respectively. The MAP approach can use a more flexible prior model than the CLR approach. The CLR approach assumes that the prior probability function is Gaussian, and can be considered as a special case of the MAP approach. In [79], Shechtman et al. used weighted smoothing regularization terms that avoided smoothing over spatial and temporal edges to reconstruct HR video sequence from multiple low resolution video sequences. A corresponding weight matrix for the smoothing regularization terms is determined by the location, orientation and magnitude of space-time edges in the low resolution video sequences. For example, the temporal regularization term will give larger weight if the regions have high details but small motion (or no motion). The corresponding HR pixel value is obtained by taking the mean of the input low resolution video sequences with respect to t. Similarly, the spatial regularization term will give larger weight if the region is textureless or smooth but large motion.

In this case, the corresponding HR pixel is obtained by taking the mean of the input low resolution video sequences with respect to x and y. Figure 2.3 illustrates the space-time regularization for SR image reconstruction. In [31], Sina et al. combined the edge information in the luminance and chrominance color channels and proposed an intercolor dependency penalty term for SR image reconstruction.

Regularization approaches simultaneously solve the motion estimation, interpolation and restoration problem for SR image reconstruction, and this avoids error accumulation over the process. Besides that, they also provide a more robust and flexible method to model the noise characteristics and a priori knowledge for HR images.

2.4.4 Learning Approach

Reconstruction-based SR can be interpreted as an inverse mathematical model of the image system as shown in Figure 2.2. In practice, a real-world low resolution image consists of other image formation distortions such as motion blur, sensor noise etc. These problems have been addressed independently. Fergus et. al. [32] recovered blur kernels of the image with a gradient prior. Liu et. al. [52] estimated the noise based on a piecewise smooth image prior and the camera response functions. However, the problems become extremely difficult when all distortions appear in a single image. It is also impractical to recover all the unknown distortion effects. Instead of recovering all the unknown modules of image system such as motion, alignment, noise etc., learning-based SR synthesize a high-resolution image from low resolution image(s) with the help of an exemplar training set to learn a suitable model.

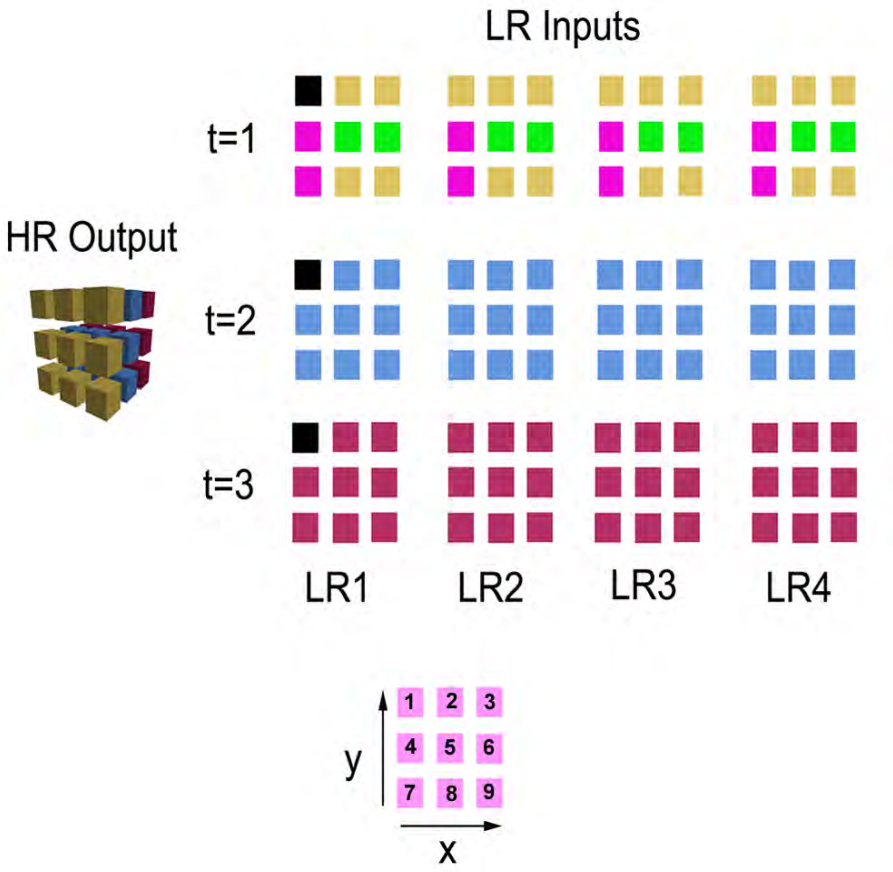


Figure 2.3: Shechtman's space-time regularization. The figure shows four 3x3x3 LR input patches and a 3x3x3 HR output patch. It illustrates that the pixel value in the HR output frame can be obtained by taking the mean of the input patches in respect of t (first pixel), x (fifth and sixth pixels) and y (fourth and seventh pixels). The weight matrix in the regularization equation is used to avoid smoothing (averaging) over spatial and temporal edges. When the temporal regularization term is given heavier weight, it implies that the pixel values in respect of t are similar (no motion are taken in that region or there are no temporal edges in that region). Thus, smoothing in temporal domain will not lose the detail information. Similarly to the spatial regularization term.

In [34], [4], [51] and [43], a Markov random field (MRF) model is used to learn the relationship between neighboring patches in low resolution and high resolution images and the relationship between the training data and input data. Freeman et. al. [34] proposed to perform a learning based image superresolution with a reference database with high resolution (HR) and low resolution (LR) image pairs. They used a MRF to model the relationship between the HR patches and LR patches and the relationship between adjacent HR patches. However, the results quality is often limited by the quality of the training patches. The results usually consists of some irregularities. Fattal [46] proposed an image superresolution approach to impose the edge constraint and conserve local intensities with respect to the LR input image. Sun et. al. [84] proposed a patch similarity measurement to select data with similar textural details to the input image for their image superresolution approach.

For learning-based approach, the input can even be just a single low resolution image. In [38], [33], Glasner et. al. and Freeman et. al. observed that recurrence of patches in a single image can be used for multiple image superresolution. The learning-based superresolution can be applied on those patches recurred across different scales of the same image. The method do not require any training dataset for learning the relationship between the HR patches and LR patches. The method can use the input image itself as exmaple patches for image superresolution. Even though it is a practical advantage, the input images have to be high resolution and consist of sufficient repeated patches inside the images.

Single facial image SR is also known as face hallucination. It was first proposed by Baker et. al. [5]. It is a learning based facial image SR. Since

input images are always facial images, facial feature or facial model can be used as the prior knowledge in the SR. Therefore, the facial structure can be preserved based on the prior knowledge learned from the training dataset. In [19], a MAP estimator is used to estimate the high resolution image that lies near to a PCA face sub-space. Mohammed et. al. [64] also used a PCA model and a non-parametric model for texture synthesis to generate a high resolution novel facial images. Jia and Kong [43] build a tensor model with different person, illumination, pose and expression to super-resolve the input face and synthesize a new facial expression of the face. Sun et. al. [84] proposed a patch similarity measurement to select data with similar textural details to the input image for their image SR approach.

In [44], Joshi et. al. presented a data-driven based personal photo enhancement system. However, it is only applied on frontal images with same expression. They addressed that the image deblurring and color correction issue are required in pre-processing stage because images are obtained from different sources even thought the subject is the same person. In addition, these global correction methods are not able to enhance the image. A MRF local model with the prior images is used to enhance the image. In general, using an appropriate dataset as prior knowledge to model the input image is the key factor of the MRF model. It significantly improves the quality of the superresolution image.

Generally, data selection plays an important role in learning-based image superresolution especially face hallucination. Human being are more sensitive to the changes in the structure and textural details of facial images than other images such as building, trees, scenery images.

Chapter 3

Batch Algorithm for Facial Structure From Motion with Additional Shape Constraints

3.1 Introduction

Recovering 3D structure from a sequence of images is one of typical interest topics in the computer vision community. In the past two decades, factorization algorithms have been widely applied to structure from motion (SFM) problems. It was first introduced to reconstruct rigid structure under arbitrary motion by Tomasi and Kanade [90]. Basically, the factorization algorithm for SFM decomposes the image feature tracks (*measurement matrix*) into motion of the camera and the 3D shape matrix via Singular Value Decomposition (SVD) and rank theorem. However, it is an ill-conditioned problem. Their linear transformations also yield valid motions and bases. Therefore, it is not possible to recover structure from the image sequence without some prior knowledge. Additional constraints such

as orthogonality of rotation matrix are required to recover the structure.

Generally, orthographic camera model is chosen as the camera model for the factorization algorithm because it is a good approximation to the perspective camera model when the reconstructed target is far from the camera and the depth variation within the target is relatively small. [85] and [72] also proposed extended factorization algorithms for perspective and paraperspective models, respectively.

Recently, recovery of different kinds of structures such as multiple linearly moving objects [39], articulated objects [98], model based non-rigid objects [16], [14], [91], [97] are reported. Model based non-rigid object recovery is attractive because many interesting non-rigid objects in nature such as human face can be represented by models. Reconstructing 3D human faces is very useful in face recognition. Compared to 2D face images, 3D face are invariant to pose changes. The pose changes significantly affect the performance of face recognition algorithms. Therefore, we can use non-rigid factorization to decompose the pose and deformation of the non-rigid structure from a image sequence.

To model the deformation of these non-rigid objects, the weighted combination of basis shapes has been applied in non-rigid SFM [16]. Using this model, Jing Xiao et al. [97] showed a closed-form solution for non-rigid SFM with rotation constraints and basis constraints. The solution is exact only when the data is noise free. The method does not work satisfactorily with noisy data[15].

In this chapter, a batch algorithm and a non-linear shape constraint optimization are proposed to improve the existing closed-form solution under noisy environments. The batch algorithm partitions the matrix and

recovers 3D structures from each partition separately. Then we apply the optimization algorithm to refine the closed-form solution of each partition based on shape constraints.

3.2 Overview of Factorization Algorithm for Non-rigid SFM

Here the camera model is assumed to be the weak perspective projection model. We also assumed that the motion is non-degenerate. Let the 2D image coordinates of P feature points over F frames denoted as $\mathbf{W} = \{\mathbf{w}_{fp} = (u_{fp}, v_{fp}) | f = 1, \ldots, F, p = 1, \ldots, P\}$, the $2F \times P$ *measurement matrix*:

$$\mathbf{W} = \begin{bmatrix} u_{11} & \ldots & u_{1P} \\ v_{11} & \ldots & v_{1P} \\ \vdots & u_{fp} & \vdots \\ \vdots & v_{fp} & \vdots \\ u_{F1} & \ldots & u_{FP} \\ v_{F1} & \ldots & v_{FP} \end{bmatrix} \tag{3.1}$$

The camera projection matrix is written as:

$$\mathbf{R}_f = \begin{bmatrix} r_{f1} & r_{f2} & r_{f3} \\ r_{f4} & r_{f5} & r_{f6} \end{bmatrix} \quad f \in \{1, \ldots, F\} \tag{3.2}$$

The non-rigid structure is represented by a linear combination of K 3D shape bases. Let $\mathbf{S}_f = \{\mathbf{s}_{fp} = (x_p, y_p, z_p) | p = 1, \ldots, P\}$ denote the 3D non-rigid structure of the f^{th} frame. Let $\mathbf{B} = \{\mathbf{b}_k = (x_{kp}, y_{kp}, z_{kp}) | k = 1, \ldots, K, p = 1, \ldots, P\}$ denote as the 3D shape bases. Then, the 3D non-

rigid structure in each frame can be represented as:

$$\mathbf{S}_f = \sum_{k=1}^{K} c_{fk}\mathbf{b}_k \quad f \in \{1, \ldots, F\} \tag{3.3}$$

where c_{fk} are the weights. Then, $\mathbf{W} = \mathbf{MB} + \mathbf{T}$ where $\mathbf{M}$ is a $2F \times 3K$ motion matrix, $\mathbf{B}$ is a $3K \times P$ 3D structure matrix and $\mathbf{T}$ is a $2F \times 1$ translation vector. When K=1, the structure is rigid. The motion matrix is the product of the weighting coefficients and the corresponding camera projection matrices. We can write this as

$$\mathbf{M} = \begin{bmatrix} c_{11}\mathbf{R}_1 & \ldots & c_{1K}\mathbf{R}_1 \\ \vdots & c_{fk}\mathbf{R}_f & \vdots \\ c_{F1}\mathbf{R}_F & \ldots & c_{FK}\mathbf{R}_F \end{bmatrix} \tag{3.4}$$

The translation vector can be obtained by computing the mean of the P feature points. The *registered measurement matrix*, $\hat{\mathbf{W}}$ is given by subtracting $\mathbf{T}$ from $\mathbf{W}$. The world origin now is placed at the centroid of the feature points, i.e.

$$\frac{1}{P}\sum_{p=1}^{P} \mathbf{w}_{fp} \quad \forall f \in \{1, \ldots, F\} \tag{3.5}$$

When the data is noiseless, the rank of $\hat{\mathbf{W}}$ is $3K$. Applying SVD, $\hat{\mathbf{W}}$ can be decomposed into a motion matrix, $\hat{\mathbf{M}}$ and a 3D basis matrix, $\hat{\mathbf{B}}$. However, it is only up to an arbitrary $3K \times 3K$ invertible transformation, $\mathbf{G}$. The exact motion matrix, $\mathbf{M}$ and 3D basis matrix, $\mathbf{B}$ can be written as:

$$\begin{aligned} \mathbf{M} &= \hat{\mathbf{M}} \cdot \mathbf{G} \\ \mathbf{B} &= \mathbf{G}^{-1} \cdot \hat{\mathbf{B}} \end{aligned} \tag{3.6}$$

The *corrective transformation matrix*, $\mathbf{G}$ is compound of K $3K \times 3$ matrix, $\mathbf{G}_k$. Then, $\mathbf{Q}_k = \mathbf{G}_k\mathbf{G}_k^T$. Computing the $\mathbf{Q}_k$ requires additional constraints. We have

$$\hat{\mathbf{M}}\mathbf{Q}_k\hat{\mathbf{M}}^T = \begin{bmatrix} c_{1k}\mathbf{R}_1 \\ \vdots \\ c_{1k}\mathbf{R}_F \end{bmatrix} \begin{bmatrix} c_{1k}\mathbf{R}_1 & \dots & c_{1k}\mathbf{R}_F \end{bmatrix} \tag{3.7}$$

Since rotation matrices are orthonormal, we have $\mathbf{R}_i\mathbf{R}_i^T = \mathbf{I}_{2\times 2}$. In [7], it was showed that using only these rotation constraints is insufficient to uniquely determine $\mathbf{Q}_k$. Thus, they also assume the first K images to be basis images. The corresponding weighting coefficients are then

$$c_{ij} = \begin{cases} 1 & \text{when } i = j \\ 0 & \text{when } i \neq j \end{cases} \tag{3.8}$$

We can now obtain a closed-form solution for each $\mathbf{Q}_k$ by optimizing the rotation and basis constraints. For the details of proof, the reader is referred to [7].

3.3 Batch Algorithm Using Matrix Partitioning

In practice, a large number of frames from video sequence are available, and using all the frames in SVD algorithm to minimize $\|\mathbf{W}-\mathbf{MB}\|_F$ may bring no advantage, firstly, because there is a large amount of redundancy in the video frames (this is just increasing the computational cost). and secondly, minimizing $\|\mathbf{W} - \mathbf{MB}\|_F$ does not guarantee that the recovered structure

is optimal. The solutions of the motion matrix $\mathbf{M}$ and the bases $\mathbf{B}$ also depend on the constraints we used on solving the corrective transformation matrix, $\mathbf{G}$.

Hence, we introduce a batch algorithm where a registered measurement matrix is partitioned into N submatrices. Then, the closed-form solution method is applied to each separately. This yields N estimates instead of a single estimate for the structures from a large number of frames. We hence expect that the proposed algorithm will improve the confidence in the result. We then propose to use these in a shape constrained non-linear optimization technique to find the best shape estimate.

Let $\Omega_i \subset \{1, \ldots, F\}$, $i = 1, \ldots, N$ be a subset of frame indexes. Then, let $\mathbf{W}_{\Omega_i} = \{(u_{fp}, v_{fp}) | f \in \Omega_i, p = 1, \ldots, P\}$ denote a row subspace of the matrix, where $|\Omega_i| \geq \max(\frac{K^2+K}{2}, 3K)$. The union of all subsets Ω_i contains all the elements of $\{1, \ldots, F\}$. All subsets are disjoint. Hence, the information in every frame is used for recovery of the structure.

Here, we assume that K is known. The set of K basis images which give the smallest condition number is the set of the most independent basis images. Thus, we can selected them as the K basis images.

Since the rank of $\hat{\mathbf{W}}_{\Omega_i}$ has to be at least $3K$, the number of frames in each partition can be determined in such a way that reasonable amount of the energy of $\hat{\mathbf{W}}_{\Omega_i}$ remains in the first $3K$ eigen-subspaces. Then each $\hat{\mathbf{W}}_{\Omega_i}$ can be decomposed by the non-rigid factorization algorithm discussed in Section 2 as:

$$\mathbf{W}_{\Omega_i} = \mathbf{M}_{\Omega_i} \mathbf{B}_i \quad i = 1, \ldots, N \tag{3.9}$$

The recovered structures are exact for noiseless data.

When $K = 1$ (rigid case), the motion matrix $\mathbf{M}$ and $\mathbf{B}$ are simplified

as rotation matrix $\mathbf{R}$ and the rigid structure matrix $\mathbf{S}$. When $K \geq 2$ (non-rigid case), we do not only need to recover the bases $\mathbf{B}$, but also the weighting coefficients in the motion matrix $\mathbf{M}$ for recovering the 3D structure. $\mathbf{M}$ can be obtained as

$$\mathbf{M}_i = \mathbf{W}\mathbf{B}_i^+ \quad i = 1, \ldots, N \tag{3.10}$$

where $\mathbf{B}_i^+$ is the pseudo-inverse of $\mathbf{B}_i$. Since the rotation matrix $\mathbf{R}_f$ is orthonormal, $||\mathbf{R}_f|| = 1$. The corresponding coefficients for each frame can be easily extracted out from motion matrix.

Let N sets of the estimated structures of the f^{th} frame denote as $\{\tilde{\mathbf{S}}_f\}_i$. Given the 3D shape bases $\mathbf{B}_i$ and the corresponding coefficients, each recovered structure can be computed by (5.5). Since each set of the recovered structures, $\{\tilde{\mathbf{S}}_f\}_i$ is independently estimated from the corresponding $\mathbf{W}_{\Omega_i}$, the reference coordinate systems of each two sets of the recovered structures are different up to a 3×3 orthonormal transformation. The orthonormal transformation can be obtained by applying Procrustes method.

