Face Recognition at a Distance
Low-resolution and alignment problems

Yuxi Peng
FACE RECOGNITION AT A DISTANCE
LOW-RESOLUTION AND ALIGNMENT PROBLEMS

Yuxi Peng
Graduation committee:

prof.dr. J.N. Kok, University of Twente
prof.dr. R.N.J. Veldhuis, University of Twente
dr. L.J. Spreeuwers, University of Twente
prof.dr. A. Stein, University of Twente
prof.dr. B.P. Veldkamp, University of Twente
prof.dr. R. Raghavendra, Norwegian University of Science and Technology
prof.dr. F. Roli, University of Cagliari
dr. A.C.C. Ruifrok, Netherlands Forensic Institute
FACE RECOGNITION AT A DISTANCE
LOW-RESOLUTION AND ALIGNMENT PROBLEMS

DISSEMINATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra
on account of the decision of the graduation committee,
to be publicly defended
on Friday the 8th of February, 2019 at 16.45 hours

by

Yuxi Peng

born on 2 October, 1987
in Hunan, China
This dissertation has been approved by:

Promoter: prof.dr. R.N.J. Veldhuis
and
Co-promoter: dr. L.J. Spreeuwers
Acknowledgements

It took a long time to finish this work. I have worked with a lot of people who contributed their time and efforts to my research. It is a pleasure for me to express my gratitude to them all.

First and foremost, I would like to express my sincere gratitude to my supervisors Raymond Veldhuis and Luuk Spreeuwers for the continuous support of my Ph.D study, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis.

Further, I would like to thank the members of my graduation committee for reviewing my thesis.

Next, I would like to thank my colleagues from both SAS group and SCS group for their friendship and support. My thanks first goes to our technician Geert-Jan and our secretaries Bertine, Suse and Sandra for providing a lot of help and inspirations which help me integrate to the group and also to Dutch culture. I would like to thank Chris Zeinstra, Meiru, Tauseef, Chris van Dam, Dan, Abhishek, Jen-Hsuan, Chanjuan, Xiaoying, Roeland and Erwin for sharing research and life experience and (fun) ideas.

In the second year of my research I have collected a small face dataset which plays an important role in my work. I want to thank everyone who participated and gave me the consent to use their face images for my research.

I would like to thank my family and my friends for supporting me spiritually throughout writing this thesis and my life in general. I specially thank my parents for their encouragement. Although they are far away in China, they are always supportive to my work.

Last but not the least, special thanks to my beloved husband Atze for his support and care, and especially for the (in)genius design of my thesis cover.
Acknowledgements
3.2.2 Introduction .................................................. 20
3.2.3 Super-resolution methods ..................................... 22
3.2.4 Experimental results ......................................... 23
3.2.5 Conclusion ..................................................... 29

3.3 Comparison of super-resolution benefits for downsampled images and real low-resolution data ......................... 30
3.3.1 Abstract ....................................................... 30
3.3.2 Introduction .................................................... 31
3.3.3 Downsampled vs. real low-resolution image ................. 32
3.3.4 Super-resolution methods ..................................... 32
3.3.5 Experimental results ......................................... 34
3.3.6 Conclusion ..................................................... 39

3.4 Discussion ...................................................... 39

4 Low-resolution face recognition and the MixRes classifier 41
4.1 Introduction ...................................................... 41
4.2 Likelihood ratio based mixed resolution facial comparison ... 42
4.2.1 Abstract ....................................................... 42
4.2.2 Introduction .................................................... 42
4.2.3 Mixed-resolution Likelihood ratio based similarity score 44
4.2.4 Experiments .................................................... 48
4.2.5 Conclusion ..................................................... 51

4.3 Low-resolution face alignment and recognition using mixed-resolution classifiers ........................................... 51
4.3.1 Abstract ....................................................... 51
4.3.2 Introduction .................................................... 52
4.3.3 Approaches to LoRes-HiRes comparison ..................... 54
4.3.4 Evaluation using down-sampled or real LoRes images ... 58
4.3.5 Proposed methods ............................................. 62
4.3.6 Experiments .................................................... 67
4.3.7 Conclusion ..................................................... 77
4.3.8 Acknowledgements ............................................. 78

4.4 Discussion ...................................................... 78

5 The importance of proper alignment in low-resolution face recognition 81
5.1 Introduction ...................................................... 81
5.2 Low-resolution face recognition and proper alignment ....... 82
5.2.1 Abstract ....................................................... 82
5.2.2 Introduction .................................................... 82
5.2.3 The difference between down-sampled and real low-resolution images ........................................ 86
5.2.4 Matching-score based registration ......................... 98
5.2.5 Conclusion ................................................. 103
5.3 Discussion .................................................. 104

6 Low-resolution face recognition for a range of resolutions 105

6.1 Introduction .................................................. 105
6.2 Designing a low-resolution face recognition system for long-range surveillance ................................. 106
6.2.1 Abstract .................................................. 106
6.2.2 Introduction .............................................. 106
6.2.3 Scenario and hypothesis .................................. 108
6.2.4 MixRes classifier ........................................ 109
6.2.5 Experiments .............................................. 109
6.2.6 Conclusion ............................................... 114
6.3 Discussion .................................................. 115

7 Conclusion .................................................... 117

7.1 Answers to the research questions .......................... 118
7.2 Future work ................................................. 121

References ...................................................... 123

List of Publications ............................................. 131
Summary

Existing face recognition techniques are very successful in recognizing high-resolution facial images. However, their performance is not sufficient on low-resolution facial images. A very common source of low-resolution facial images is the recordings of surveillance cameras. Low-resolution facial images are difficult to recognize not only because limited information is contained in the small number of pixels, but also because the recordings of low-resolution facial images are usually under uncontrolled situations without user cooperation, which results in motion blur, variation of pose and illumination, and occlusion.

In a typical forensic case, such as a robbery in a pizza shop, surveillance quality facial images are recorded. To help the police with the forensic case, an automatic face recognition system can be used, for example, to search the police database to see if any facial image in the database matches the person on the surveillance images. In this case, the gallery images are the mug-shots in the police database. They are pre-stored high-resolution images. The probe images are the surveillance quality facial images, which are of low-resolution. In this thesis we focus on dealing with this scenario.

One problem is that the gallery and probe images are of very different resolutions, but most face recognition systems require them to be of the same resolution. Existing methods solve the resolution mismatch problem mainly by three approaches: 1. applying super-resolution to low-resolution probes, 2. down-sampling high-resolution gallery images and 3. direct low-resolution to high-resolution comparison.

The first approach is to apply super-resolution to low-resolution probes, and then perform comparison with the gallery images in the high-resolution space. However, our experimental results show that although the super-resolution methods enhance the visual quality of facial images, they have no benefits for face recognition on real low-resolution facial images that were captured at far distances.

The second approach is to down-sample high-resolution gallery images and perform comparison in the low-resolution space. It is simple and requires lower computational costs than using high-resolution images. Our experimental results show that this approach can outperform applying super-resolution methods.
The third approach is direct comparison of low-resolution probes and high-resolution galleries. Compared to the other two approaches, this approach has advantages that it avoids losing information as a result of down-sampling high-resolution or adding artefacts by super-resolution. We proposed a novel method, mixed-resolution biometric comparison, following this approach. The method is based on the likelihood ratio framework where in the derivation of the expression for the likelihood ratio, the combined statistics of the low and high resolution images is taken into account. Our experiments on surveillance quality images demonstrate that this method significantly outperforms the state-of-the-art.

In literature on low-resolution face recognition, what in some papers is considered as low-resolution, is still considered as high-resolution in other papers. To harmonize the terminology in low-resolution face recognition, we propose a resolution scale. We define the range of low-resolution and further divide it into Upper Low Resolution, Moderately Low Resolution and Very Low Resolution. Most face recognition methods, including commercial systems that are developed for high-resolution face recognition methods, can perform well on Upper-Low-Resolution images. The performance of commercial systems become worse for Moderately-Low-Resolution images while methods that are aimed at low-resolution face recognition still perform well. Very-Low-Resolution facial images are very difficult to recognize while simple holistic methods perform the best.

Because the lack of low-resolution images, most of the existing low-resolution face recognition methods are trained and tested using down-sampled images. In this thesis, we test various face recognition methods and demonstrate that down-sampled images are not fully representative of realistic low-resolution images. Face recognition methods perform much better on down-sampled images than on real low-resolution images. We further demonstrate that, under controlled situations (in absence of pose and illumination variations), inaccurate alignment is the major problem that causes the poor recognition performance on real low-resolution images; when images are captured under uncontrolled situations, alignment still plays an important role, but performance also degrades due to pose and illumination variations. In addition, we propose to use matching-score based registration to achieve better alignment and hence better face recognition performance. Our experimental results show that matching-score based registration significantly improves the face recognition performance of most of the face recognition methods.

In the scenario of low-resolution galleries and high-resolution probes, if the surveillance camera is monitoring a large area, the probe images will be of
various low-resolutions depending on the distances between the subject and the camera. This is a difficult task because a face recognition system is usually designed for a certain resolution, and it is not guaranteed that it also works for other resolutions. Therefore, we investigate whether it is beneficial for the recognition performance of long range face recognition to combine several classifiers that are tuned to images of different resolutions. Our experimental results show that if a classifier is only trained on images captured at a single distance, it could not perform well on the images from a very different distance. However, if we combine the images captured at various distances for training, a single classifier can perform at least as well as a combination of different classifiers when each of them are trained on images captured at a single distance.

In conclusion, we propose solutions to compare low-resolution probes with high-resolution galleries which significantly outperform the state-of-the-art on surveillance quality facial images. We emphasise that realistic low-resolution material should be used for training and testing. We focus attention on developing face recognition methods that can actually be useful for real-life applications. We bring an important step forward of low-resolution face recognition for forensic search.
List of Figures

1.1 Surveillance images from a robbery [1]. .......................... 2
1.2 Resolution scale. .................................................. 3
1.3 Face recognition process. ......................................... 4

3.1 (a) Sample face images from the FRGC v1.0 database at various resolutions, (b) identification rates .......................... 24
3.2 Sample face images from the SCface database captured at: (a) distance1, (b) distance2, (c) distance3 and (d) mug-shots. ..... 25
3.3 Identification rates for various resolutions with bicubic interpolation on the FRGC database .................................. 26
3.4 Comparison of DSR method with NMCF method on the FRGC database: (a) LR 7 × 6, (b) LR 14 × 12. ......................... 27
3.5 Comparison of DSR method with NMCF method on the SCface database. .................................................. 28
3.6 SR images constructed by DSR method (a) on the FRGC database (b) on the SCface database. ............................... 28
3.7 Sample image frames from the HumanID database. .......... 35
3.8 ROC curves of PCA results, downsampled vs. real: (a) downsampled, Config. 1; (b) downsampled, Config. 2; (c) downsampled, Config. 3; (d) real, Config. 1; (e) real, Config. 2; (f) real, Config. 3. ................................. 36
3.9 Face recognition results, GAR@FAR=0.1: (a) PCA, (b) LDA, (c) LBP. .................................................. 36
3.10 ROC curves of SR results using RL/DSR methods with PCA, Config. 2: (a) dowsampled, RL; (b) dowsampled, DSR; (c) real, RL; (d) real, DSR. ................................. 37
3.11 Super-resolution RL/DSR/NMCF results, GAR@FAR=0.1: (a) PCA, Config. 2; (b) PCA, Config. 3; (c) LDA, Config. 2; (d) LDA, Config. 3; (e) LBP results; (f) NMCF method. (ds = downsample) ................................................. 38