3.4 Non-linear Shape Constraint Optimization

Here, we introduce an objective function which is optimized to enforce non-linear shape constraints and estimate the best recovered structure $\mathbf{S}_f$ from the set of estimated structures $\{\tilde{\mathbf{S}}_f\}_i$ from each partition. It is given as:

$$\min \sum_{n=1}^{N} \sum_{i=1}^{P} \sum_{j=1}^{P} \|s_{fi} s_{fj}^T - \tilde{s}_{fin} \tilde{s}_{fjn}^T\|^2 \quad \forall f \in \{1, \ldots, F\} \tag{3.11}$$

where N is the number of partitions. This optimization minimizes the inner products of every two feature points. In other words, we are optimizing the errors in the lengths and the mutual angles of the feature points, so we named it *metric optimization*. The metric optimization plays a role in structure refinement of the factorization method. A general-purpose quasi-Newton method [25], [83], [49], [74] is used to find the optimum solution of (5.4).

The initialization is critical for non-linear optimization problems. To avoid the solution of the metric optimization from being trapped at an unsuitable local minimum, we choose the least mean square of $\{\tilde{\mathbf{S}}_f\}_i$ as the initial value for the metric optimization.

3.5 Summary of The Proposed Approach

Our proposed algorithm is summarized as follows:

1. Partition the measurement matrix $\mathbf{W}$ into N submatrices.
2. Choose the K basis images from each subset based on their condition numbers. The set of the K basis images with the smallest condition number is the set of the most independent basis images.
3. Apply non-rigid factorization algorithm proposed by Jing Xiao et al. [7]
4. Extract the weight c_{fk} from the motion matrix $\mathbf{M}$.
5. Compute the structures by Eq. (5.5).
6. Optimize the estimated structures obtained in Step 5 by the objective function in Eq. (5.4).

Chapter 4

Facial Image Inpainting with a Learned Guidance Vector Field

In certain circumstance, a big portion of image may be corrupted during data transmission. This problem needs to be solved first before applying image superresolution. To overcome the issue, we need to retouch the corrupted image. It is known as image inpainting. The objective of image inpainting is to recover the missing region. The good news on our work is that input images are the facial images. Since input is a facial image, the structure of the face have to be preserved. Moreover, the boundary at the missing region need to be unnoticeable after retouching. Therefore, we propose a PCA based approach with a guidance vector field to retouch the missing region of the input image. The details are discussed in the following sections.

4.1 PCA Based Image Inpainting

As we mentioned, the input is a facial image and we need to preserve the structure of the input face. Thus, learning facial structure is required. We can use Principal Component Analysis (PCA) based image inpainting which learns the PCA model of the missing region from training images and then recovers the missing region from the model. In this technique, a region of interest (ROI), $\mathcal{F}$, is defined as shown in Fig. 6.2 to enclose the missing region, $\mathcal{M}$, within a bounding band of known pixels, $\mathcal{B}$. The corresponding ROI vectors, $\mathbf{F}$ from the training images are used to extract a PCA model, $\mathcal{L}$, represented by a set of orthonormal bases, $\mathbf{I}_i$ and mean, $\bar{\mathbf{F}}$. An ROI, $\mathcal{F}$, can be reconstructed from $\mathcal{L}$ as

$$\hat{\mathbf{F}} = \sum_i \alpha_i \mathbf{I}_i + \bar{\mathbf{F}} \tag{4.1}$$

where α_i are the projection weights of $\mathbf{F}$ onto the basis vectors. $\hat{\mathbf{F}}$ will differ from the original $\mathbf{F}$ in two ways: firstly, $\mathcal{M}$ will be filled in by the model, which is desirable, and secondly the original uncorrupted pixels in $\mathcal{B}$ will be altered. The latter effect is undesirable, and hence it is neccessary to impose a consistency constraint and replace all the $\mathcal{B}$ pixels in $\hat{\mathbf{F}}$ by the original pixels. This results in an updated $\hat{\mathbf{F}}$, but there is a noticeable discontinuity between the filled region, $\mathcal{M}$, and the bounding pixels, $\mathcal{B}$. In order to obtain a better estimate for $\mathcal{M}$, reconstruction of the ROI, $\hat{\mathbf{F}}$, from $\mathcal{L}$ and imposition of the consistency constraint can be iterated in a POCS algorithm until convergence. This will improve the estimate of pixels in $\mathcal{M}$; however, the discontinuity between $\mathcal{M}$ and $\mathcal{B}$ is ignored in this process, and hence may be present even though the inpainted region

looks acceptable. Hence, we propose to learn a GVF from the training data and use it in a Poisson equation based solution to obtain a smooth transition at the boundary of the missing region.

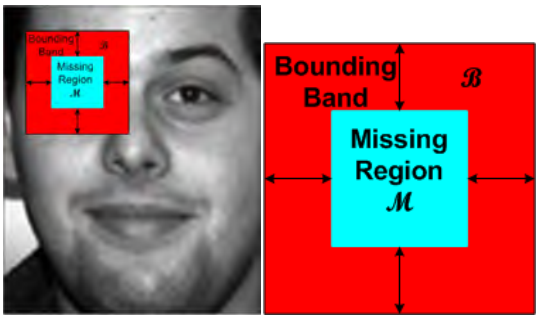


Figure 4.1: Illustration of missing region, $\mathcal{M}$, and bounding band, $\mathcal{B}$, in ROI.

4.2 Guidance Vector Field Image Inpainting

Guidance Vector Field (GVF) image inpainting was originally used for seamlessly editing image regions by Pérez et. al. [69]. The solution is obtained by solving a set of Poisson equations with Dirichlet boundary conditions with a prescribed GVF. Here, the problem can be formulated as a minimization problem:

$$\begin{aligned} \min_{f|_{\mathcal{M}}} \quad & \sum_{\{i,j\}\bigcap\mathcal{M}\neq 0}(f_i - f_j - v_{ij})^2 \\ \text{s.t.} \quad & f_i = f_i^* \qquad\qquad \forall i \in \partial\mathcal{M} \end{aligned} \tag{4.2}$$

where f_i denotes the value of the pixel at location i, $\mathcal{M}$ (as before) denotes a closed missing region, $f|_{\mathcal{M}}$ denotes the values of the pixels in $\mathcal{M}$, v_{ij} denotes the gradient at the mid point of location i and j, $\partial\mathcal{M}$ denotes the boundary subset and f_i^* denotes the given input value of pixel i. The solution can be obtained through the Gauss-Seidel solver. This inpainted

solution will have a gradient field that is close to the given GVF, with a seamless transition between $\mathcal{M}$ and the boundary, $\partial\mathcal{M}$.

However, the structure of the missing region depends highly on the GVF chosen. In the case of faces which have a unique structure in different regions, it is not possible to obtain the GVF from an undamaged region. The GVF can only learned from other sources, and here, we propose a new approach to iteratively learn the GVF from training images.

4.3 Iterative Learning of Guidance Vector Field for Image Inpainting

Our proposed method for learning the GVF for image inpainting is a POCS based iterative image inpainting technique which seeks the solution which satisfies two different constraints. The first is a structure constraint which extends the concept of the PCA based image inpainting in Section 4.1. For this, we model the high frequency details of the ROI, $\mathcal{F}$, by a set of orthonormal bases obtained by PCA. It is defined as:

$$\nabla\hat{\mathbf{F}} = \sum_i \beta_i \mathbf{h}_i + \bar{\mathbf{H}} \tag{4.3}$$

where $\bar{\mathbf{H}}$ denotes the mean of the gradient of the training ROIs, $\nabla\hat{\mathbf{F}}$ denotes the reconstructed gradient of $\mathbf{F}$, $\mathbf{h}_i$ denote the orthonormal basis vectors of the high frequency detail space and β_i are the projection weights.

The second constraint imposes smoothness through GVF based inpainting as discussed above. The $\nabla\hat{\mathbf{F}}$ obtained from the first (PCA model) constraint is used as the GVF for inpainting. Then, the solution obtained from

this second constraint is input to the PCA model and reconstructed. This process is iteratively repeated until convergence. The solution will satisfy the facial structure constraints, and also have a seamless boundary between $\mathcal{M}$ and $\mathcal{B}$. The algorithm is summarized in Algorithm 1. The algorithm complexity is $O(m^2)$ where m is $|\mathcal{F}|$.

Input: ROI from n training images, $\mathbf{L}_i$, $i \in \{1, \ldots, n\}$, ROI of a corrupted image, $\mathbf{F}$, with $\mathcal{M} = \{\mathbf{0}\}$.
Output: ROI of an inpainted Image, $\hat{\mathbf{F}}$.

1. Compute the gradients $\{\nabla \mathbf{L}_i\}$ of $\{\mathbf{L}_i\}$.
2. Apply PCA on $\{\nabla \mathbf{L}_i\}$ to obtain $\{\mathbf{h}\}$ and $\bar{\mathbf{H}}$.
3. Compute the gradient $\nabla \mathbf{F}$ of $\mathbf{F}$.
4. Compute $\beta_i =< \nabla \mathbf{F}, \mathbf{h}_i >$.
5. $\nabla \hat{\mathbf{F}} = \sum_i \beta_i \mathbf{h}_i + \bar{\mathbf{H}}$.
6. Obtain the GVF, $\{v_{ij}\}$, from the corresponding elements of $\nabla \hat{\mathbf{F}}$.
7. Solve Equation 4.2 by Gauss-Seidel solver to obtain $\hat{\mathbf{F}}$.
8. $error = \sum_i (\mathbf{F}(i) - \hat{\mathbf{F}}(i))^2$.
9. $\mathbf{F} \leftarrow \hat{\mathbf{F}}$.
10. **Repeat** step 3-10 until $error < \delta$ or $t > T$ (maximum iteration).

Algorithm 1: Algorithm for POCS based image inpainting.

4.4 Patch Selection for Learning PCA Model

The face is a non-rigid structure which changes significantly with facial expression. For example, there is a large difference between the happy face (open mouth with teeth visible) and the neutral face (mouth closed). To account for this variation, we manually mark points on all the training

images (this can be easily automated, for example, by using active shape model), and form a mesh by Delaunay triangulation. Following this, all training images are warped onto the triangulated mean face mesh. Here, we assume that only one triangular patch is corrupted and needs to be inpainted (though it is easy to generalize this). Hence, a corrupted patch will have three neighboring triangular. Here, instead of using the neighboring patches from all the training images, we propose a patch selection algorithm to select the appropriate training images from which the patches will be used for learning.

Let $\mathcal{M}$ and $\{\mathcal{B}_i | i = 1, \ldots, m\}$ (here $m = 3$) denote the missing region of the input image and its neighboring patches, respectively, and let $\mathcal{M}^j$ and $\mathcal{B}_i^j$ denote the corresponding patches in the training image j. The subscript i of $\mathcal{B}_i$ denotes the index of the patch. The distance between each $\mathcal{B}_i^j$ and $\mathcal{B}_i$ is measured by the $L2$-norm, $d(\mathcal{B}_i^j, \mathcal{B}_i)$. The k (we use $k = 5$) training images with the smallest d are selected and the corresponding indexes of these k training images form an index set, $\mathcal{I}_i$. The indexes of the images used to form the training set is given by $\cap \mathcal{I}_i$. The algorithm is summarized in Algorithm 2.

Input: ROI from n training images, $\mathbf{L}_j$, $j \in \{1, \ldots, n\}$, ROI of a corrupted image, $\mathbf{F}$

Output: Selected training images, $\{\mathbf{L}_s | \{s\} \subset \{j\}, |\{s\}| \leq k\}$.

forall the i **do**

  forall the j **do**

    Compute $d(\mathcal{B}_i^j, \mathcal{B}_i)$.

  end

  1. Sort the training data based on d.
  2. $\mathcal{I}_i \leftarrow$ { corresponding indexes of the k nearest training images}

end

Images with index in $\cap \mathcal{I}_i$, $\{\mathbf{L}_s\}$, are selected as training images for Algorithm 1.

Algorithm 2: Algorithm for patch selection.

Chapter 5

Image Superresolution on Generic Face

5.1 Edge Model

A low resolution image can be interpreted as an image which has been obtained by blurring a high resolution image and then downsampling it, causing loss of high frequency details. Simple interpolation techniques can only increase the size of the image, but the blurred quality of the image essentially remains unchanged. Hence, it is necessary to estimate and add back the high frequency details. These missing details in the image can be defined as

$$\mathbf{E} = \mathbf{I_{HR}} - \tilde{\mathbf{I}}_{\mathbf{LR}} \tag{5.1}$$

where $\mathbf{I_{HR}}$ denotes the (unknown) original high resolution image, $\mathbf{I_{LR}}$ is the given low resolution, downsampled image and $\tilde{\mathbf{I}}_{\mathbf{LR}}$ is its interpolated version.

We propose to estimate $\mathbf{E}$ from a training data set and add it to $\tilde{\mathbf{I}}_{\mathbf{LR}}$ to obtain the superresolved image. Let $\mathcal{H}$ be the space of high resolution training images. We then blur and downsample the images in $\mathcal{H}$, to obtain low resolution training images in space, $\mathcal{L}$. Ideally, the blurring used should be the same as that present in $\mathbf{I}_{\mathbf{LR}}$, though in practice, a reasonably good guess will be viable.

Both spaces can be represented by an orthonormal basis obtained by principal component analysis (PCA):

$$\mathbf{I}_{\mathbf{HR}} = \sum_i \alpha_i \mathbf{h}_i + \bar{\mathbf{H}} \tag{5.2}$$

$$\mathbf{I}_{\mathbf{LR}} = \sum_i \beta_i \mathbf{l}_i + \bar{\mathbf{L}} \tag{5.3}$$

where $\bar{\mathbf{H}}$, $\bar{\mathbf{L}}$ are the means of the images in $\mathcal{H}$ and $\mathcal{L}$, respectively. $\mathbf{h}_i$ and $\mathbf{l}_i$ are the corresponding orthonormal basis vectors of the two spaces, and α_i, β_i are the projection weights to represent images $\mathbf{I}_{\mathbf{HR}}$, $\mathbf{I}_{\mathbf{LR}}$ in the two spaces.

To represent the high frequency details lost in the $\mathcal{H} \rightarrow \mathcal{L}$ transformation, we construct an interpolated space, $\tilde{\mathcal{L}}$, by interpolating $\{\mathbf{l}_i\}$ to $\{\tilde{\mathbf{l}}_i\}$, where the latter have the same vector dimension as $\{\mathbf{h}_i\}$. The interpolation causes the vector length to change from $\|\mathbf{l}_i\| = 1$ to $\|\tilde{\mathbf{l}}_i\| = 2$, though $\{\tilde{\mathbf{l}}_i\}$ remain orthogonal. Hence, the $\{\tilde{\mathbf{l}}_i\}$ are re-normalised to unit length to provide an orthonormal basis for $\tilde{\mathcal{L}}$. Assuming negligible aliasing errors in downsampling and reconstruction errors in interpolation, $\tilde{\mathcal{L}}$ is simply a blurred or lowpassed subset of $\mathcal{H}$, $\tilde{\mathcal{L}} \subset \mathcal{H}$.

We can now define the space of missing high frequency details, $\mathcal{E}$ (or

simply edge space) as $\mathcal{H} - \tilde{\mathcal{L}}$, and we can write

$$\mathbf{E} = \mathbf{I_{HR}} - \tilde{\mathbf{I}}_{\mathbf{LR}} = \sum_i \alpha_i \mathbf{h}_i - \sum_i \beta_i \tilde{\mathbf{l}}_i + \bar{\mathbf{H}} - \bar{\tilde{\mathbf{L}}} \tag{5.4}$$

In practice, the mean images of the high resolution and low resolution spaces will be about the same, so that $\bar{\mathbf{H}} - \bar{\tilde{\mathbf{L}}} \approx \mathbf{0}$. Also, given an image, $\mathbf{I_{HR}}$, and its low resolution versions, $\mathbf{I_{LR}}$, $\tilde{\mathbf{I}}_{\mathbf{LR}}$, we can expect that $< \mathbf{h}_i, \tilde{\mathbf{l}}_j > \approx 1$ for $i = j$ and ≈ 0 for $i \neq j$. Thus, $\beta_i \approx \tilde{\beta}_i \approx \alpha_i$, where $\beta_i =< \mathbf{I_{LR}}, \mathbf{l}_i >$, $\tilde{\beta}_i =< \tilde{\mathbf{I}}_{\mathbf{LR}}, \tilde{\mathbf{l}}_i >$ and $\alpha_i =< \mathbf{I_{HR}}, \mathbf{h}_i >$. This yields

$$\mathbf{E} = \sum_i \beta_i \mathbf{e}_i \tag{5.5}$$

where $\mathbf{e}_i \triangleq \mathbf{h}_i - \tilde{\mathbf{l}}_i$ is a basis defining the edge space (or Laplacian space). Figure 5.1 illustrates the geometric relationship between $\mathbf{h}_i$, $\tilde{\mathbf{l}}_i$ and $\mathbf{l}_i$ in 2D.

Using the above ideas, given an image $\mathbf{I_{LR}}$, it is first projected onto the space $\mathcal{L}$ to obtain $\{\beta_i\}$, and also interpolated to yield $\tilde{\mathbf{I}}_{\mathbf{LR}}$. The high frequency details for this image are estimated by Equation 5.5, and an SR image is obtained as

$$\hat{\mathbf{I}}_{\mathbf{HR}} = \mathbf{E} + \tilde{\mathbf{I}}_{\mathbf{LR}} \tag{5.6}$$

However, we must ensure that $\hat{\mathbf{I}}_{\mathbf{HR}}$ produced by using high frequency details from the training set is consistent with the given $\mathbf{I_{LR}}$. We consider this next.

5.2 Backprojected Error Correction

Directly adding $\mathbf{E}$ to the interpolated image, $\tilde{\mathbf{I}}_{\mathbf{LR}}$, can produce artifacts in the SR image, especially if $\mathbf{I_{LR}}$ is not a member of the training set, and

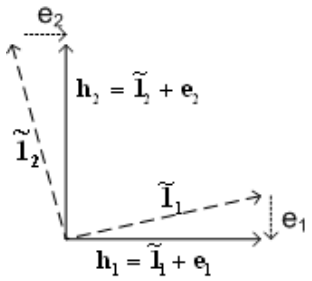


Figure 5.1: A 2D geometric illustration of the relationships between the vectors.

the blur present in $\mathbf{I_{LR}}$ is different from the blur used to generate the low resolution space, $\mathcal{L}$. Eliminating these artifacts needs another constraint arising from the given image, $\mathbf{I_{LR}}$. Thus the SR image we seek must satisfy the model constraint - the high frequency information added must be consistent with the model - and the data constraint which ensures that the SR image is consistent with the given image. These two constraints are alternately used in the iterative POCS method to obtain the SR image. In several POCS based methods for extrapolation, some original data is available to serve as a hard constraint. This is not the case here; all that is available are the blurred pixels in $\mathbf{I_{LR}}$. Hence, for the data constraint, we use a correction using backprojected error similar to [22].