3.12 Reconstructed SR images by RL method from (a) real LR images, resolution 50×50, 30×30 and 25×25; (b) downsampled images, resolution 30×30, 25×25 and 20×20; (c) downsampled images, resolution 15×15, 10×10 and 5×5. ................. 38

4.1 Block diagram of the classifier according to (4.8). ................. 47

4.2 Sampled images from SCface database in our experiments. First row: HR, second row: LR ......................................................... 48

4.3 ROC curves for comparing MRBC to CBD for HR vs. LR setting. ................................................................. 49

4.4 ROC curves using MRBC for all the three settings. .............. 50

4.5 Sample images from our own database. ......................... 59

4.6 Verification rate (VR) at FAR 10% on images from various distances using different classifiers: (a) PCA (b) LDALLR (c) LBP (d) CLPM. .............................................. 61

4.7 Block diagram of the mixed-resolution classifier according to (4.23). ................................................................. 66

4.8 Histograms illustrating the Normal distribution for facial features for MixRes. Top row: histograms of HiRes feature elements; Bottom row: histograms of LoRes feature elements. .. 69

4.9 Sample images from the SCface database in the experiments in Section 4.3.6.1. First row: HiRes, second row: LoRes. ........ 70

4.10 ROC curves for comparing MixRes to CLPM and FaceVACS. (a) multi-gallery (b) single gallery. Probe: dist2, 15×12 pixels. Gallery: dist3, 30×24 pixels. ................................. 72


4.12 Sample images from the SCface database in the experiments in Section 4.3.6.2. First row: mug-shots, second row: dist3, third row: dist2, last row: dist1. ......................... 74

5.1 Resolution scale. .............................................................. 85

5.2 Down-sampled and real low-resolution image examples. (a) High-resolution, IPD 96 pixels (b) down-sampled, IPD 10 pixels (c) real, IPD 10 pixels (d) real with pose, IPD 12 pixels (subject from the Human ID data set) ......................... 87
5.3 Face recognition processing flow of (a) real low-resolution facial images (b) down-sampled facial images (commonly) (c) down-sampled images (proposed) .................................................. 88
5.4 Sample images from the UT-FAD data set (with IPD in the brackets) ................................................................. 89
5.5 Comparing ‘REAL’, ‘ADS’ and ‘DSA’ settings on the UT-FAD data set. X axis: IPD (pixels); Y axis: verification rate (VR) at FAR 0.1 .......................................................... 92
5.6 Sample images after pre-processing from the Human ID data set. ................................................................. 94
5.7 Comparing ‘REAL’, ‘ADS’ and ‘DSA’ settings on the Human ID data set. X axis: IPD (pixels); Y axis: verification rate (VR) at FAR 0.1 .......................................................... 95
5.8 Comparing ‘REAL’, ‘ADS’ and ‘DSA’ settings on the Human ID data set using COX training. X axis: IPD (pixels); Y axis: verification rate (VR) at FAR 0.1 .......................................................... 97
5.9 Comparing ‘REAL’, ‘ADS’ and ‘REALM’ settings on the UT-FAD data set. X axis: IPD (pixels); Y axis: verification rate (VR) at FAR 0.1 .......................................................... 101
5.10 Comparing ‘REAL’, ‘ADS’ and ‘REALM’ settings on the Human ID data set. X axis: IPD (pixels); Y axis: verification rate (VR) at FAR 0.1 .......................................................... 102
5.11 Comparing ‘REAL’, ‘ADS’ and ‘REALM’ settings on the Human ID data set with COX training. X axis: IPD (pixels); Y axis: verification rate (VR) at FAR 0.1 .......................................................... 102
6.1 Sample images of each resolution after pre-processing. ...... 111
6.2 Verification results of training and testing with images of different resolutions. X axis: probe image resolution, Y axis: Verification Rate (VR) at FAR 0.1 ........................................... 111
6.3 Verification results of classifiers trained with different resolution divisions of the training data. X axis: probe image resolution, Y axis: Verification Rate (VR) at FAR 0.1 ........................................... 113
6.4 Comparison of DIV4, Train70 and Train23 to the commercial face recognition system. X axis: probe image resolution, Y axis: Verification Rate (VR) at FAR 0.1 ........................................... 114
List of Tables

1.1 Commonly used face datasets. $N_s$: number of subjects. $N_i$: number of images. ........................................ 6


2.2 Papers using real low-resolution data as probe. $N_C$: number of subjects for testing. $N_G$: number of images per subject in the gallery set. $N_P$: number of images per subject in the probe set. R-1: Rank-1 recognition rate. Lo: low-resolution. Hi: high-resolution. Sup: super-resolution. As for the SCface database, the images from dist1, dist2 and dist3 are captured at a distance of 4.2 m, 2.6 m, 1.0 m. ........................................ 15

3.1 Identification rates [%] for PCA, LDA and LBP on the SCface database. ........................................ 25

3.2 Comparison of the identification rates [%] of three schemes on the FRGC database. ................................. 29

4.1 Parameters of MRBC in each experiment. The meaning of the parameters can be found in Section 4.2.3. ............. 49

4.2 Comparison of MRBC to CBD. The values are in the format: average value (standard deviation). ....................... 49

4.3 MRBC results using single gallery image per subject. The values are in the format: average value (standard deviation). The verification rates are obtained at false acceptance rate 0.1. .... 50

4.5 Papers using real LoRes data as probe. \( N_C \): number of subjects for testing. \( N_C \): number of images per subject in the gallery set. \( N_P \): number of images per subject in the probe set. L–H: LoRes–HiRes; S–H: SupRes–HiRes. As for the SCface database, the images from dist1, dist2 and dist3 are captured at a distance of 4.2 m, 2.6 m, 1.0 m.