Let $\mathfrak{F}$ denote the blurring and down sampling operation which produces $\mathbf{I_{LR}}$ and let $\mathfrak{G}$ denote the interpolation operation to produce $\tilde{\mathbf{I}}_{\mathbf{LR}}$. The error, ε, is defined as:

$$\varepsilon = \mathbf{I_{LR}} - \mathfrak{F}(\hat{\mathbf{I}}_{\mathbf{HR}}) \tag{5.7}$$

Obviously, $\varepsilon = \mathbf{0}$ only if $\mathfrak{F}(\hat{\mathbf{I}}_{\mathbf{HR}}) = \mathbf{I_{LR}}$. Thus, ε is backprojected by $\mathfrak{G}$ to correct $\hat{\mathbf{I}}_{\mathbf{HR}}$ obtained from the first constraint. The new SR image, $\hat{\mathbf{I}}^{new}_{\mathbf{HR}}$ is defined as

$$\hat{\mathbf{I}}^{new}_{\mathbf{HR}} = \hat{\mathbf{I}}^{old}_{\mathbf{HR}} + \mathfrak{G}(\varepsilon) \tag{5.8}$$

In the above error correction rule, we use $\mathfrak{G}(\varepsilon)$ instead of $\mathfrak{F}^{-1}(\varepsilon)$. This is because the blur operator in $\mathfrak{F}$ is unknown, and hence the corresponding deblurring operator cannot be specified in $\mathfrak{F}^{-1}$. For simplicity, we ignore the deblurring and simply approximate $\mathfrak{F}^{-1}$ by an interpolation operation, $\mathfrak{G}$.

Backprojecting the error, ε, serves to reduce the artifacts produced by the first constraint while causing the estimate to deviate somewhat from the learned model. Hence, $\mathbf{E}$ is reestimated with $\hat{\mathbf{I}}_{\mathbf{HR}}^{new}$. The POCS algorithm is iteratively applied with these two constraints until the iterations converge.

5.3 POCS Algorithm

Our POCS based SR method using the learned edge model and backprojected error basically consists of a pre-processing (training) part and an iterative part. The training computes the principal components to represent the spaces and the main SR algorithm consists of the iterative part. They are summarized below.

Input: a set of n high resolution training images, $\mathbf{I}^i_{\mathbf{HR}}$,

$i \in \{1, \ldots, n\}$, a low resolution image, $\mathbf{I}_{\mathbf{LR}}$

Output: Super-resolved Image, $\hat{\mathbf{I}}_{\mathbf{HR}}$

Pre-Processing Part:

1. Simulate blurring and down sampling on high resolution training images to generate a set of n low resolution training images, $\mathbf{I}^i_{\mathbf{LR}}$, $i \in \{1, \ldots, n\}$.

2. Interpolate the $\mathbf{I}^i_{\mathbf{LR}}$ to obtain $\tilde{\mathbf{I}}^i_{\mathbf{LR}}$, $\tilde{\mathbf{I}}^i_{\mathbf{LR}} = \mathfrak{G}(\mathbf{I}^i_{\mathbf{LR}})$, $i \in \{1, \ldots, n\}$.

3. Apply PCA on $\{\mathbf{I}^i_{\mathbf{HR}}\}$ and $\{\mathbf{I}^i_{\mathbf{LR}}\}$ to obtain $\{\mathbf{h}_i\}$ and $\{\mathbf{l}_i\}$, respectively.

4. Interpolate and normalize $\mathbf{l}_i$ to obtain $\tilde{\mathbf{l}}_i$.

5. Compute $\mathbf{e}_i = \mathbf{h}_i - \tilde{\mathbf{l}}_i$.

Iterative Part:

Compute $\beta_i =< \mathbf{I_{LR}}, \mathbf{l}_i >$

Interpolate: $\tilde{\mathbf{I}}_{\mathbf{LR}} = \mathfrak{G}(\mathbf{I_{LR}})$

repeat

1. $\mathbf{E} = \sum_i \beta_i \mathbf{e}_i$
2. $\hat{\mathbf{I}}_{\mathbf{HR}} = \tilde{\mathbf{I}}_{\mathbf{LR}} + \mathbf{E}$
3. $\varepsilon_t = \mathbf{I_{LR}} - \mathfrak{F}(\hat{\mathbf{I}}_{\mathbf{HR}})$
4. $\hat{\mathbf{I}}_{\mathbf{HR}} \leftarrow \hat{\mathbf{I}}_{\mathbf{HR}} + \mathfrak{G}(\varepsilon_t)$
5. $\beta_i =< \hat{\mathbf{I}}_{\mathbf{HR}}, \mathbf{h}_i >$
6. $error_t \triangleq \sum_{i,j} \varepsilon_t^2(i,j)$

until $|error_t - error_{t-1}| < \delta$ or $t > T$(maximum iteration).

Chapter 6

Image Superresolution on Specific Face

In this chapter, we propose an image superresolution approach on specific face. Here, we assume that the training images acquired by different cameras but consists of a same person's face with different expressions and poses. It is very common in practice. Some examples are shown in Figure 6.1. Nowadays, most people have a lot of personal photos which recorded their memorable events in their personal computer. Since we are not professional photographers, some of them are not desirable. Those low quality images usually are taken by low resolution cameras such as webcams, inexpensive pocket cameras, mobile phones etc. In certain circumstance, the images have to be taken by cameras from a long distance. The subjects in these images usually are unclear due to their low resolution and poor quality of the lenses and camera sensors. Therefore, simple image interpolation approaches which enlarge the size of the images are not able to fully resolve the problem here. The textural details need to be enhanced but we only have a single image with unique pose and expression. However,

there are a lot of other high quality images in the album. Although the expressions and poses in these images are not exactly same as the low quality image to be enhanced, these high resolution image still can be used as examples to improve the low quality images. In our approach, we establish a high resolution training data set of a same person's face with different expression and poses. An image retrieval based on facial action unit is introduced to retrieval training data with similar pose and expression to the input image. The details of the image retrieval are discussed in section 6.1. A learning-based image super-resolution and image enhancement based on MRF is proposed to reconstruct the superresolution image of the input. The details are given in section 6.3.

Figure 6.1: Samples of input images captured under different lighting conditions and camera models

6.1 Intelligent Image Selection

Retrieval appropriated training data plays an important role in image superresolution. Training images which are similar to the input image can narrow down the search space for estimating the superresolution image of the input. It also can avoid inappropriate patches been selected to generate the irregularities in the output of the image superresolution. Moreover, smaller search space improves the computational time of the optimization problem in image superresolution. In our proposed approach, there are three retrieval criteria based on pose, shape and texture. First the pose of input face is estimated based on its shape ratios. The input is classified into the corresponding pose category to narrow down the search space. Then, the shape and texture at each portion of face are estimated separately and represented as an expression descriptor. The training images with similar expression and pose are selected based on the descriptor. The flowchart of the image retrieval system is showed in Figure 6.2.

6.1.1 Pose Discrimination

Pose discrimination is used to find the facial image with similar pose. Thus, it is not required to recover the exact pose angle of the face. The method primarily relies on a general assumption that faces are a planar object. Each facial image is marked n facial feature points, p_i. An example is showed in Figure 6.3. These feature points are tracked by the Kanade-Lucas-Tomasi (KLT) Feature Tracker[89]. The corresponding feature points in the low resolution input query image are manually marked due to the poor image quality. Some feature detectors can be applied on certain images if the images have a good quality and lighting condition.

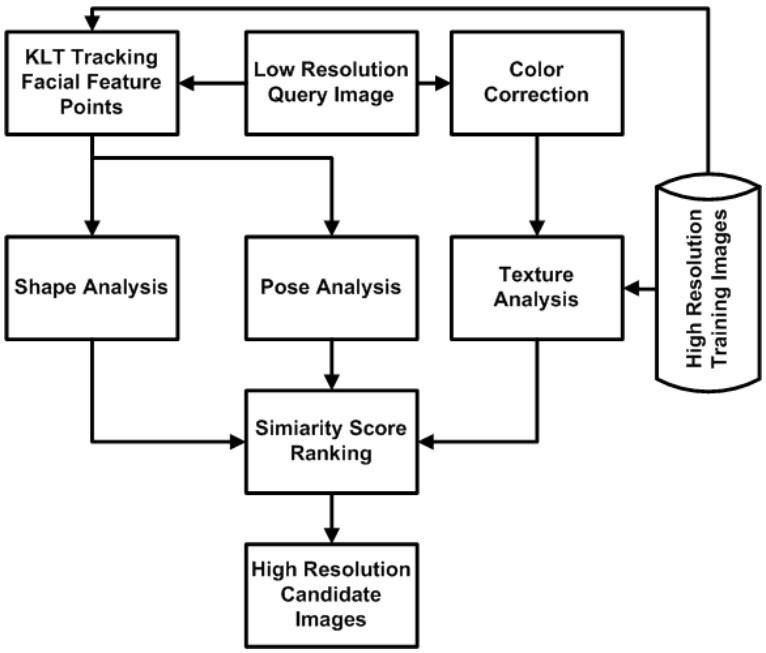


Figure 6.2: The flowchart of the high resolution candidate image retrieval system for image superresolution

As we can see in Figure 6.4, the structures of facial images are very different when the pose is changed. Even though each images has been marked with the feature points, the images are very difficult to be aligned properly due to the sparsity of the feature points and the significate pose changes. Therefore, we need to select the high resolution images with similar pose to the query image as the training image. Since the images have been tracked by the facial feature points, small changes in pose can be tolerate. Image aligned and warping can be applied to remove the small change.

Facial pose changes can be interpreted as the head rotation about three orthogonal axes. There are roll, pitch and yaw which refer to rotations about the respective axes as shown in Figure 6.5(a). Since faces can be aligned by the feature points, the translation can be resolved by aligned the means of the feature points. $p'_i = p_i - \frac{1}{n}\sum_n p_i$. Changes in roll and

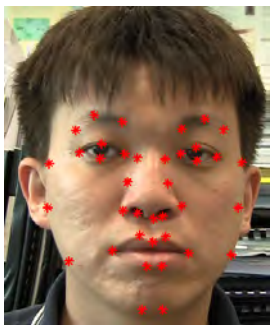


Figure 6.3: A high resolution training image with 38 marked feature point.

Figure 6.4: Some high resolution training images with different poses.

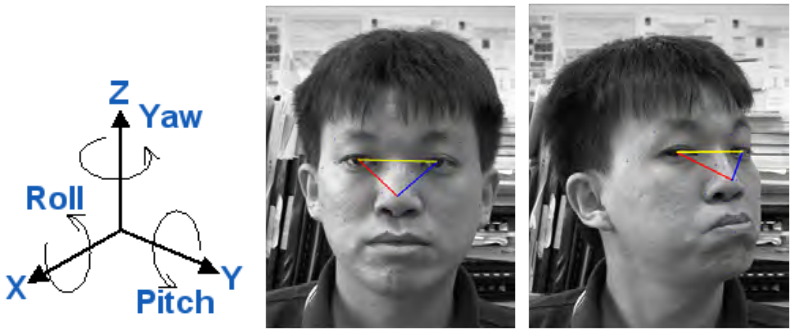
pitch are very limited. The structures of faces change in 2D images due to roll and pitch are not significate. Thus, we only need to measure the yaw changes here. To measure the similarity of yaw, a pose ratio is defined as

$$R_1 = \frac{l}{r} \tag{6.1}$$

where l is the distance of the center of left eye to the nose tip and r is the distance of the center of right eye to the nose tip.

6.1.2 Shape Analysis

Since faces are non-rigid objects, the shape structure of faces is varied with the different facial expressions especially the region at eyes and mouth. Some facial images with different expressions are showed in Figure 6.6.



(a) (b) (c)

Figure 6.5: (a) Illustration of 3D body rotations about three orthogonal axes. These rotations are referred to as yaw, pitch and roll. (b) and (c) illustrate the ratio R_1 are affected by pose change. When face turn to right, the distance from the center of left eye to the nose tip (red line) is longer than the center of right eye to the nose tip (blue line).

Even though the pose of a face remains unchanged, the structure and textural details of the face can be changed significantly due to its expression. To select appropriate candidate images as training data, we need to retrieve images with similar expression. In this session, a shape analysis to retrieve images with similar expression is investigated. The texture analysis is discussed in Section 6.1.4.

As discussed in Section 6.1.1, the training images have been tracked n feature points by KLT method. These feature points can be used on the shape analysis here. Since the changes in eyes and mouth affect the structure of face significantly, the similar expression here is defined based on the shape of the eye and mouth regions.

First, feature points at eye and mouth regions are extracted. The feature points at left eye region, right eye region and mouth region are denoted as $\mathbf{P}_i^l$, $\mathbf{P}_i^r$ and $\mathbf{P}_i^m$, respectively. Each set of the feature points is used to fit an ellipse. Fitting the ellipse can be done by principal component analysis (PCA). The maximum spread is along the major axis of the ellipse, $\mathbf{a}$ which

Figure 6.6: High resolution training images with same pose but different expressions.

is also known as the first principal component direction. The minor axis of the ellipse, **b**, is the second principal component direction. It is orthogonal to the first principal component direction. The mean and covariance of the k feature points at the region of interest, $\mathbf{P}_i$, are defined as

$$\bar{\mathbf{P}}_i = \frac{1}{k}\sum_{i=1}^{k}\mathbf{P}_i \tag{6.2}$$

$$\Sigma = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{P}_i - \bar{\mathbf{P}}_i)(\mathbf{P}_i - \bar{\mathbf{P}}_i)^T \tag{6.3}$$

The mean, $\bar{\mathbf{P}}_i$ is the center of the ellipse. The principal components can be obtained as the eigenvectors of Σ.

The two eigenvalues obtained from PCA are denoted as a and b. The ratio of them is defined as

$$D_{shape} = \frac{b}{a} \tag{6.4}$$

The ratio is scale invariant. Thus, the ratio is not affected by image resolution. Some examples are showed in Figure 6.7.

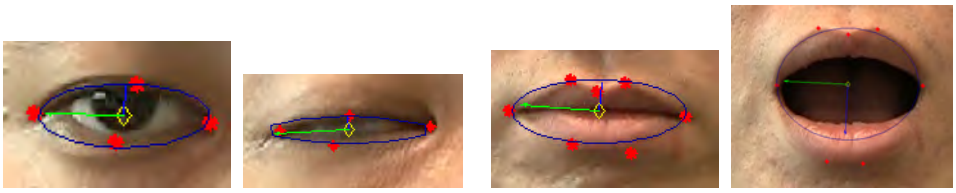


Figure 6.7: Fitting ellipse on the regions of interest (eyes and mouth).

6.1.3 Color Constancy

Since input images and training images are acquired from different cameras and under different lighting conditions, having a consistent color distribution of images is important in texture analysis. Figure 6.8 shows the images obtained under different conditions. The color distributions of these images are very different. The skin color of the same person varies significantly according to the lighting conditions and camera sensors.

(a) (b) (c) (d)

Figure 6.8: Images capture from different camera and under different lighting conditions. (a) Image captured by HD camera (b) Image captured by mobile phone camera (c) Image captured by video camera (indoor environment) (d) Image captured by video camera (outdoor environment)

To overcome this issue, a histogram equalization is applied on the input image first. A modified input image with similar color distribution of

the training images is generated for the texture analysis in the next step. The idea of the histogram equalization is illustrated in Figure 6.9. In our approach, the histogram equalization is applied on CIELAB color space. After input and training images are transformed from RGB to CIELAB color color space, their histograms in each color space are computed independently and normalized. The three normalized histograms from CIELAB color space can be interpreted as their color probability density functions (pdf). Assuming these color pdf are independent, the equalization transformation can be done independently.

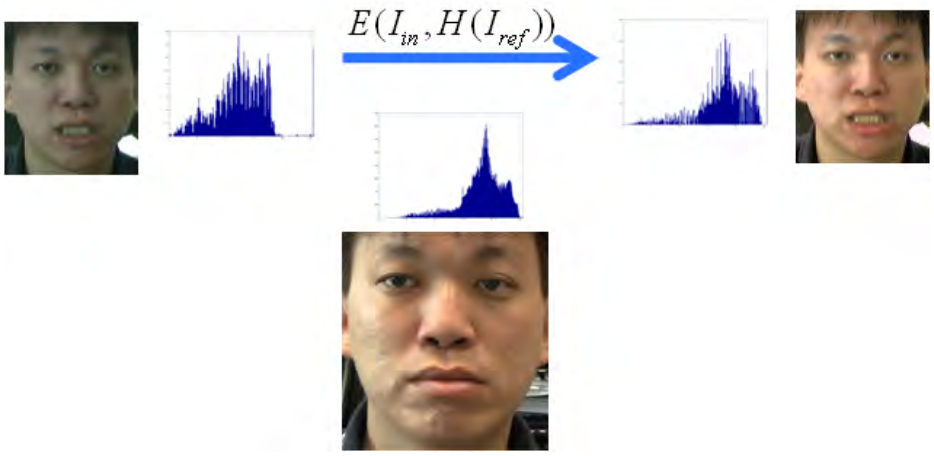


Figure 6.9: Illustration of the color correction by histogram equalization

The input histogram and the reference histogram are denoted as $H(m)$, $H(n)$, respectively. Assuming that the resolution of image is N and the scale is $[p_0, p_k]$, the equalization transformation $\mathfrak{T}(p)$ can be derived as

$$\mathfrak{T}(p) = \frac{p_k - p_0}{N} \int_{p_0}^{p} H(u)du + p_0 \tag{6.5}$$

It is noted that the equalization transformation is monotonic. Let $f(m)$ and $g(n)$ denote the equalized input histogram and the equalized reference histogram, respectively. $g^{-1}(n)$ denote the inverse function of $g(n)$. Since

$f(m)$ and $g(n)$ are equalized, $g^{-1}(f(m))$ transforms the input image to an image with similar histogram of the reference image. To resolve the inverse transformation above, we build a lookup table to map the histograms to minimize

$$|C_m(k) - C_n(\mathfrak{T}(k))| \tag{6.6}$$

where C_m is the cumulative histogram of input image, C_n is the cumulative sum of reference histogram for all intensities k. Moreover, $C_n(\mathfrak{T}(k))$ cannot overshoot $C_m(k)$ by more than half the distance between the histogram counts at k. The algorithm is summarized in Algorithm 3.