4.6 Parameters of MixRes and its combination with MSBR and ET. \( M_w \) and \( N_w \) are the number of feature vectors of HiRes and LoRes training after the first dimensionality reduction. \( D \) is the number of feature vectors after the second dimensionality reduction.

4.7 Resolutions (pixels) of gallery and probe images in each subsection.

4.8 Comparison of MixRes to CBD, DSR, CLPM and FaceVACS. The values are in the format: average value (standard deviation). The verification rates are obtained at FAR = 10%. Probe: dist2, 15 × 12 pixels. Gallery: dist3, 30 × 24 pixels.


4.10 Results on the SCface database with different experiment settings. Gallery: mug-shots, 80 × 80 pixels. Probe: dist1, 32 × 32 pixels. The values are in the format: average value (standard deviation). The verification rates (%) are obtained at FAR = 10%. The abbreviation \( ds \) stands for down-sampled images. \( d1 \), \( d2 \) and \( d3 \) stand for dist1, dist2 and dist3, respectively.

4.11 Results on our own database. Gallery: 1m, 80 × 80 pixels. Probe: 8m, 32 × 32 pixels. The verification rates are obtained at FAR = 10%.

5.1 Parameter settings.

5.2 Three probe settings.
5.3 Details of our experimental data from the Human ID data set.  
Ni: total number of images. Ns: number of subjects.  . . . . . .  94
5.4 Three probe settings with matching-score based registration.  . 100

6.1 Number of images and subjects of each resolution used in our  
experiments. Res: resolution. Ni: total number of images. Ns:  
number of subjects.  . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
6.2 Division of resolutions for training in the second experiment.  . 112
Chapter 1

Introduction

1.1 Background

In 2015, a man walked into a pizza shop in the USA and ordered a steak, a cheese sandwich and a meatball sub [1]. When the man was asked for payment, he bent over to check his pockets and came up holding a handgun. He pointed the gun at the employee of the pizza shop and grabbed the cash from the store register, then turned and ran out. When the police arrived, they took witness statements and arranged to get video surveillance pulled from any nearby businesses that might have caught footage of the robber. The surveillance photos of the suspected robber were released to get information from the public, as shown in Fig. 1.1.

This is a typical forensic case where surveillance quality facial images are presented. In forensic cases there are three types of usage of the facial images. The first one is search, which means to search the police database to see if any facial image in the database matches the person on the surveillance images. The second type is called intelligence investigation. The surveillance images are compared with another crime scene to see if they originate from the same person. The third one is individualization, which is to provide evidence whether a certain suspect is in these photos or not. For all the three types, an automatic face recognition system can be used. Existing face recognition techniques are very successful in recognizing high-resolution facial images. However, their performance is not sufficient on the surveillance quality images like in Fig. 1.1. Those images are usually of low-resolution, blurry, with various poses and occlusion. In this thesis, we will study low-resolution face recognition
Fig. 1.1: Surveillance images from a robbery [1].

techniques that may support forensic case work.

1.2 Low-resolution face recognition problems

A face recognition system identifies or verifies the identity of a person from a digital facial image. An input image with unknown identity is called a probe image. An image which is pre-stored in the database with known identity is called a gallery image. The face recognition system compares a probe image and a gallery image and outputs a decision whether they are from the same subject or not (verification), or it compares a probe image with all the gallery images and outputs the identity of a matching gallery image as the probe identity (identification).

Low-resolution face recognition means that at least the probe images are of low-resolution. Gallery images are not necessarily of low-resolution. In real life applications, it is very common that the gallery images are pre-stored high-resolution images, for example, mug-shots of convicted criminals or suspects. Low-resolution facial images are difficult to recognize not only because limited information is contained in the small number of pixels, but also because the recordings of low-resolution facial images are usually without user cooperation in uncontrolled situations, which results in motion blur, variation of pose and
illumination, and occlusion. In this work, we focus on the low-resolution problem, but also take into account small variations in pose and illumination.

1.3 The scale of resolutions

In literature on low-resolution face recognition, what in some papers is considered low-resolution, is still considered high-resolution in other papers. For example, the Interpupillary Distances (IPD) of low-resolution images described in [2] are from 2 pixels to 8 pixels, and the IPD of high-resolution images is 16 pixels. While in [3], the IPD of low-resolution images are 20 pixels. To harmonize the terminology in low-resolution face recognition, we propose the resolution scale as shown in Fig 1.2. In this graph we use IPD as a reference when we measure resolution. The IPD is the distance in pixels between the centers of the eyes.

Four biometric standards are shown in this graph: ISO/IEC 19794-5:2005 [4], ANSI/INCITS 385-2004 [5], the European norm EN 50132-7 [6], and ICAO Doc 9303 [7]. These standards provide recommended facial image resolutions for recognition. We use them in this graph as separating points for resolution divisions. We consider an IPD of 50 pixels as the separation of high resolution and low resolution. We further divide low resolution into Upper Low Resolution (ULR), Moderately Low Resolution (MLR) and Very Low Resolution (VLR). ULR (IPD 25 to 50 pixels) is a relatively higher resolution and not a very difficult task for most existing face recognition methods. Images of MLR (IPD 13 to 25 pixels) are harder to recognize than ULR. Methods that are designed for low-resolution face recognition outperform high-resolution face recognition methods at MLR. VLR (IPD below 13 pixels) is extremely difficult for face recognition and most face recognition methods perform poorly on these images. The image resolutions of several commonly used face data sets, FRGC v2.0 [8], LFW [9], SCface [10], and HumanID [11], are also included in this graph. Detailed information of this graph can be found in Chapter 5.