1. Compute the CIELAB color histograms of input image and normalize them

2. Equalize each histogram independently by Equation 6.5

3. Compute the CIELAB color histograms of reference image and normalize them

4. Equalize each histogram independently by Equation 6.5

5. For each bin k, find $C_m(k)$ and the corresponding bin j in $C_n(j)$ that minimize Equation .

6. Build the lookup table, $lookup[k] = j$ and convert input image to an image with the reference histogram

Algorithm 3: Algorithm for color correction.

6.1.4 Texture Analysis

The color distribution of the input image are corrected by the method mentioned in Section 6.1.4. Hence, the modified input image and training

images can be assumed to have similar color distribution. In this section, a facial texture analysis is investigated for our image retrieval system. Our objective is to represent the input image and training images by its local spatial information. The Pyramid Histogram of Orientation Gradients (PHOG) proposed by Bosch et. al. [12] is used as the texture descriptor.

Histogram of Orientated Gradients (HOG) [26] descriptor consists of a histogram of edge orientation gradients weighted by its corresponding magnitude within an image subregion quantized into K bins. First, the edge image is computed by Canny edge detector [18]. Next, the gradient magnitude and gradient orientation are computed. The edge image, the gradient magnitude image and the gradient orientation image are divided into 2^l cells for l level. At each cell, the histogram of orientation gradients is computed.

PHOG descriptor is concatenated HOG descriptor at each pyramid resolution level. The pyramid at level l has 2^l image subregions along each dimension or 4^l subregions in total. Each subregion is represented by a K-vector corresponding to the K bin of the histogram. Thus, the dimensionality of PHOG descriptor is $K \sum_{l=0}^{L} 4^l$ where L is the total number of the pyramid levels.

Since the texture information at the center is more important than boundary region, we proposed an Extended PHOG (EPHOG) to emphasize the representation of the center subregion. At the center of the given images, another PHOG descriptor is applied on it. EPHOG is concatenated the original PHOG and the new PHOG descriptor which represents the center of the image. The idea of EPHOG is illustrated in Figure 6.10.

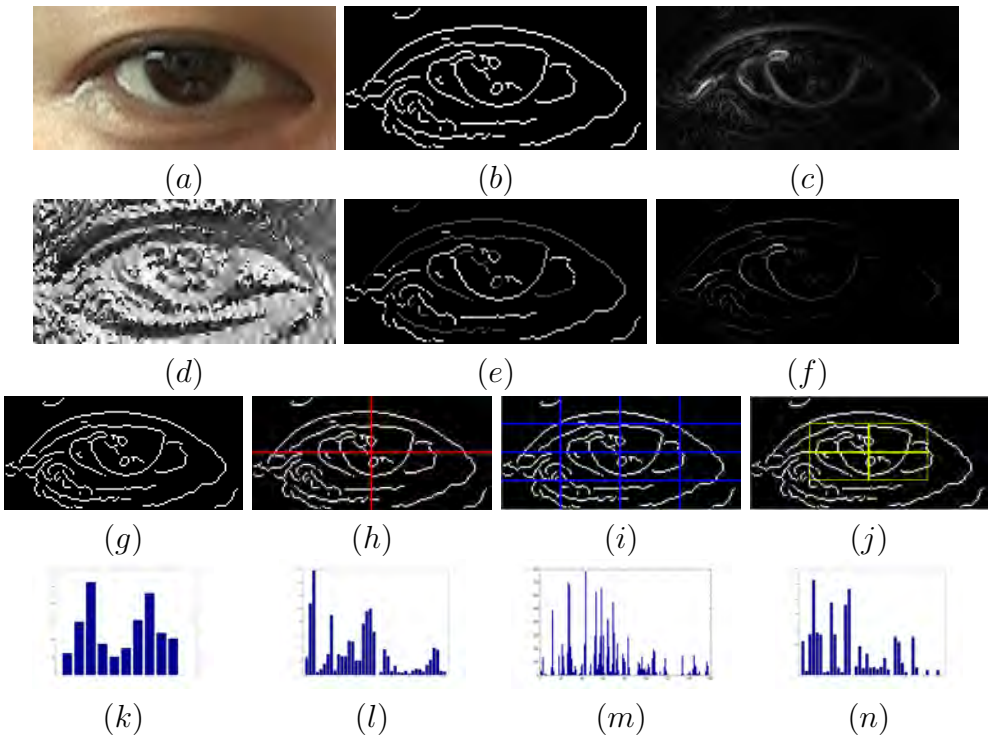


Figure 6.10: (a) A color image, (b) Canny edge of image (a), (c) Gradient magnitude of image (a), (d) Gradient angle of image (a). (e)the corresponding bin number of the gradient angle along the edges. (f) the gradient magnitude along the edges. (g)-(i)are the grids for $l = 0$ to $l = 2$ pyramid levels in the conventional PHOG. (j) is the grid for level $l = 1$ in the proposed EPHOG. Only the four cells at the center of the images are computed. (k)-(n) are the corresponding HOG descriptors of (g)-(j).

6.1.5 Similarity Measurement for Image Retrieval System

The image retrieval system aims to find a set of high resolution training images which are similar to input image in pose, shape and textural details. Firstly, the pose ratio, R_1 is used to classify the training images into five classes. The R_1 of input image is computed and the corresponding training images are selected for the shape and texture analysis. To integrate the similarity of shape in 6.1.2 and texture descriptors in Section 6.1.4, a

similarity score is defined as

$$P(D_S^I, D_T^I|x) = P(D_S^I|x)P(D_T^I|x) \tag{6.7}$$

where the shape score is defined as

$$P(D_S^I|x) = 1 - \frac{(D_S^x - D_S^I)^2}{\max_{\{x\}}(D_S^x - D_S^I)^2} \tag{6.8}$$

The D_S^x and D_S^I are the shape ratio defined in Equation 6.4. The maximum of $(D_S^x - D_S^I)^2$ is used for normalization. The interval of the shape score is $[0, 1]$. The texture score is defined as

$$P(D_T^I|x) = 1 - \frac{\chi^2(D_T^x, D_T^I)}{\max_{\{x\}} \chi^2(D_T^x, D_T^I)} \tag{6.9}$$

D_T^x and D_T^I are the EPHOG descriptor vectors of the training image, x and input image, I, respectively. The χ^2 distance is used to measure the distance between two EPHOG decriptor vectors. The smaller χ^2 distance implies the more similar between two images. High resolution training images with high score are selected as the candidates for image superresolution in Section 6.3.

6.2 Image Alignment

Since the images are captured by different cameras, the images are unlikely aligned properly. Moreover, face is a non-rigid object. Expression changes distorts the structure of face significantly and make the alignment more difficult. To overcome this issue, we triangulate the feature points by Delaunay Triangulation and apply an affine warping on each corresponding

triangle between high resolution training images and input image. The image alignment would not only ensure the input image is aligned with the training image, but also help to narrow down the search space in the image superresolution approach in the next section.

6.3 MRF Model for Face Hallucination

After the high resolution training images are selected and aligned, we need to match the patches in input image with the patches in these candidates subject to the constraint that the overlap region between two adjacent patches are smooth. Thus, a patch-based nonparametric Markov random field (MRF) model is proposed to minimize the energy function.

$$E(x|\theta) = \sum_{s \in \mathcal{V}} \theta(x_s, p_s) + \sum_{(s,t) \in \mathcal{E}} \rho_{st}(x_s, x_t, p_s, p_t) \tag{6.10}$$

where set $\mathcal{V}$ denote the image patches obtained from the selected training data, p_s at coordinate x_s, θ_s denote the data penalty function and ρ_{st} denote the smoothness function of patch p_s and patch p_t and s and t denote the patch indices. To minimize the energy function, $E(x|\theta)$, we use the sequential tree-reweighted message passing algorithm proposed in [46].

The input image is a poor quality low resolution image. It can be blur, noisy, over- or underexposure image. In order to generating a visually-pleasing and aesthetically attractive facial image, a data penalty function based on the gradient and color information is proposed. In general, our eyes are more sensitive to certain colors than others. Using typical L2-norm to measure the difference between two color patches is not appropriate because our color perception is non-uniform. In addition, our eyes is also

more sensitive to the region with large gradient changes such as edges. Moreover, edges are important components for preserving the structure of the face. Thus, we imposed a color cost and an edge cost into the penalty function, $\theta(x_s, p_s)$.

$$\theta(x_s, p_s) = D_I(x_s, p_s) D_G(x_s, p_s) \tag{6.11}$$

The smoothness function, ρ_{st} is a constraint to ensure that the overlapping region of the adjacent patches must be as similar as possible. It is defined as

$$\rho_{st}(x_s, x_t, p_s, p_t) = D_I^{\Omega}(x_s, x_t, p_s, p_t) D_G^{\Omega}(x_s, x_t, p_s, p_t) \tag{6.12}$$

ρ_{st} is also imposed a color cost, D_I^{Ω} and an edge cost D_G^{Ω} on the overlapping region, Ω.

6.3.1 Color Constraint

The color cost function,$D_I(x_s, p_s)$, is defined as

$$D_I(x_s, p_s) = 1 - \exp\left(-\lambda \Delta E_{CIE_{00}}(I(x_s, p_s), I_y)\right) \tag{6.13}$$

I_y is the patch extracted from the selected training images and $I(x_s, p_s)$ is the input patch to be optimized. $E_{CIE_{00}}$ is an extension of the L2-norm color difference function with five additional corrections on lightness, chroma, hue and chroma-hue interaction to resolve the perceptual uniformity issue defined by CIE [78]. The scale of D_I is $[0, 1]$. If two patches are similar in color, the color penalty is small.

Similarly, the color smoothness function, D_I^{Ω} is defined as

$$D_I^{\Omega}(x_s, p_s, x_t, p_t) = 1 - \exp\left(-\lambda \Delta E_{CIE_{00}}(I^{\Omega}(x_s, p_s), I^{\Omega}(x_t, p_t))\right) \quad (6.14)$$

where $I^{\Omega}(x_s, p_s)$ and $I^{\Omega}(x_t, p_t)$ are the overlapping regions of two adjacent patches.

6.3.2 Edge Constraint

Since we would like to preserve the edge information especially the strong edge information, an edge cost function is defined as

$$D_G = \begin{cases} 1 - \exp\left(-\lambda_1 \|G_x - G_y\|_2\right), & \|G_x\| > \varepsilon; \\ 1 - \exp\left(-\lambda_2 \|G_x - G_y\|_2\right), & \|G_x\| \leq \varepsilon; \end{cases} \quad (6.15)$$

where G_x and G_y are the gradient magnitude in input image I_x and training image, I_y, respectively. It is noted that $\lambda_1 \gg \lambda_2$ to preserve the strong edge. In addition, small $\|G_x\|$ is likely noise.

Similarly, the edge smoothness function, D_G^{Ω} is defined as

$$D_G^{\Omega}(x_s, p_s, x_t, p_t) = \begin{cases} 1 - \exp\left(-\lambda_1' \|G_{x_s} - G_{x_t}\|_2\right), & \|G_{x_s}\| > \varepsilon, \|G_{x_t}\| > \varepsilon; \\ 1 - \exp\left(-\lambda_2' \|G_{x_s} - G_{x_t}\|_2\right), & \text{otherwise}; \end{cases} \quad (6.16)$$

Chapter 7

Experiments and Results

7.1 Experiments on Structure From Motion

The proposed factorization algorithm with metric optimization is evaluated quantitatively and qualitatively on synthetic data and facial expression images, respectively. In the quantitative evaluation, the new approach was applied on rigid and non-rigid synthetic data sets. In the qualitative evaluation, a set of human facial expressions was used to examine the performance of the approach. The results are presented below.

7.1.1 Quantitative Evaluation on Synthetic Data

In this section, three approaches were evaluated on synthetic data. The first approach is Jing Xiao et al's [97] non-rigid factorization algorithm. The second approach applies the batch algorithm to estimate the 3D structures from each partition. The optimum structure is the mean of the estimated 3D structures which was the smallest mean square distance to the 3D estimated structures. The third approach is the batch algorithm with metric

optimization. Two experiments were carried out to examine the performance of the algorithms.

In the first experiment, 5 rigid object datasets with Gaussian white noise were generated. The strength level of noise is defined as $\frac{\|\mathbf{noise}\|}{\|\mathbf{W}\|}$. 5%, 10% and 20% strength level of noise were added to the datasets. Thus,15 experiments used to evaluate the performance of the algorithm. Each dataset had 50 3D feature points and 100 frames with random projection matrices. A 200×50 measurement matrix $\mathbf{W}$ represented the image feature tracks.

In the second experiment, 5 non-rigid object datasets formed by 3 shape bases were generated. Each dataset had 25 3D feature points and 203 random projection matrices. A 406×25 measurement matrix $\mathbf{W}$ represented the image feature tracks. For the non-rigid dataset, Gaussian white noise was added at strength levels of 5%, 10% and 20% to evaluate the performance of the algorithms.

To make the experiments comparable, all the synthetic datasets were partitioned into 10 subsets and the batch algorithm of Section 3 was applied. For the rigid case, each subset contained 50 3D feature points and 10 random projection matrices. For the non-rigid case, each subset contained 25 3D feature points and 13 random projection matrices (3 basis images + 10 non-basis images). They formed 10 smaller measurement matrices $\mathbf{W}_i$. Then we applied non-rigid factorization algorithm on each $\mathbf{W}_i$ to recover 3D structures. Metric optimization was applied on these 3D estimated structures by quasi-Newton optimization algorithm.

For the rigid case, the relative measurement error, $\frac{1}{P}\sum_{p=1}^{P}\frac{\|\mathbf{b}_p-\mathbf{b}_p^{truth}\|}{\|\mathbf{b}_p^{truth}\|}$ was evaluated for examining the performance of our approach. The results are shown in Figure 7.1. For the non-rigid case, the mean of the relative er-

rors between the optimal structure and the ground truth, $\frac{1}{PF}\sum_{p=1}^{P}\sum_{f=1}^{F}\frac{\|\mathbf{s}_p-\mathbf{s}_p^{truth}\|}{\|\mathbf{s}_p^{truth}\|}$, was used instead. The results are shown in Figure 7.2. From the figure, the relative error of the proposed algorithm is significantly lower than factorization algorithm [97]. The variance of the error is also small, showing that the method is more stable and robust than the original factorization algorithm.

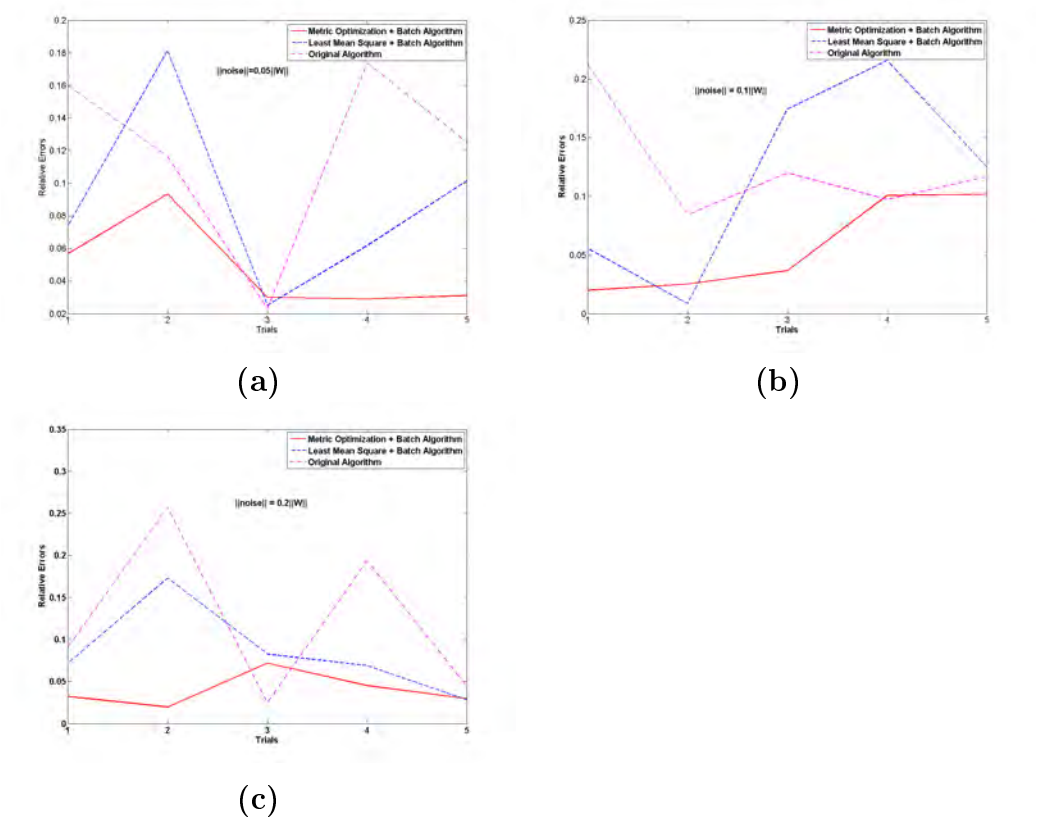


Figure 7.1: Relative errors of the three different appoaches of the factorization algorithms on rigid synthetic data under different levels of Gaussian white noise. (a, b and c).

7.1.2 Qualitative Evaluation on Facial Expressions

Recognizing facial expressions is one of the current challenging problems. Thus, we are motivated to evaluate our approach with facial expressions. In

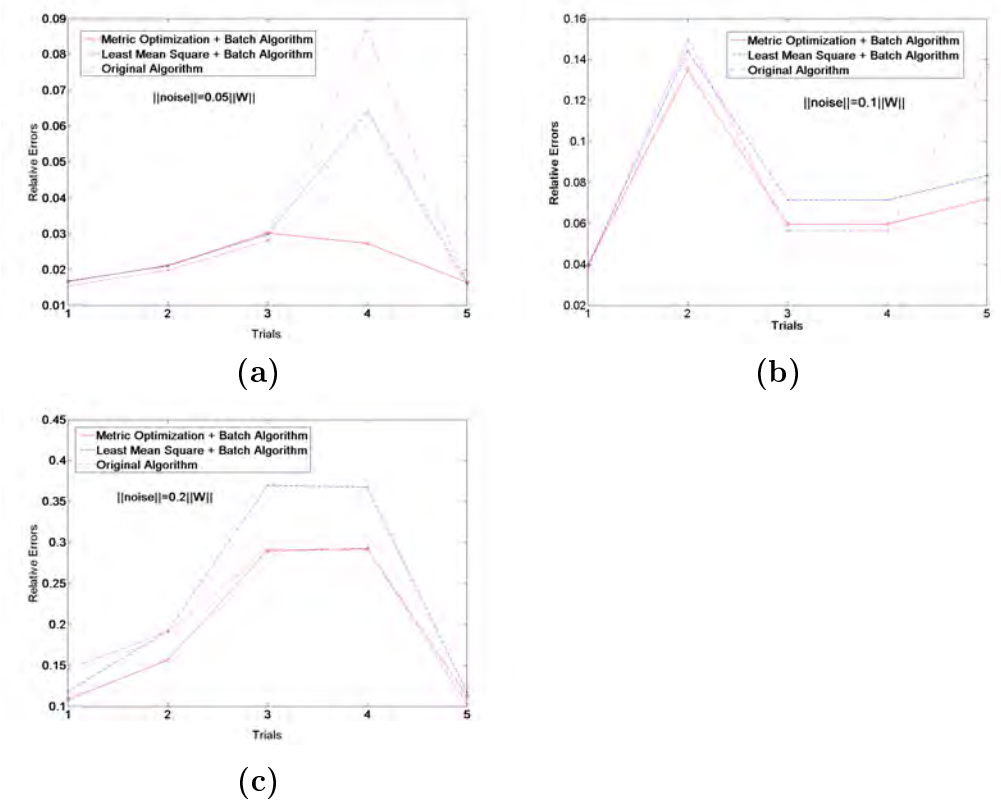
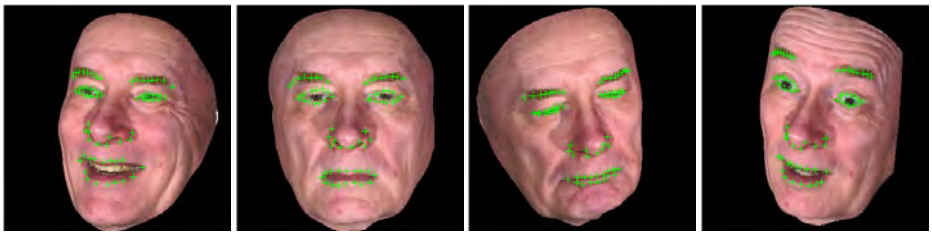


Figure 7.2: Relative errors of the three different appoaches of the factorization algorithms on non-rigid data under different levels of Gaussian white noise (a, b and c).

this experiment, a 3D face model with four different expressions captured from the 3D Facial Expression Database [100] at the State University of New York was used to examine the qualitative performance of our proposed approach. The four expressions are happy, neutral, sad and surprise. First, we manually selected 68 feature points on the 3D models. Then, the 3D models were rotated about x-axis from $-10°$ to $+10°$ in $2°$ steps, about y-axis from $-20°$ to $+20°$ in $1°$ steps and about z-axis from $-10°$ to $+10°$ in $2°$ steps. In each step, we generated an image of the 3D model. Therefore, we have 4961 images for each expression. Some images with different expressions are shown in Figure 7.3. The ground truth of the 3D feature

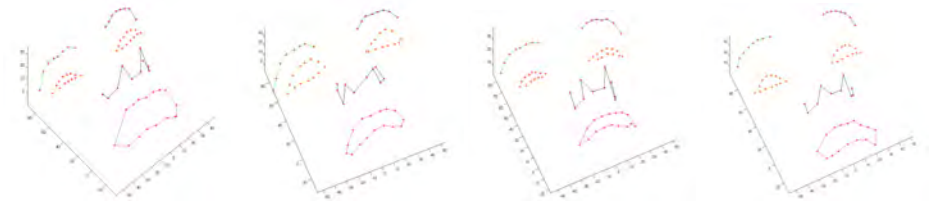
points of each expression is shown in Figure 7.4.

In this experiment, three different levels of Gaussian white noise were added to $\mathbf{W}$, with strength levels of 0%, 5% and 10%. Then, $\mathbf{W}$ was partitioned into 41 subsets for the batch algorithm and factorization was used for each subset with metric optimization. The results are shown in Figure 7.5, Figure 7.6 and Figure 7.7, respectively. The results show that the proposed algorithm successfully recovered the face expressions over the video sequence with low level of noise. In the future work, we propose to handle the structure recovery with texture information, and improve performance under larger noise.



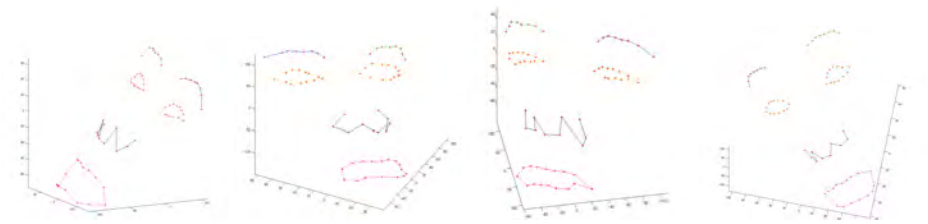
(a) **(b)** **(c)** **(d)**

Figure 7.3: (a) Happy expression image with rotation about $x = -10°$, $y = 20°$ and $z = 10°$ (b) Neutral expression image with rotation about $x = 0°$, $y = 0°$ and $z = 0°$ (c) Sad expression image with rotation about $x = -10°$, $y = -20°$ and $z = -10°$ (d) Surprise expression image with rotation about $x = -10°$, $y = 20°$ and $z = 10°$.



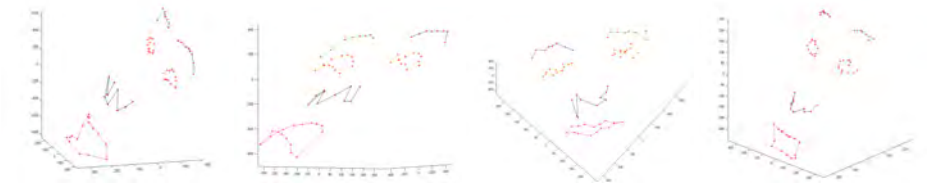
(a) **(b)** **(c)** **(d)**

Figure 7.4: (a) Ground truth of 3D happy expression (b) Ground truth of 3D neutral expression (c) Ground truth of 3D sad expression (d) Ground truth of 3D surprise expression.



(a) (b) (c) (d)

Figure 7.5: (a) Reconstructed 3D happy expression (b) Reconstructed 3D neutral expression (c) Reconstructed 3D sad expression (d) Reconstructed 3D surprise expression under 0% Gaussian white noise.



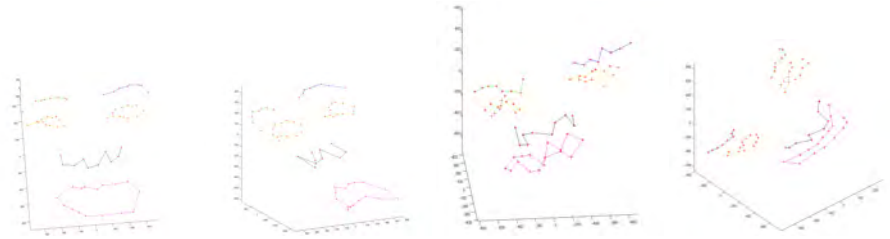
(a) (b) (c) (d)

Figure 7.6: (a) Reconstructed 3D happy expression (b) Reconstructed 3D neutral expression (c) Reconstructed 3D sad expression (d) Reconstructed 3D surprise expression under 5% Gaussian white noise.

7.2 Experiments on Image Inpainting

In our image inpaiting experiments, 156 aligned images obtained from the Yale Face database B [35] and FERET database [71] and 31 images from an image sequence with different expressions were used. Five experiments were performed to evaluate our proposed image inpainting approach.

In the first experiment, the results when a test subject was included or excluded from the training set were compared. In the former case, all 15 aligned faces were used for training and a test subject was selected from one of them. For the latter case, 155 of the 156 aligned faces were used for training and the remaining image was used for test. The experimental results are shown in Fig. 7.8. It can be seen that regardless of whether the



(a) (b) (c) (d)

Figure 7.7: (a) Reconstructed 3D happy expression (b) Reconstructed 3D neutral expression (c) Reconstructed 3D sad expression (d) Reconstructed 3D surprise expression under 10% Gaussian white noise.

subject is in the training database or not, the inpainting is realistic, though when the subject is in the training set, the inpainted region is closer to the original.

In the second experiment, our approach in Algorithm 1 is compared with image inpainting obtained based on PCA only and only on Poisson image inpainting. Here, 155 of the 156 aligned images are used for training and the remaining image is used for test. Since the GVF for Poisson image inpainting is unknown, we simply assumed the GVF as zeros. The results are shown in Fig. 7.9. For PCA-based image inpainting, the boundary of the filled missing region can be clearly seen although the structure of the missing region is well synthesized from the face eigenspace. The results are shown in Fig. 7.10. With the Poisson image inpainting, the inpainted boundary is seamless compared to the PCA-based image inpainting but it fails to recover the structure of the missing region because the GVF is not known. The results produced by our approach not only retain the structure of the face, but also have a smoother and visually pleasing seamless boundary.

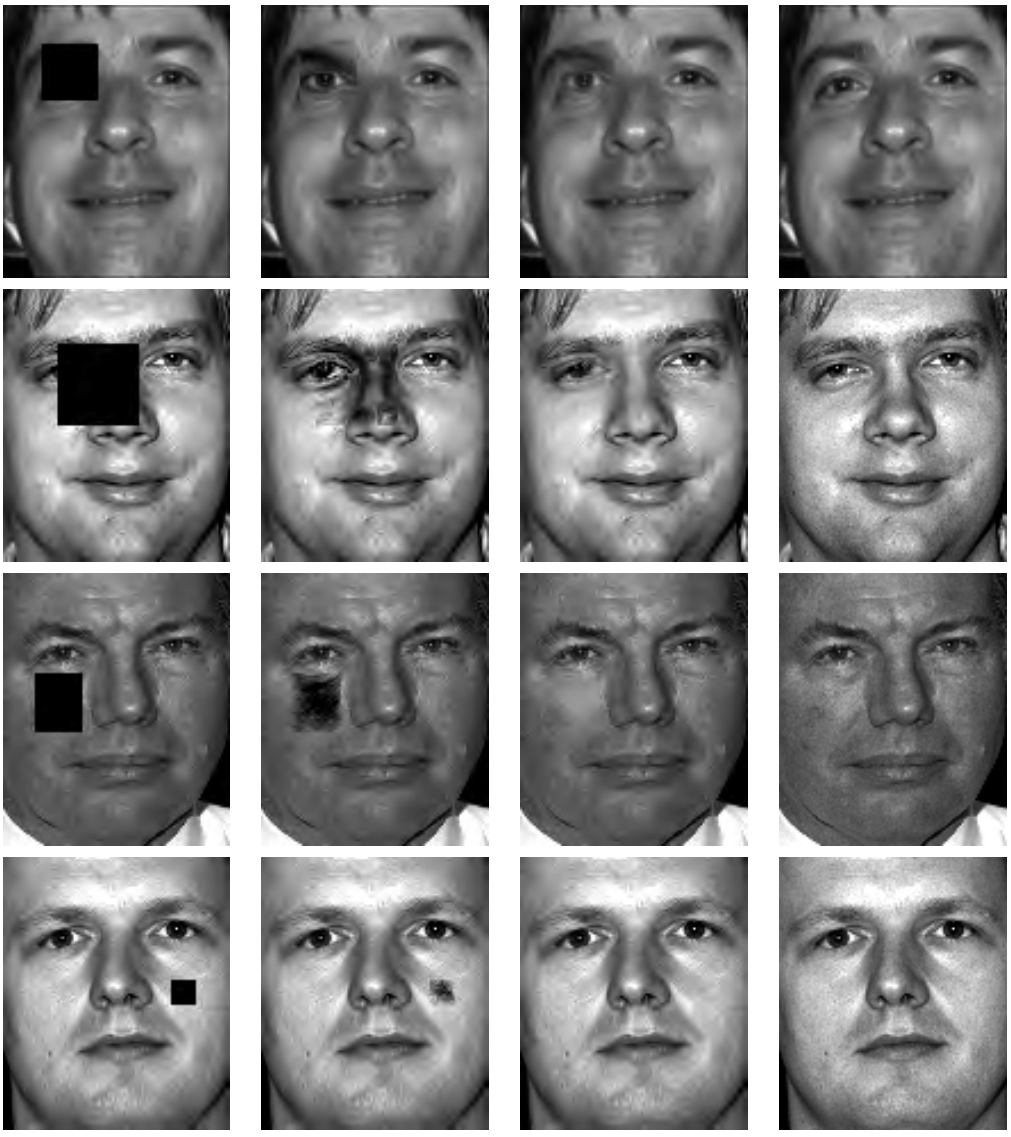
In the third experiment, we quantitatively evaluate our approach in Algorithm 1. Our approach is compared with Poisson image inpainting

(a) (b) (c) (d)

Figure 7.8: Image inpainting when a subject is in the training database or not: (a) the corrupted images, (b) inpainted images when the subject is not in the training database, (c) inpainted images when the subject is in the training database and (d) original images.

(a) (b) (c) (d)

Figure 7.9: Comparison with other image inpainting approaches: (a) the corrupted images, (b) Poisson image inpainting, (c) inpainted images by iterative learning GVF (Algorithm 1) and (d) the original images.



(a) (b) (c) (d)

Figure 7.10: Comparison with other image inpainting approaches: (a) the corrupted images, (b) inpainted images by PCA-based approach, (c) inpainted images by iterative learning GVF (Algorithm 1) and (d) the original images.

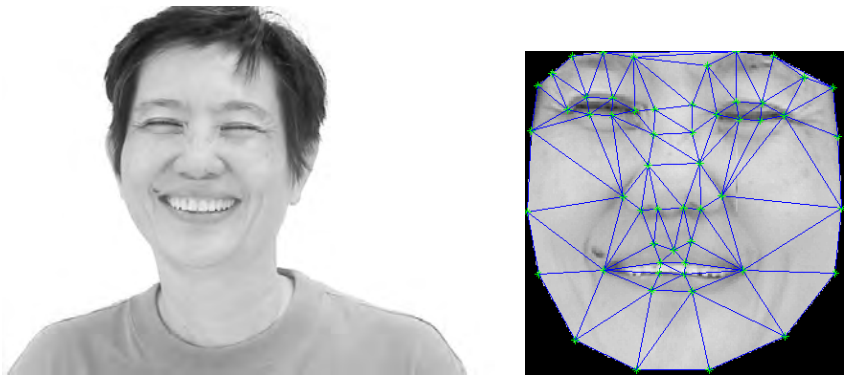
Corrupted Region	size	Ours	PCA	Poisson [69]
Smooth	21×11	2.9385	41.7922	4.0238
Smooth	36×31	2.4181	58.1195	8.3347
Non-smooth and structural	46×46	6.0953	35.1778	9.9527

Table 7.1: Mean squared error results on our approach, PCA-based inpainting and Poisson inpainting [69]

and PCA-based inpainting approach. Three sets of 10 corrupted images are used in this experiment. The corrupted regions of each set of the images are different. The size and corrupted region are indicated in Table 7.1. The corresponding mean squared errors are also shown in Table 7.1. From the experiments, the results showed that our approach outperformed the other two approaches.

In the fourth experiment, 31 images from a video sequence with different expressions were used to evaluate our inpainting method with patches selected for the PCA model as discussed in Section 4.4. 55 markers were labeled on the 31 images, and the mean of the 55 markers over the training images was used to form a face mesh by Delaunay triangulation. All 31 face images were warped onto the triangulated mean face mesh with 55 markers, and this was used to form the training set. Similar to the second experiment, comparison between a test image included or excluded from the training set was done. When the test image was included in the training database, all 31 warped image were available for training, and in the other case, the test was excluded from the training set. One of the face images and its corresponding warped face image are shown in Fig. 7.11.

One of the triangular patches was removed from a test image to simulate damage. The adjacent patches of the damaged patch, $\mathcal{B}_i$ and the corresponding patches in the training images, $\mathcal{B}_i^j$ were extracted and the patch



(a) (b)

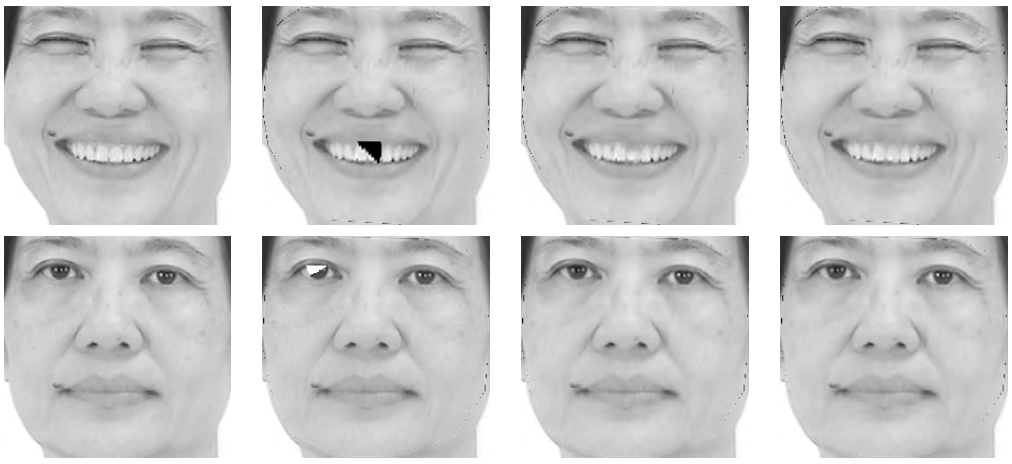
Figure 7.11: Samples of training images: (a) the original image obtained from an image sequence, (b) the corresponding warped image.

selection method of Section IV was applied to choose the best patches for learning. The selected training data obtained from Algorithm 2 was used in our inpainting approach to recover the damaged region. The results are shown in Fig. 7.12. It can be seen that with our approach, the quality of inpainted images is realistic even if the subject is not in the training database. The experiment also showed that the computational time is significantly reduced because fewer images were used for training the PCA model. Besides that, better representation of missing region improves the inpainting result, even though facial expression differ

7.3 Experiments on Face Hallucination

7.3.1 Generic Faces

The same set of 15 aligned images from the Yale Face Database B [35] which was used in image inpainting experiments was used to evaluate our image superresolution for generic face again. The top row of Figure 7.13 shows five 120×100 high resolution images, $\mathbf{I}^i_{\mathbf{HR}}$, from $\mathcal{H}$. The middle



(a) (b) (c) (d)

Figure 7.12: Image inpainting by patch-based learning model: (a) the original images, (b) corrupted images, (c) inpainted images when the test image is not in the training set and (d) inpainted images when the test image is in the training set.

row shows corresponding 60×50 images, $\mathbf{I}^i_{\mathbf{LR}}$, from $\mathcal{L} = \mathfrak{F}\mathcal{H}$, obtained by 2×2 averaging, and replacing those pixels by the average value. The bottom row shows bilinearly interpolated images, $\tilde{\mathbf{I}}^i_{\mathbf{LR}} = \mathfrak{G}(\mathbf{I}^i_{\mathbf{LR}})$. Figure 7.14 shows the first four principal components from the $\mathcal{H}$ and $\tilde{\mathcal{L}}$ spaces. The last row of Figure 7.14 shows the differences between the corresponding principal components, which form the elements $\mathbf{e}_i$ of the high frequency Laplacian edge space, $\mathcal{E}$. The 4×4 matrix in Figure 7.14 shows the values of $< \mathbf{h}_i, \tilde{\mathbf{l}}_j >$, $i, j = 1, \ldots, 4$. It is apparent that $< \mathbf{h}_i, \tilde{\mathbf{l}}_j > \approx 1$, if $i = j$ and ≈ 0, otherwise.