![Resolution scale](image_url)
In this work, we address problems of low-resolution face recognition on VLR images.

1.4 The process of face recognition

Face recognition follows the process as shown in Fig. 1.3. Preprocessing makes sure the input images can be used by the face recognition system. It usually includes face detection, landmark localization, image alignment, histogram equalization and so on. It is very important that during the preprocessing the facial images are registered to the same standard, for example, the eyes and noses are in the same pixel location in the images. During the next stage, the features that are useful for recognition are extracted from the images. Then the features from two images are compared and result in a score. This score is used to determine the input probe image’s identity. Usually a threshold is set: if the score is higher than the threshold, this probe image has the same identity as the gallery image. This is called Verification: verify if two images have the same identity. We can also compare the probe image with all the gallery images, and pick the identity of the gallery image which has the highest score as the identity of the probe image. This is called Identification: identify an image’s identity.

Fig. 1.3: Face recognition process.

1.5 Approaches to low-resolution face recognition

Low resolution face recognition has low-resolution facial images as inputs. We have previously mentioned that we focus on a common scenario, where the probe images are of low-resolution and the gallery images are of high-resolution. A problem of performing face recognition for this scenario is that the gallery and probe images are of very different resolutions, but most face recognition systems require them to be of the same resolution. To solve the
problem, we can enhance the resolution and quality of the low-resolution probe images and then compare them with the high-resolution gallery images using available general face recognition methods designed for high resolution. We can also first down-sample the gallery images and then compare the probe images with them in the low-resolution space. It is also possible to design a method which can directly compare images of different resolutions. The last two approaches both need dedicated solutions for low-resolution face recognition.

The approach that enhances the resolution of an image is called super-resolution. It was initially designed for visual enhancement. However, a good visual appearance does not imply better face recognition performance, because the enhancement cannot use data from the specific subject and therefore uses general facial data, which does not improve individualisation. Therefore, constraints have been introduced in super-resolution methods, aiming at improving face recognition performance, when reconstructing super-resolution images [12,13].

Low-resolution face recognition methods can be designed for low-resolution-to-low-resolution comparison or low-resolution-to-high-resolution comparison. If the gallery images are of high-resolution, they have to be down-sampled before performing low-resolution-to-low-resolution comparison. However, the high-frequency information, which can be useful in face recognition, is lost in the down-sampling process. To make maximum use of the gallery image information, recently researchers have put more effort into low-resolution-to-high-resolution comparison to seek for an optimal solution [14,15].

A face recognition system is usually designed for a certain resolution, and it is not guaranteed that it also performs well for other resolutions. Commercial face recognition systems usually have a recommended resolution to ensure good performance. If the input image is much smaller than that size, the performance drops. However, if a surveillance camera is monitoring an open area, for example a parking lot, people have various distances to the camera and therefore the captured faces will have various resolutions. There are two approaches to deal with this situation. The first one is to improve acquisition devices so that the images captured at the farthest distance have a high enough resolution for recognition [16–18]. The second approach is to design a face recognition system for images captured at various distances, but there is only a limited number of publications on this topic [19,20].

The performance of a face recognition method needs to be evaluated using test images, and face recognition methods usually requires many facial images for
training. The quality of the images is very important for the performance of a face recognition system. A face recognition system usually performs well if the training and testing images are captured at similar situations, and it will perform poorly when the training and testing images are very different. So a low-resolution face recognition system requires low-resolution training images to ensure good performance on low-resolution facial images. Therefore, low-resolution face data sets are very important for developing low-resolution face recognition methods. There are many publicly available data sets of facial images, but there is a lack of low-resolution data sets. In Table 1.1 we listed some commonly used face data sets.

Table 1.1: Commonly used face datasets. \( N_s \): number of subjects. \( N_i \): number of images.

<table>
<thead>
<tr>
<th>Database</th>
<th>( N_s )</th>
<th>( N_i )</th>
<th>( IPD_{probe} )</th>
<th>Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRGC v2.0 [8]</td>
<td>568</td>
<td>50,000</td>
<td>300</td>
<td>illumination</td>
</tr>
<tr>
<td>Multi-PIE [21]</td>
<td>337</td>
<td>750,000</td>
<td>70</td>
<td>pose, illumination</td>
</tr>
<tr>
<td>FERET [22]</td>
<td>1199</td>
<td>14,126</td>
<td>60</td>
<td>illumination</td>
</tr>
<tr>
<td>LFW [9]</td>
<td>5749</td>
<td>13,233</td>
<td>60</td>
<td>all</td>
</tr>
<tr>
<td>ChokePoint [23]</td>
<td>29</td>
<td>64,204</td>
<td>20 to 60</td>
<td>pose, illumination</td>
</tr>
<tr>
<td>SCface [10]</td>
<td>130</td>
<td>4160</td>
<td>13, 22, 31</td>
<td>pose</td>
</tr>
<tr>
<td>Human ID [11]</td>
<td>315</td>
<td>504 videos</td>
<td>9 to 30</td>
<td>illumination</td>
</tr>
</tbody>
</table>

Multi-PIE, FERET, FRGC and LFW are high-resolution face data sets, as their IPDs are above 50 pixels. FRGC and FERET contain frontal images, and Multi-PIE has images taken at different viewing angles. LFW images are taken at uncontrolled situations. ChokePoint, Human ID and SCface data sets contain low-resolution images. When we compare the number of subjects and images in the high-resolution face data sets and those in the low-resolution data sets, we can clearly see that the low-resolution data sets are significantly smaller. ChokePoint has many images but only 29 subjects. The results on this data set are not representative for a large population.