We performed several experiments to evaluate the proposed SR technique. In the first experiment, we compared the differences between bilinear interpolation and our SR method. For the latter, we used 14 of the 15 images in the database to form the training set and used the remaining image for test. We considered super-resolving to 120×100 from 60×50 ($\times 2$) as well as from 30×25 ($\times 4$). In the first case, each HR image was 2×2 pixel

Figure 7.13: Five high resolution training images obtained from Yale Face Database B are shown in the first row and corresponding simulated low resolution images are shown in the second row. The third row images are the bilinearly interpolated images.

Figure 7.14: The first four principal components of $\mathcal{H}$ (1^{st} row), $\tilde{\mathcal{L}}$(2^{nd} row) and $\mathcal{E}$ (3^{rd} row) and the matrix of $< \mathbf{h}_i, \tilde{\mathbf{l}}_j >$, i, $j = 1, \ldots, 4$.

averaged and downsampled to 60×50 to form a low resolution subspace (LR_1). For $\times 4$ magnification, the 60×50 training images were further 2×2 averaged and downsampled to 30×25 to produce a space at this resolution (LR_2). Given a 60×50 test image, the HR and LR_1, subspaces were used in our algorithm to produce the super-resolved image. Given a 30×25 test image, we implemented our algorithm in two steps: LR_2 and LR_1 were first used to super-resolve to 60×50, after which LR_1 and HR were used to produce the final 120×100 super-resolved image. The results are shown in Figure 7.15. The super-resolved images show considerably more details than the images interpolated by bilinear interpolation. Our approach not only retains individual specific information from the low resolution images, but also enhances the details at the eyes, nose and mouth. The performance differences are especially dramatic at $\times 4$ magnification.

The above experiments can be considered to be somewhat idealized in that the exact blur in the test image is assumed to be known, and is used to generate low resolution space, $\mathcal{L}$, used in the SR process. Hence it is of interest to consider cases when the blur in the test image is different from that used to generate the space $\mathcal{L}$, as in practice, the blur in the test image may be unknown. In Figure 7.17 (b) and (c), our approach is applied on a low resolution image which contains blur from a 5×5 averaging mask, and horizontal motion blur, respectively. The SR images were produced using the same space $\mathcal{L}$ derived as in the previous experiment. Again it is apparent that the SR images produced are richer in detail compared to the interpolated images. We also compare our SR method with neighborhood embedding method [21], fast image upsampling method[77] and the sparse representation method [99]. In practice, the better our guess is of the blur

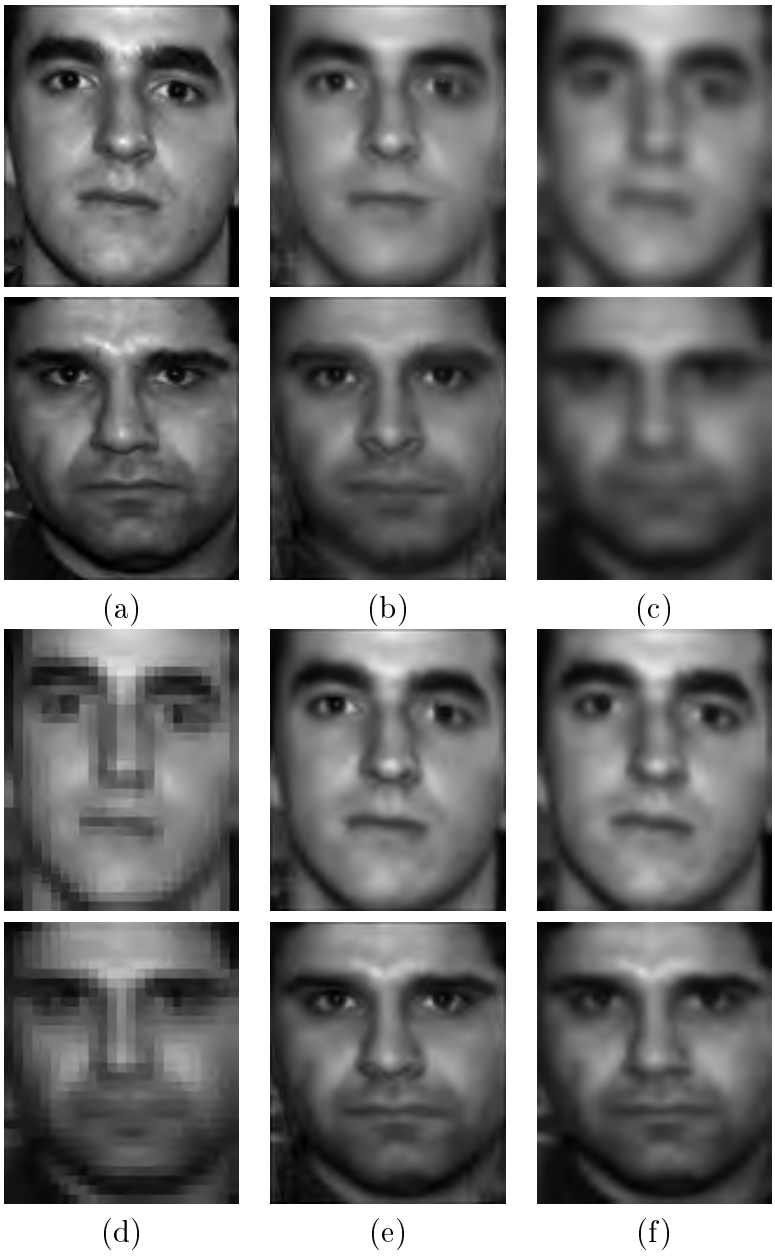
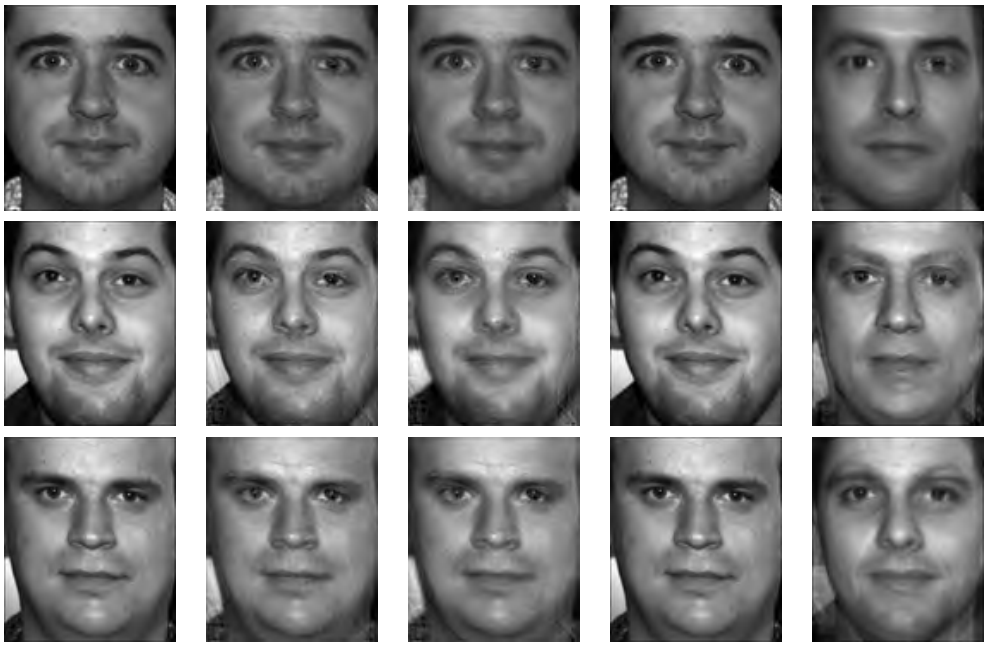


Figure 7.15: Superresolution on images with different levels of magnification: (a) The original high resolution images, (b) SR images for $\times 4$ magnification, (c) bilinearly interpolated images for $\times 4$ magnification, (d) $\times 4$ magnification by pixel replication (e) SR images for $\times 2$ magnification and (f) Bilinearly interpolated images for $\times 2$ magnification.



(a) (b) (c) (d) (e)

Figure 7.16: SR experiments when subject is in the training database or not, with different methods. (a) The original high resolution images, (b) SR images by our approach (test image is included in the training dataset), (c) SR images by our approach (test image is not included in the training dataset), (d) SR images by conventional PCA reconstruction method when the test image is in the training dataset and (e) not included in the training dataset.

Filters	Bilinear	NN[21]	UP[77]	Sparse [99]	Ours
Average	63.67	67.91	76.54	68.46	77.37
Motion Blur	63.92	70.80	73.44	72.24	75.98
Noise	87.78	94.34	127.59	107.15	97.07
Combo	53.14	55.42	60.84	59.06	66.95

Table 7.2: Sharpness measurement results on bilinear interpolation, neighborhood embedding method [21], fast image upsampling method [77], sparse representation method [99] and the proposed superresolution method

in the test image to use for generating the space $\mathcal{L}$, the better we expect the SR image quality will be. We also applied SR to a low resolution image with 20db additive noise. The result is shown in Figure 7.17 (d). Finally, we also tested our approach on a low resolution image produced by two blurring filters, motion blur and average filter, and with 20db Gaussian white noise. The result is shown in Figure 7.17 (e). We also quantitatively evaluate the experiment results by measuring the sharpness of the output SR images. The sharpness is defined as $\frac{1}{N}\sum G(x,y)^2$ where $G(x,y)$ is the image gradient at coordinate (x,y) and N is the number of pixels. Five low resolution images are used in this experiment. The result is shown in Table 7.2. It showed that our approach is able to reconstruct the edge information and textural details on the heavily distorted low resolution images by different filters. Ours is relatively more consistent compared to other approaches.

In the last experiment, a portion of low resolution image was removed to simulate damage. Thus, the missing region in the low resolution needs to be recovered before the image is super-resolved. We have developed a related technique for inpainting, and using this with the SR method considered here, our approach can further enhance the missing region. The results are shown in Figure 7.18.

(a) (b) (c) (d)

Figure 7.17: Superresolution (1^{st} row), bilinear (2^{nd} row), neighborhood embedding method [21](3^{rd} row), fast image upsampling method [77] (4^{th} row), and sparse representation method [99] (5^{th} row) on images with different blurs and noise: (a) Blurry image from a 5×5 averaging mask, (b) Blurry image from a motion blur filter, (c) 20db noisy image (d) Image with combined effects of (a), (b) and (c).

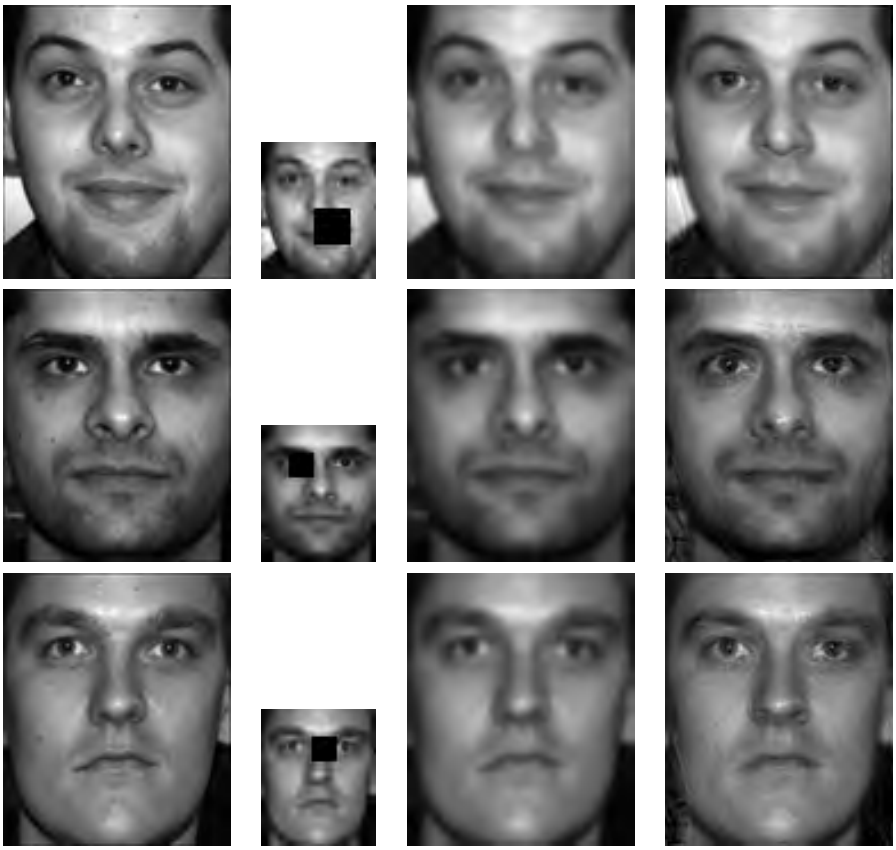


Figure 7.18: Image inpainting and superresolution: (a) The original high resolution image, (b) The damaged low resolution image, (c) The interpolated inpainted image and (d) The super-resolved inpainted image.

Figure 7.19: High resolution training images with different expressions and poses extracted from each dataset

7.3.2 Specific Faces

Experiments for our face hallucination on specific person used five sets of high resolution training images. Each dataset consists of a specific person with different poses and expressions. Each dataset has 2000-5000 high resolution images. All these high resolution images have been labeled 38 feature points for alignment algorithm. Some sample images are shown in Figure 7.19. The input images were captured from different sources such as mobile phone camera, web camera, low-end video camera etc. Due to the quality image, the low resolution input faces are manually labeled the 38 feature points. The superresolution images were enlarged the input images two times to eight times. In the first experiment, our proposed image selection method defined in section 6.1.4 was evaluated. Three measurement for texture analysis were compared in the experiment. The first approach applied the $L2$-norm to measure the texture similarity, $P(D_T^I|x)$. The second

	$L2$-norm	χ^2	χ^2+Shape Ratio
Open Eye	88.33%	70.25%	**90.75%**
Closed Eye	52.25%	61.67%	**67.67%**
Open Mouth	10.00%	77.67%	**84.67%**
Closed Mouth	28.33%	81.50%	**91.50%**

Table 7.3: Comparison results on the proposed similarity measurement, Equation 6.7 with the conventional approaches, χ^2 distance and $L2$ norm.

approach replace $L2$-norm in the first approach to the χ^2 measurement. In the third approach, the shape ratio is integrated with texture analysis as mentioned in section 6.1.4. Three training datasets were used in this experiment. 40 low resolution images with different poses and expressions were selected as query images for the performance evaluation. The top N rank of high resolution training images were retrieved from the databases. The retrieval rate (RR) is defined to evaluate the system. RR is defined as

$$RR(q) = \frac{H(q)}{N} \tag{7.1}$$

where $H(q)$ is the number of ground truth images for a query image, q, found in the top N retrieved images. The results were shown in Table 7.3. In the experiment, we found that the performance of $L2$-norm measurement was inconsistent compared to χ^2 measurement. The combination of shape and texture analysis further boosted the accuracy of the retrieval system. The closed eye did not work well in all approaches. It is mainly due to the extracted region is too small that the PHOG descriptor and shape ratio did not work properly. However, 67.67% of the retrieval rate was acceptable for our image superresolution application. A query image and its retrieved HR images are shown in Figure 7.20

In the second experiment, we compared the difference between bicubic

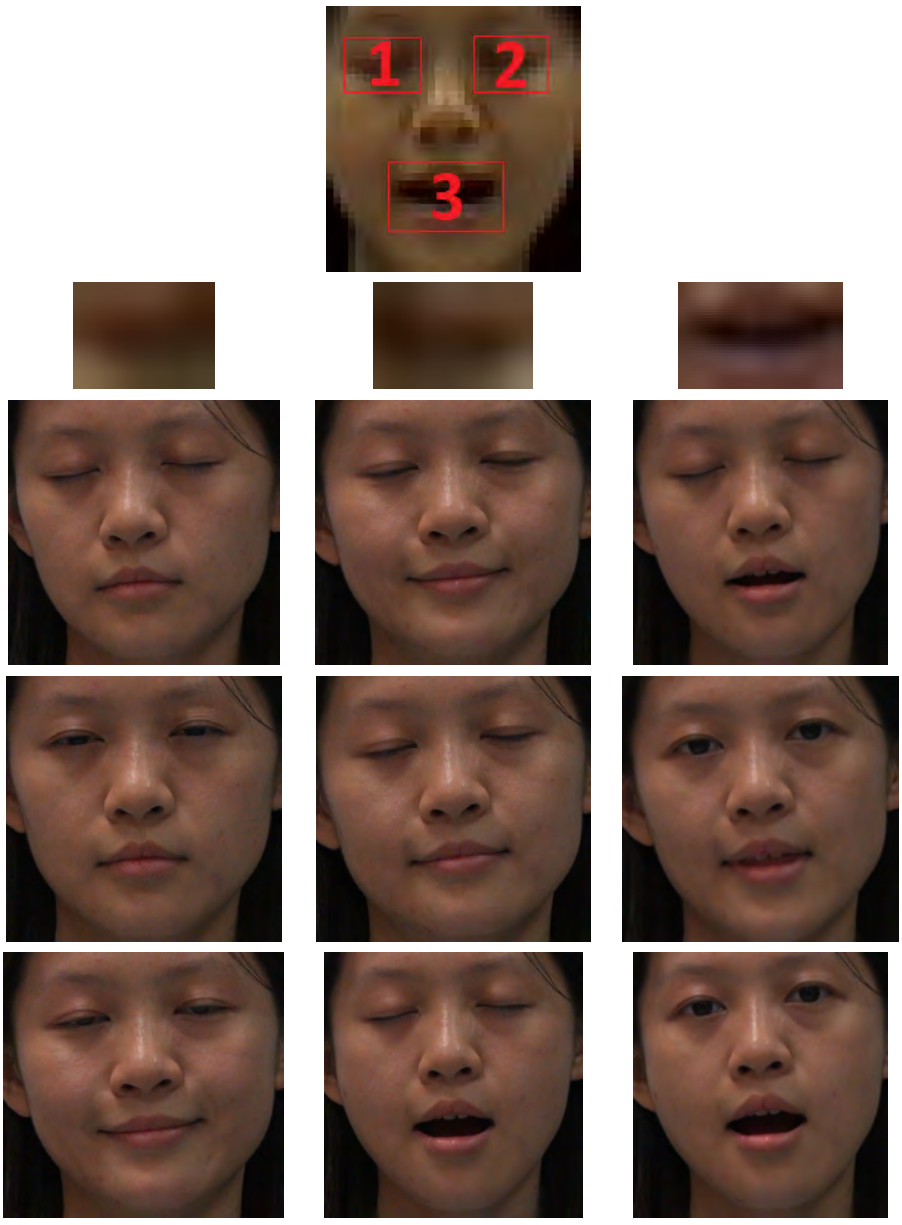


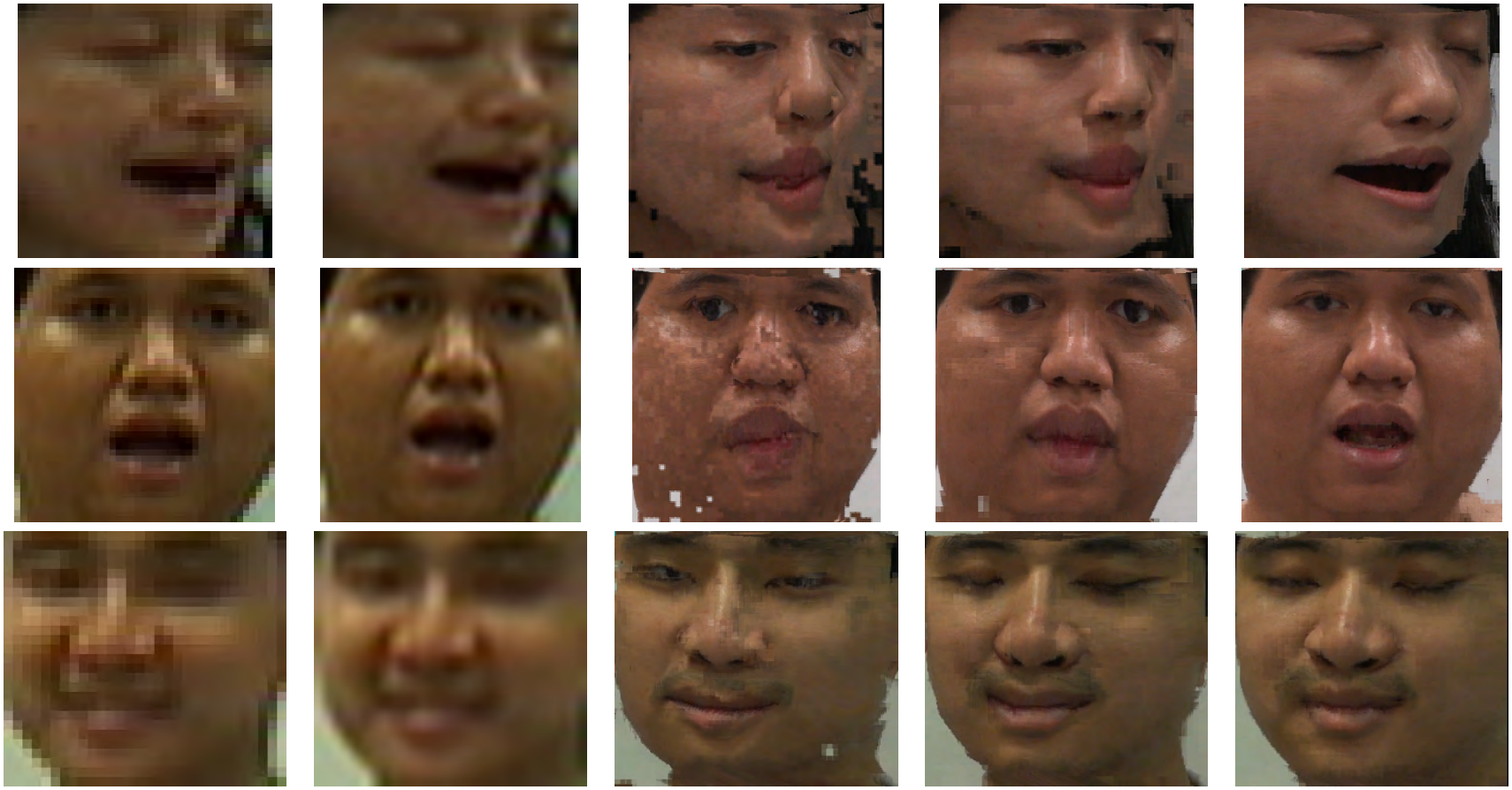
Figure 7.20: (1^{st} row)A low resolution query image (2^{nd} row)The region of interest (left to right: left eye, right eye, mouth) with bicubic interpolation for image selection and the corresponding high resolution images retrieved from the training dataset are shown in the next each column

interpolation and our proposed image superresolution on specific person. Low resolution images with different poses and expressions were used as the input images. Each input image was manually marked the 38 feature points due to the limitation of the feature detector on low resolution images. 20 high resolution images were retrieved from the corresponding dataset by the proposed retrieval system. The input image was aligned with the retrieved images by the feature points. The MRF based superresolution approach discussed in section 6.3 was applied. Some sample results with four time magnification are shown in Figure 7.21 (d). Figure 7.21 (c) showed the results by using bicubic interpolation with our proposed color correction approach. The results of our method not only preserve the facial structure, but also have richer textural details compared to the interpolated images.

In the third experiment, we evaluated the importance of our image selection method and alignment algorithm here. Three training datasets were used for the experiment. Low resolution images with different poses and expressions shown in Figure 7.22 (a) were used as the input images. Each input image was manually marked the 38 feature points due to the limitation of the feature detectors on low resolution images. The input image also applied the proposed color constancy method. Next, the input face was divided into three part, namely right eye region, left eye region and mouth region. The eyes and mouth region were used for texture analysis. 20 high resolution images were retrieved for each region of interest from the corresponding dataset by the proposed image selection algorithm. The retrieved images were then aligned with the low resolution input images by their corresponding 38 tracked feature points. The MRF based superresolution approach presented in section 6.3 was applied. The sample results

Figure 7.21: Image superresolution on a specific person with different expressions. (a)Input images (b) Bicubic interpolated results on input images (c)Bicubic interpolated results with color correction (d) Superresolution images for $\times 4$ magnification

with eight time magnification are shown in Figure 7.22 (e). Figure 7.22 (c) applied the proposed superresolution approach but the reference images are randomly selected and the alignment presented in section 6.2 is not applied. Figure 7.22 (d) applied the similar approach as Figure 7.22 (c) but the proposed alignment is applied. The intelligent image selection plays an important role in our approach. The proposed image selection method gave the appropriate reference images for our face hallucination approach to generate the high resolution image with the correct pose and expression. The superresolution images not only preserve the facial structure, but also have richer textural details compared to the bicubic interpolated images in in Figure 7.22(b).



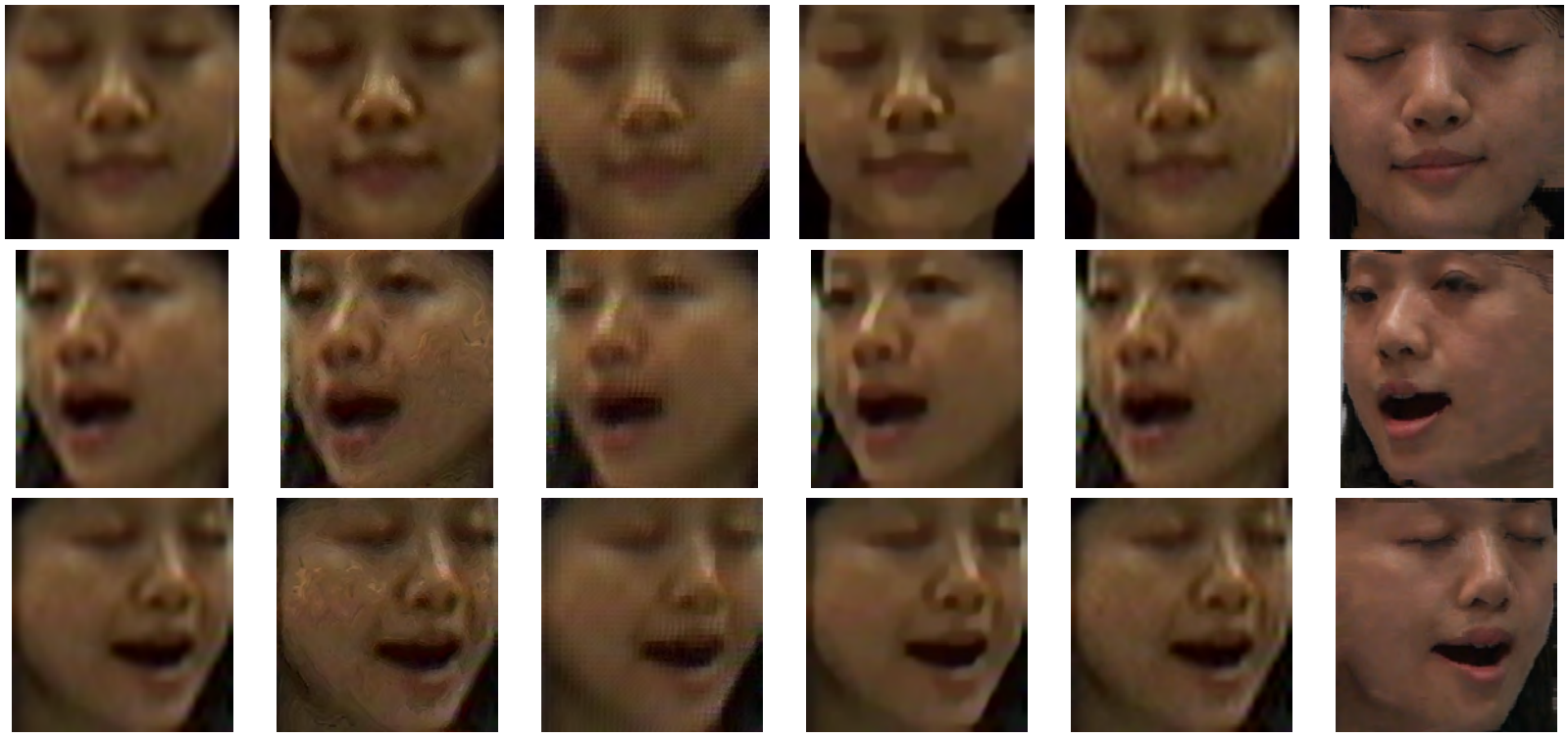
(a) (b) (c) (d) (e)

Figure 7.22: (a)Input image (b)Bicubic interpolation (c)Superresolution image by randomly selected high resolution patches (d) Superresolution image by randomly selected the patches from the aligned images (e)Superresolution image by selected patches from the aligned images with similar pose and expression for $\times 8$ magnification

In the fourth experiments, we compared the outputs of our method with the bicubic interpolation method, the VISTA approach [34], neighborhood embedding method [21], fast image upsampling method[77] and the sparse representation method [99]. Figure 7.23 shows the comparison results for $\times 4$ magnification. Figure 7.24 and Figure 7.25 show the comparison results for $\times 8$ magnification. The input images were captured by real-world low-end cameras which the blurring filter, down-sampling operator, other image distortions and noise are unknown. Those effects are also very difficult to predict from a single low resolution image with the limited information about the cameras. Using the implementations provided by Chang et. al. [21], we generated the superresolution images with 4x magnification shown in Figure 7.23 (b) and 8x magnification shown in Figure 7.24 (b) and Figure 7.25 (b) . Similarly, we also generated the corresponding superresolution images by work in Yang et. al. [99] with 4x magnification shown in Figure 7.23 (c) and 8x magnification in Figure 7.24 (c) and Figure 7.25 (c). Compare to our results shown in Figure 7.23 (d), 7.24 (d) and Figure 7.25 (d), our method significantly improved the quality and the resolution of the output images. The performance of our method is outstanding because other prior approaches enhanced interpolated images by adding high frequency details but our method replaces the poor quality input patches with the high quality patches from the selected images. Thus, ours not only increase the size of image resolution, but also improve the quality of the images.

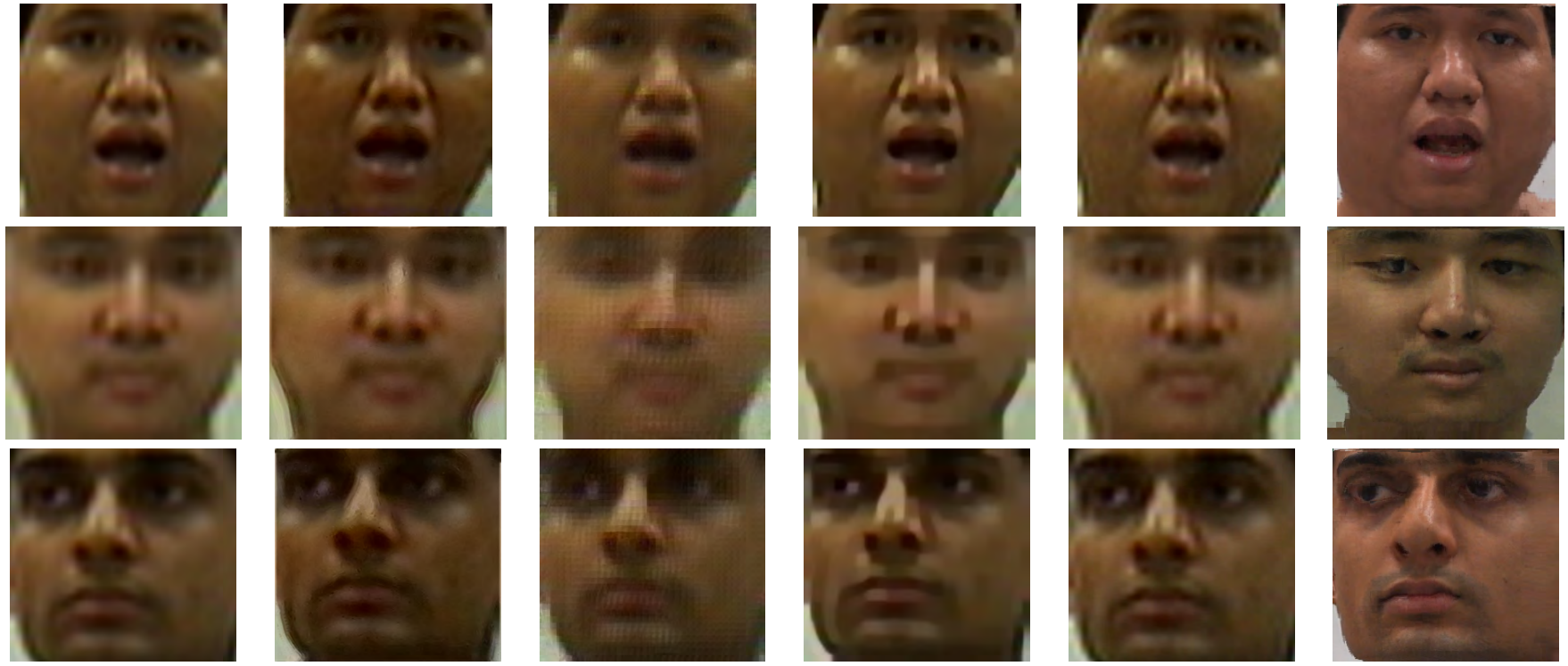
(a) (b) (c) (d) (e) (f)

Figure 7.23: (a) Bicubic interpolation (b)The VISTA approach [34] (c)Neighborhood embedding method [21] (d)Fast image upsampling method [77] (e) Sparse representation method [99] (f)Our method for $\times 4$ magnification



(a) (b) (c) (d) (e) (f)

Figure 7.24: (a) Bicubic interpolation (b)The VISTA approach [34] (c)Neighborhood embedding method [21] (d)Fast image upsampling method[77] (e) Sparse representation method [99] (f)Our method for $\times 8$ magnification



(a) (b) (c) (d) (e) (f)

Figure 7.25: (a) Bicubic interpolation (b)The VISTA approach [34] (c)Neighborhood embedding method [21] (d)Fast image upsampling method[77] (e) Sparse representation method [99] (f)Our method for $\times 8$ magnification

In the fifth experiment, an underexposure image with pepper noise was used for face hallucination. It is very commonly captured by an inexpensively pocket camera under the poor lighting environment. Since our method replaces the noisy image patches with the high quality patches from the training data, a more visually pleasing image was generated by our method compared to the other methods. The result is shown in Figure 7.26.

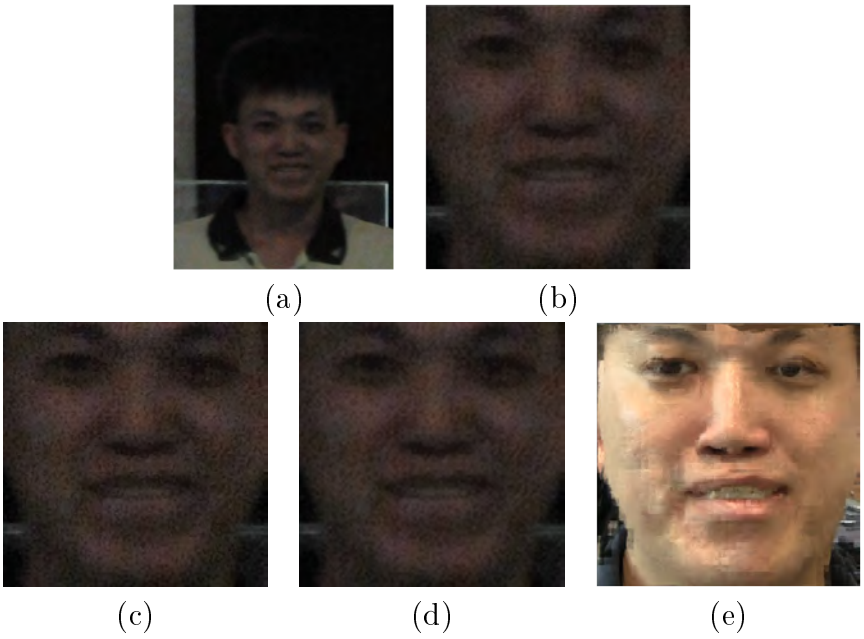


Figure 7.26: (a) Input image with pepper noise and underexposure (b) Bicubic interpolation (c)Neighborhood embedding method [21] (d) Sparse representation method [99] (e) Our method for $\times 2$ magnification

Chapter 8

Conclusion

In this thesis, we addressed the problems of facial structure from motion, facial image inpainting and face hallucination. In practice, the quality of personal photos depends a great deal on the faces in the photos being clearly visible and recognizable. Hence, it would be very useful to generate a high resolution facial image with high quality textural details from a low resolution facial image. However, generating a high resolution image from a low resolution image is an ill-posed problem. The missing information can only be recovered in a sensible manner by leveraging on prior knowledge or imposing prior constraints.