Because of the lack of low-resolution images, most of the existing low-resolution face recognition methods are trained and tested using down-sampled images. It is questionable, however, if the face recognition performance on down-sampled images reflects the performance on real low-resolution images. There are many differences between down-sampled and real low-resolution images. One im-
portant factor is that the landmarks of the down-sampled images can be accurately located at high-resolution before down-sampling, while locating the landmarks on the real low-resolution images is very difficult. Therefore the down-sampled images are usually perfectly aligned with the gallery images that they are compared with, while real low-resolution are poorly aligned.

Deep-learning is very popular in high-resolution face recognition. It has been proved to be effective on unconstrained facial images from the LFW data set [24]. Deep-learning techniques can also be applied to low-resolution face recognition. However, deep-learning requires a large amount of data for training while there is only a limited amount of realistic low-resolution facial data available as we have mentioned. Some researchers train deep-network using down-sampled images for low-resolution face recognition. However, in our experiments on VLR probe images, where we have tested a state-of-the-art deep-learning based low-resolution face recognition method [25], the results of this deep-learning method are worse than simple low-resolution face recognition methods (Chapter 5). Therefore, we decided not to use deep-learning as a major tool to solve low-resolution face recognition problems in this work. But, when enough low-resolution data becomes available, it is good to investigate deep-learning for recognition of VLR facial images.

1.6 Research questions

In the previous sections we have identified the following problems in low-resolution face recognition: resolution mismatch of gallery and probe images, using down-sampled images instead of real low-resolution images when testing new methods, the lack of low-resolution face data sets, and problem of handling images of various resolutions. Therefore the research questions addressed in this thesis are as follows:

RQ.1 Which approaches exist for comparison of low-resolution probe facial images to high-resolution gallery facial images? And what is the optimal approach?
   (a) Does super-resolution benefit low-resolution face recognition?
   (b) Can we design a method that allows direct comparison of facial data from different domains, in particular low-resolution and high-resolution facial images, that fully exploits the information from both domains without transforming images from the one domain to the other?
RQ.2 What are the limitations of using down-sampled images instead of real low-resolution images in testing and training classifiers?
(a) What is the main cause of the difference in face recognition performance when testing on real low-resolution images and down-sampled images?
(b) How can this knowledge be exploited to improve face recognition performance on real low-resolution images?
(c) What are the benefits of training a classifier with real low-resolution images as opposed to down-sampled images?

RQ.3 How to design a face recognition system that operates well on a large range of resolutions?

RQ.4 Can we improve the performance of face recognition on low resolution surveillance quality images to the extend that it contributes to forensic investigation?

1.7 Contributions

The work presented in this thesis has several contributions to the field of low-resolution face recognition:

1. Improve the state-of-the-art face recognition of low-resolution to high-resolution comparison.
(a) Identification of the limitation of super-resolution methods for face recognition.
(b) A new method, mixed-resolution biometric comparison, to directly compare low-resolution probe images to high-resolution gallery images. This method, when trained with appropriate data, is also applicable to other types of heterogeneous biometric recognition.

2. Explore the problem of using down-sampled images instead of real low-resolution images.
(a) Insights into the differences between down-sampled and real low-resolution images. Identification of the problem of poor alignment in low-resolution and provide a solution.
(b) A face data set which contains facial images taken at various distances in a controlled situation. It can be used to study the influence of distance to facial images.
(c) Demonstration of the benefit of using real low-resolution training data.

3. A method for face recognition across a range of resolutions.
4. An important step towards the applicability of surveillance data to forensic face recognition.

1.8 List of publications


1.9 Reading guide and overview of the thesis

Full publications are included in this thesis, thus there will be some duplications. In Chapter 2 to Chapter 6, the first section (Introduction) gives a reading guide, and the last section (Discussion) gives conclusions of this chapter.

This thesis is organized as follows: Chapter 2 provides a literature review of related subjects. Chapter 3 until Chapter 6 are composed of one or two publications listed in Section 1.8. Chapter 3 studies super-resolution and its benefits to low-resolution face recognition. Chapter 4 presents a novel face recognition method for low-resolution to high-resolution comparison. It also identifies various problems when testing low-resolution face recognition methods. Chapter 5 provides a comprehensive study on the importance of alignment in low-resolution face recognition. Chapter 6 studies low-resolution face recognition for a range of resolutions and provides a solution to optimize the training process. Chapter 7 concludes the thesis.
Chapter 2

Related work

2.1 Introduction

This chapter introduces literatures on super-resolution, low-resolution face recognition and low-resolution face alignment. The contents of this chapter have overlap with the introductions of our publications which are presented in the following chapters. The readers can decide to skip either part of the thesis.

2.2 Super-resolution

Super-resolution is the technique that enhances the resolution of an imaging system. It was originally developed for visual enhancement. The simplest super-resolution approach is interpolation.

Super-resolution methods can be either reconstruction-based or learning-based. Reconstruction-based super-resolution methods use multiple low-resolution images to reconstruct a single high-resolution image. For example, Schultz and Stevenson [26] proposed a method to reconstruct a high-resolution image from a low-resolution video sequence. This method is based on an observation model which models the subsampling of the unknown high-resolution data and accounts for independent object motion occurring between frames. Learning-based super-resolution methods learn a transformation matrix from training image pairs of both high-resolution and low-resolution. The learnt
transformation matrix can be applied on a single low-resolution image to reconstruct a high-resolution image. For example, Freeman and Pasztor [27] modelled the relationship between high- and low-resolution images, and between neighbouring local high-resolution regions using Markov network. The parameters of the network are learnt from training data.

Super-resolution techniques that are specially developed for enhancing the quality of facial images are also called face hallucination. Face hallucination uses prior knowledge of typical face features. In 2000, the first face hallucination method is proposed by Baker and Kanade [28]. They used Gaussian pyramid to model the relation between low-resolution and high-resolution images. Bayesian maximum a posteriori framework is used to build the objective function with the prior predicted by Gradient Prior Prediction Algorithm. Then Gradient Descent is used for optimization. This method is applicable for input of both a single image or multiple images, and the reconstructed images have better visual quality than the ones reconstructed by [26]. Wang and Tang [29] proposed a face hallucination method using eigen-transformation. The input low-resolution image is represented as a linear combination of the low-resolution images in the training set by Principal Component Analysis. The super-resolution image is reconstructed using the corresponding high-resolution training images with the same coefficients.

Super-resolution for face recognition is however different than face hallucination. Sometimes, face hallucination even introduces artifacts in the reconstructed images which confuse the face recognition classifiers. Therefore, new methods have been developed for improving face recognition performance on low-resolution images. Those methods usually add constrains which make use of the class information of the images, and often the super-resolution process is applied on face features instead of original facial images. For example, Gunturk et al. [3] proposed to apply super-resolution in an eigen-domain that reconstructs only the necessary information for recognition. Zou and Yuen [13] developed a data constraint to minimize both the distances between the constructed super-resolution images and the corresponding high-resolution images and the distances between super-resolution images from the same class. The first part of the constraint can be used alone to reconstruct super-resolution images. The second part of the constraint, which is the class information, is aiming for better recognition performance, but results in bad quality images in visual aspect. Hennings-Yeomans et al. [30] built a model for super-resolution based on Tikhonov regularization and a linear feature extraction stage. This model can be applied when images from training, gallery and probe sets have varying resolutions. This approach was extended in [12] by adding a face prior
to the model and using relative residuals as measures of fit. Bilgazyev et al. [31] proposed a method that uses dual-tree complex wavelet transform to extract high frequency components of training images. Then the super-resolution features are represented as a weighted combination of the high-resolution training images and the weights are the same as the ones that represent the input low-resolution probe using corresponding low-resolution training images. Zhang et al. [32] proposed a super-resolution method in morphable model space, which provides high-resolution information required by both reconstruction and recognition.

There is also another type of approach of comparing low-resolution to high-resolution images which is similar to super-resolution: map high-resolution and low-resolution image features into a common space where the feature dimensionalities are the same so that the comparison is possible. For example, in [33], Biswas et al. proposed an approach using multidimensional scaling to transform the high-resolution gallery and low-resolution probe images to a common space so that the distance between transformed features of low-resolution images can be as close as possible to the corresponding high-resolution images. Li et al. [34] proposed a method to obtain coupled mappings that project both high-resolution and low-resolution image features to a unified feature space in which direct comparison of high-resolution and low-resolution is possible. The objective function is built to cluster the projections of low-resolution images and their corresponding high-resolution images in the new feature space. In our early-published papers, we classify this type of approach also as super-resolution, because of the following two reasons. Firstly, this approach maps the low-resolution features to a higher dimensionality which is similar to feature based super-resolution methods. Secondly, the application of this approach is the same as the recognition-oriented super-resolution methods and super-resolution methods are always compared with in the recognition performance testing experiments. However, this approach is also different from super-resolution as it aims at a common space and the high-resolution gallery images also need to be transformed. Thus, in our later work and also in this thesis, we classify this approach as direct low-to-high comparison in low-resolution face recognition. We will discuss this approach in Section 2.3.

2.3 Low-resolution face recognition

Face recognition for low resolution is different from face recognition for high resolution. Firstly, it is much harder to detect landmarks reliably and ac-
accurately in low-resolution images. Secondly, low-resolution images contain far less discriminative information than high-resolution images. There is a small but growing body of literature that specifically addresses the problem of low-resolution face recognition. Important references of low-resolution face recognition are listed in Table 2.1 and Table 2.2. Table 2.1 lists papers that present experiments on down-sampled probe images and Table 2.2 lists papers using real low-resolution probe images. We separate the papers in two tables because the evaluation of the two types of data is different. For more information of the difference between down-sampled and real low-resolution images, please refer to Chapter 4 and Chapter 5. In the tables, we include the methods, the experimental settings and rank-1 recognition rates. As we can see, each paper presents experiments in a different setting even when they use the same database. In Chapter 4, we will discuss about the experimental protocols.