We proposed an extension of the closed-form non-rigid factorization algorithm proposed by Xiao et al. [97]. A batch algorithm partitions the measurement matrix to make several independent estimates of the 3D structure and then uses a metric optimization process to fuse the estimates and recover an optimized 3D structure. The batch algorithm allows the system to process the data in parallel because the factorization algorithm can be applied on each partition separately. Thus, it is suitable for real-time applications such as surveillance and biometric authentication systems. More-

over, it is able to recover non-rigid structure from an arbitrary number of images. The algorithm does not require repeated computation with the factorization algorithm using the whole measurement matrix every time new data is available. Hence, the computations are more effective by using the proposed approach. The metric optimization is another significant contribution of the current work. It fuses the multiple solutions obtained from different partitions to find an optimal 3D structure so that errors in the lengths and the mutual angles of the feature points is minimized. The experiments showed that the proposed approach is more accurate and robust than the existing factorization algorithm for both rigid and non-rigid objects under different levels of Gaussian white noise.

Next, we proposed a new POCS based facial image inpainting technique which iteratively learns the guidance vector field for solving the Poisson equation with Dirichlet boundary conditions. Experiments showed that the performance of the new approach is much better than solely applying PCA-based image inpainting or GVF image inpainting. The new approach not only retains the structure of the missing region, but also smoothens the boundary between the missing region and its neighbours. The robustness of the new approach has been demonstrated by experiments. A patch selection scheme for the learning model significantly reduced the complexity of modeling the missing region space and improved the effectiveness of the new approach.

We also contributed two face hallucination techniques for two different applications. The first proposed approach is a POCS based face hallucination algorithm with a learned edge model. It is applied on generic faces. The new approach not only enhances the missing high frequency details,

but it is also faithful to the structural details in the low resolution input image. The robustness and effectiveness of the approach are evident through the experiments in a wide variety of situations of practical interest. The idea is applicable to a given class of objects, so that a model can be learned to estimate the missing high frequency details. The model is used in conjunction with a data-based constraint in the POCS algorithm to produce the superresolution image.

The second proposed approach is a MRF based face hallucination algorithm for specific person. Here, we assumed that the subject is known. In addition, a set of high resolution images with different poses and expressions is available as training data. Nowadays, having a lot of personal photos is common to most of people. Moreover, high resolution images can be easily obtained from high definition videos. The challenge of the problem is that the facial pose and expression in the input image is not always same as the training images. Firstly, we proposed an image retrieval system based on texture and shape analysis to select the high resolution images with similar pose and expression from the training dataset. From the experiments, we showed that the retrieval rate is more than 84% in most of the cases. It is good enough for our face hallucination algorithm.

Next, we applied a color correction algorithm to overcome the color distortion problem due to the lighting conditions and low-end camera sensors. Then, the query image is aligned with the selected high resolution images by the affine warping. Lastly, we proposed a color and gradient based MRF model to enlarge and enhance the low resolution input image. From the experiments, we showed that our new approach outperformed the conventional interpolation techniques. In additional, our approach is

able to handle the under-exposure image with noise. According to our best knowledge, other image superresolution approaches do not work on this issue.

8.1 Future Work

We would like to extend our current work onto 3D models and video sequence. Using the factorization algorithm to recover 3D structure has certain limitations due to the limited availability of corresponding points over the low resolution image sequence. The factorization algorithm is only able to recover the depth information of the salient feature points because it is very difficult to find correspondences on smooth and textureless surface. Calibrating the camera for finding corresponding points is very difficult for non-rigid objects because the deformation of the objects and camera motion occur simultaneously. Thus, some interpolation techniques are required to improve the geometric quality of the recovered structure. The textural details of the 3D facial models can be done by our current image superresolution approach for image enhancement. For video sequence, we can impose an additional temporal coherent constraint into our MRF model to extend our work onto video superresolution.

Moreover, all our contributions in image inpainting and image superresolution can be integrated into a system for the facial image enhancement. The computational cost and robustness of our current image retrieval system can be further improved for a real-time application on eletronic devices eg. digital cameras, mobile phones, tablet PC etc.

Bibliography

[1] http://www.expo2005.or.jp/ml/en/08/index.html.

[2] Henrik Aanæs, Rune Fisker, Kalle Åström, and Jens Michael Carstensen. Robust factorization. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 24(9):1215–1225, 2002.

[3] C. Allène and N. Paragios. Image renaissance using discrete optimization. In *18th International Conference on Pattern Recognition*, pages 631–634, Hong-Kong, Aug 2006.

[4] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 24(9):1167–1183, 2002.

[5] Simon Baker and Takeo Kanade. Hallucinating faces. In *Fourth International Conference on Automatic Face and Gesture Recognition*, March 2000.

[6] Y. Bar-Shalom and T. Formann. *Tracking and Data Association*. Academic Press, 1988.

[7] William A. Barrett and Alan S. Cheney. Object-based image editing. In *SIGGRAPH '02: Proceedings of the 29th annual conference on*

Computer graphics and interactive techniques, pages 777–784, New York, NY, USA, 2002. ACM.

[8] B.G. Baumgart. *Geometric Modeling For Computer Vision.* PhD thesis, Stanford University, 1974.

[9] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110:346–359, June 2008.

[10] M. Bertalmio, A. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *In Proc. Int. Conf. Computer Vision and Pattern Recognition 2001.*, volume 1, 2001.

[11] Andrew Blake and Michael Isard. *Active Contours.* Springer, 1998.

[12] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2007.

[13] Johan G. Bosch, Steven C. Mitchell, Boudewijn P. F. Lelieveldt, Francisca Nijland, Otto Kamp, Milan Sonka, and Johan H. C. Reiber. Automatic segmentation of echocardiographic sequences by active appearance motion models. *IEEE Trans. Med. Imaging*, 21(11):1374–1383, 2002.

[14] Matthew Brand. Morphable 3d models from video. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2:456–463, 2001.

[15] Matthew Brand. A direct method for 3d factorization of nonrigid motion observed in 2d. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2:122–128, 2005.

[16] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2:690–696, 2000.

[17] A.M Buchanan and A.W. Fitzgibbon. Damped newton algorithms for matrix factorization with missing data. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2:316–322, 2005.

[18] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[19] D. Capel and A. Zisserman. Super-resolution from multiple views using learnt image models. *Proceedings of Computer Vision and Pattern Recognition 2001.*

[20] T.F Chan and J. Shen. Mathmatical models for local nontexture inpainting. 62:1019–1043, 2002.

[21] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 275–282, 2004.

[22] P. Chatterjee, S. Mukherjee, S. Chaudhuri, and G. Seetharaman. Application of Papoulis-Gerchberg method in image super-resolution and inpainting. *The Computer Journal*, 52(1):64–79, 2009.

[23] Kong-Man (German) Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette across time part i: Theory and algorithms. *International Journal of Computer Vision*, 62(3):221–247, 2005.

[24] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[25] A. Criminisi, P. Perez, and K. Toyama. Object removal by exemplar-based inpainting. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 02:721, 2003.

[26] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. In *Proceedings of the CVPR*, 2005.

[27] R. Deriche and G. Giraudon. A computational approach for corner and vertex detection. *International Journal of Computer Vision*, 1(1):167–187, 1993.

[28] D.W.Jacobs. Linear fitting with missing data for structure-from-motion. *CVIU*, 82(1):57–81, 2003.

[29] A. Efros and W.T. Freeman. Image quilting for texture synthesis and transfer. In *SIGGRAPH '01: ACM SIGGRAPH 2001 Papers*, pages 341–346. ACM, 2001.

[30] P. Ekman and W.V. Friesen. Facial action coding sytem: A technique for the measurement of facial movement. *Consulting Psychologists Press*, 1978.

[31] Sina Farsiu, Michael Elad, and Peyman Milanfar. Multi-frame demosaicing and super-resolution of color images. *IEEE Trans. on Image Processing*, 15:141–159, 2006.

[32] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T. Roweis, and William T. Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH 2006 Papers*, SIGGRAPH '06, pages 787–794, New York, NY, USA, 2006. ACM.

[33] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Trans. Graph.*, 28(3):1–10, 2010.

[34] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000.

[35] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

[36] R. Gerchberg. Super-resolution through error energy reduction. *Journal of Modern Optics*, 21(9):709–720, 1974.

[37] D. Geronimo, A.M. Lopez, A.D. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1239 – 1258, july 2010.

[38] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *ICCV*, 2009.

[39] Mei Han and Takeo Kanade. Reconstruction of a scene with multiple linearly moving objects. *International Journal of Computer Vision*, 59(3):285–300, 2004.

[40] C. Harris and M. Stephens. A combined corner and edge detector. *Fourth Alvey Vision Conference*, pages 147–151, 1988.

[41] Paul Harrison. A non-hierarchical procedure for re-synthesis of complex textures. In *WSCG 2001 Conference proceedings*, pages 190–197, 2001.

[42] A. Heyden and K. Astrom. Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 438–443, 1997.

[43] K. Jia and S.G. Gong. Generalized face super-resolution. 17(6):873–886, June 2008.

[44] Neel Joshi, Wojciech Matusik, and Edward H.Adelson. Personal photo enhancement using example images. *ACM Transactions on Graphics*, 29(2):1–15, 2010.

[45] S.P. Kim and N.K. Bose. Reconstruction of 2-d bandlimited discrete signals from nonuniform samples. 1990.

[46] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1568–1583, 2006.

[47] Seth C Koterba, Simon Baker, Iain Matthews, Changbo Hu, Jing Xiao, Jeffrey Cohn, and Takeo Kanade. Multi-view aam fitting and camera calibration. In *Proc. International Conference on Computer Vision*, volume 1, pages 511 – 518, October 2005.

[48] Kiriakos N. Kutulakos and Steven M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.

[49] A. Levin, A. Zomet, and Y. Weiss. Learning how to inpaint from global image statistics. In *Proceedings of International Conference on Computer Vision*, volume II, pages 305–313.

[50] Tony Lindeberg. Feature detection with automatic scale selection. *Int. J. Comput. Vision*, 30:79–116, November 1998.

[51] C. Liu, H.Y. Shum, and W. T. Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007.

[52] Ce Liu, W. T. Freeman, Richard Szeliski, and Sing Bing Kang. Noise estimation from a single image. In *CVPR (1)*, pages 901–908, 2006.

[53] Y. Liu, T. Belkina, J. H. Hays, and R. Lublinerman. Image defencing. In *Proceedings of CVPR 2008*, June 2008.

[54] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[55] M. A. Lourakis and R. Deriche. Camera self-calibration using the kruppa equations and the svd of the fundamental matrix: The case of varying intrinsic parameters. Technical Report RR-3911, INRIA, March 2000.

[56] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.

[57] B.D Lucas and Takeo Kanade. *In Proceedings of 7th International Conference on Artificial Intelligence*, pages 674–679, 1981.

[58] P. Marziliano and M. Vetterli. Reconstruction of irregularly sampled discrete-time bandlimited signals with unknown sampling locations. *IEEE Transactions on Signal Processing*, 48:3462–3471, 2000.

[59] Iain Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198 – 213, February 2002.

[60] Iain Matthews, Jing Xiao, and Simon Baker. On the dimensionality of deformable face models. *Carnegia Mellon University Technical Report, CMU-RI-TR-06-12*, 2006.

[61] Iain Matthews, Jing Xiao, and Simon Baker. On the dimensionality of deformable face models. *Carnegia Mellon University Technical Report, CMU-RI-TR-06-12*, 2006.

[62] Stephen J. Maybank and Olivier D. Faugeras. A theory of self-calibration of a moving camera. *International Journal of Computer Vision*, 8(2):123–151, 1992.

[63] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1615–1630, October 2005.

[64] Umar Mohammed, Simon J. D. Prince, and Jan Kautz. Visio-lization: generating novel facial images. In *ACM SIGGRAPH 2009 papers*,

number 57 in SIGGRAPH '09, pages 57:1–57:8, New York, NY, USA, 2009. ACM.

[65] N. Nguyen, P. Milanfar, and G. Golub. A computationally efficient superresolution image reconstruction algorithm. *IEEE Transactions on Image Processing*, 10(4):573–583, 2001.

[66] M.K. Ozkan, A.M Tekalp, and M.I. Sezan. Pocs-based restoration of space-varying blurred images. *IEEE Transactions Image Processing*, 3(4):450–454, 1994.

[67] A. Papoulis. A new algorithm in spectral analysis and band-limited extrapolation. *IEEE Transactions on Circuits and Systems*, 22(9):735–742, 1975.

[68] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: A technical overview. *IEEE Signal Processing Magazine*, 3:21–36, May 2003.

[69] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. In *SIGGRAPH '03: ACM SIGGRAPH 2003 Papers*, pages 313–318, 2003.

[70] G. Peyré, S. Bougleux, and L. Cohen. Non-local regularization of inverse problems. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 57–68, Berlin, Heidelberg, 2008. Springer-Verlag.

[71] P. J. Phillips, H. Moon, P. J. Rauss, , and S. Rizvi. The feret evaluation methodology for face recognition algorithms. *IEEE Trans-*

actions on Pattern Analysis and Machine Intelligence, 22(10):1090–1104, 2000.

[72] Conrad J. Poelman and Takeo Kanade. A paraperspective factorization method for shape and motion recovery. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 19(3):206–218, March 1997.

[73] Marc Pollefeys. *Self-Calibration and Metric 3D Reconstruction From Uncalibrated Image Sequences*. PhD thesis, Katholieke Universiteit Leuven, 1999.

[74] S. Roth and M.J. Black. Fields of experts: a framework for learning image priors. volume 2, page 860, June 2005.

[75] R.Y.Tsai and T.S.Huang. Multiframe image restoration and registration. *Advances in Computer Vision and Image Processing*, pages 317–339, 1984.

[76] B. Schmid and R. Mohr. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, 1997.

[77] Qi Shan, Zhaorong Li, Jiaya Jia, and Chi Keung Tang. Fast image/video upsampling. *ACM Transactions on Graphics (SIGGRAPH ASIA)*, 2008.

[78] Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research and Applications*, 30(1):21–30, 2005.

[79] Eli Shechtman, Yaron Caspi, and Michal Irani. Space-time super-resolution. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 27(4):531–545, 2005.

[80] Heung-Yeung Shum, Katsushi Ikeuchi, and Raj Reddy. Principal component analysis with missing data and its application to polyhedral object modeling. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 17(9):854–867, 1995.

[81] S.P.Kim and W.Y.Su. Recursive high-resolution reconstruction of blurred multiframe images. *IEEE Transactions On Image Processing*, 2:534–539, 1993.

[82] H. Stark and P. Oskoni. High resolution image recovery from image-plane arrays using convex projections. 6:1715–1726, 1989.

[83] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum. Image completion with structure propagation. In *SIGGRAPH '05: ACM SIGGRAPH 2005 Papers*, pages 861–868, New York, NY, USA, 2005. ACM.

[84] J. Sun, J.J. Zhu, and M.F. Tappen. Context-constrained hallucination for image super-resolution. pages 231–238, 2010.

[85] Richard Szeliski and Sing Bing Kang. Recovering 3d shape and motion from image streams using non-linear least squares. Technical Report Series CRL 93/3, Cambridge Research Lab, March 1993.

[86] Z. Tauber, Z. N. Li, and M. S. Drew. Review and preview: Disocclusion by inpainting for image-based rendering. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37(4):527–540, July 2007.

[87] S. Thurnhofer. *Quadratic Volterra Filters for Edge Enhancement and Their Applications in Image Processing.* PhD thesis, University of California, Santa Barbara, 1995.

[88] S. Thurnhoffer and S.K. Mitra. Edge-enhanced image zooming. *OPTICAL Engineering*, 35(7):1862–1870, 1996.

[89] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. *Carnegie Mellon University Technical Report CMU-CS-91-132*, 1991.

[90] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[91] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3d shape from 2d motion. *In Proc. NIPS 2003.*

[92] B. Triggs. Autocalibration and the absolute quadric. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 609–614, 1997.

[93] D.S. Turaga and T. Chen. Model-based error concealment for wireless video. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(6):483–495, 2002.

[94] H. Ur and D.Gross. Improved resolution from sub-pixel shifted pictures. *CVGIP: Graphical Models and Image Processing*, 54(2):181–186, 1992.

[95] V.Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. *SIGGRAPH 1999*, pages 187–194.

[96] Greg Welch and Gary Bishop. An introduction to the kalman filter. *SIGGRAPH 2001 Course 8*, 2001.

[97] Jing Xiao, JinXiang Chai, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision*, 67(2):233–246, 2006.

[98] Jingyu Yan and Marc Pollefeys. A factorization-based approach to articulated motion recovery. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2:815–821, 2005.

[99] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Trans. on Image Processing*, 19(11):2861–2873, November 2010.

[100] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew Rosato. A 3d facial expression database for facial behavior research. *In Proc. 7th International Conference on Automatic Face and Gesture Recognition*, pages p211–216, 2006.

[101] Z. Zhang, R. Deriche, O. Faugeras, and Q.T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78, 1995.

List of Publications

Conferences

1. **Loke Yuan Ren**, Kumar Pankaj, Ranganath Surendra, Huang Weimin, Object Matching Across Multiple Non-overlapping Fields of View Using Fuzzy Logic, *In Proceedings of the First International Workshop on Sensor Networks and Applications*, 2005, pp65-70. (**Oral**)

2. **Loke Yuan Ren**, Ranganath Surendra, Batch Algorithm With Additional Shape Constraints For Non-Rigid Factorization, *In Proceedings of the 18th British Machine Vision Conference*, 2007. (**Poster**)

3. **Loke Yuan Ren**, Ranganath Surendra, Image Inpainting with A Learned Guidance Vector Field, *In Proceedings of the 7th International Conference on Information, Communications and Signal Processing*, 2009. (**Oral**)

4. **Loke Yuan Ren**, Tan Ping, Ashraf A. Kassim, Image Superresolution on Personal Photo Albums, *In Proceedings of the ACCV Workshop on Face Analysis: The Intersection of Computer Vision and Human Perception*, 2012. (**Oral**)

Journal

1. **Loke Yuan Ren**, Kumar Pankaj, Ranganath Surendra, Huang Weimin, Object Matching Across Multiple Non-overlapping Fields of View Using Fuzzy Logic, *ACTA Automatica Sinica*,32(6):pp978-987, 2006.