<table>
<thead>
<tr>
<th>Method</th>
<th>Database</th>
<th>$N_C$</th>
<th>$N_G$</th>
<th>$N_P$</th>
<th>IMsize</th>
<th>Rank-1</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKE [14]</td>
<td>Multi-PIE</td>
<td>229</td>
<td>7</td>
<td>13</td>
<td>$6 \times 6$</td>
<td>88%</td>
<td>LoRes–HiRes</td>
</tr>
<tr>
<td>CDA [2]</td>
<td>Multi-PIE</td>
<td>100</td>
<td>10</td>
<td>10+</td>
<td>$6 \times 6$</td>
<td>79%</td>
<td>LoRes–HiRes</td>
</tr>
<tr>
<td>MDS [35]</td>
<td>Multi-PIE</td>
<td>237</td>
<td>1</td>
<td>1</td>
<td>$12 \times 10$</td>
<td>81%</td>
<td>LoRes–HiRes</td>
</tr>
<tr>
<td>DTCWT [31]</td>
<td>Multi-PIE</td>
<td>202</td>
<td>1</td>
<td>20</td>
<td>$20 \times 20$</td>
<td>99%</td>
<td>SupRes–HiRes</td>
</tr>
<tr>
<td>SDA [36]</td>
<td>Multi-PIE</td>
<td>149</td>
<td>1</td>
<td>1+</td>
<td>$12 \times 12$</td>
<td>70%</td>
<td>LoRes–HiRes</td>
</tr>
<tr>
<td>S2R2 [12]</td>
<td>Multi-PIE</td>
<td>224</td>
<td>1</td>
<td>13</td>
<td>$6 \times 6$</td>
<td>73%</td>
<td>SupRes–HiRes</td>
</tr>
<tr>
<td>MFF [37]</td>
<td>FERET</td>
<td>200</td>
<td>1</td>
<td>1</td>
<td>$12 \times 12$</td>
<td>84%</td>
<td>SupRes–HiRes</td>
</tr>
<tr>
<td>NMCF [38]</td>
<td>FERET</td>
<td>1195</td>
<td>1</td>
<td>1</td>
<td>$12 \times 12$</td>
<td>84%</td>
<td>LoRes–HiRes</td>
</tr>
<tr>
<td>CLPM [34]</td>
<td>FERET</td>
<td>1195</td>
<td>1</td>
<td>1</td>
<td>$12 \times 12$</td>
<td>90%</td>
<td>LoRes–HiRes</td>
</tr>
<tr>
<td>GDAMM [39]</td>
<td>AR</td>
<td>126</td>
<td>7</td>
<td>7</td>
<td>$8 \times 7$</td>
<td>72%</td>
<td>SupRes–HiRes</td>
</tr>
<tr>
<td>MM [32]</td>
<td>CMU video</td>
<td>68</td>
<td>1</td>
<td>16</td>
<td>$23 \times 23$</td>
<td>81%</td>
<td>SupRes–HiRes</td>
</tr>
<tr>
<td>EigenSR [3]</td>
<td>CMU video</td>
<td>68</td>
<td>1</td>
<td>16</td>
<td>$10 \times 10$</td>
<td>74%</td>
<td>SupRes–HiRes</td>
</tr>
<tr>
<td>EigenTr [29]</td>
<td>XM2VTS</td>
<td>295</td>
<td>1</td>
<td>1</td>
<td>$10 \times 10$</td>
<td>59%</td>
<td>SupRes–HiRes</td>
</tr>
<tr>
<td>DSR [40]</td>
<td>FRGC v2.0</td>
<td>311</td>
<td>8</td>
<td>2</td>
<td>$7 \times 6$</td>
<td>78%</td>
<td>SupRes–HiRes</td>
</tr>
</tbody>
</table>

Existing methods solve the resolution mismatch problem mainly by following the three approaches: applying super-resolution to low-resolution probes, down-sampling high-resolution gallery images and direct low-resolution to
Table 2.2: Papers using real low-resolution data as probe. \( N_C \): number of subjects for testing. \( N_G \): number of images per subject in the gallery set. \( N_P \): number of images per subject in the probe set. R-1: Rank-1 recognition rate. Lo: low-resolution. Hi: high-resolution. Sup: super-resolution. As for the SCface database, the images from dist1, dist2 and dist3 are captured at a distance of 4.2 m, 2.6 m, 1.0 m.

<table>
<thead>
<tr>
<th>Method</th>
<th>Database</th>
<th>Gallery</th>
<th>Probe</th>
<th>( N_C )</th>
<th>( N_G )</th>
<th>( N_P )</th>
<th>R-1</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKE [14]</td>
<td>SCface</td>
<td>mug-shot</td>
<td>dist1</td>
<td>130</td>
<td>1</td>
<td>5</td>
<td>8%</td>
<td>Lo–Hi</td>
</tr>
<tr>
<td>DSR [40]</td>
<td>SCface</td>
<td>dist2</td>
<td>dist1</td>
<td>130</td>
<td>5</td>
<td>5</td>
<td>22%</td>
<td>Sup–Hi</td>
</tr>
<tr>
<td>CBD [15]</td>
<td>SCface</td>
<td>dist3</td>
<td>dist2</td>
<td>100</td>
<td>4</td>
<td>1</td>
<td>53%</td>
<td>Lo–Hi</td>
</tr>
<tr>
<td>CDA [2]</td>
<td>local video</td>
<td>photo</td>
<td>video</td>
<td>161</td>
<td>5</td>
<td>5</td>
<td>53%</td>
<td>Lo–Hi</td>
</tr>
<tr>
<td>DTCWLT [31]</td>
<td>local video</td>
<td>photo</td>
<td>video</td>
<td>34</td>
<td>1?</td>
<td>1</td>
<td>56%</td>
<td>Sup–Hi</td>
</tr>
</tbody>
</table>

high-resolution comparison.

The first approach is to apply super-resolution to low-resolution probes. Since most face recognition systems are designed for high-resolution images, many researchers reconstruct high-resolution versions from the low-resolution probes using super-resolution techniques to make use of the information contained in the high-resolution gallery images and conduct comparison in the high-resolution space. The details of this approach was already discussed in the previous section. Though applying super-resolution to low-resolution probes makes the comparison to high-resolution galleries possible in the high-resolution space, the process of super-resolution usually brings in artifacts or noise (even when they are designed for recognition) which might influence the face recognition performance.

Down-sampling high-resolution gallery images and conducting comparison in the low-resolution space is a very simple way for low-resolution to high-resolution comparison. It requires lower computational costs than using high-resolution images. Although some information in the high-resolution images will be lost in the down-sampling process, it has been reported by some researchers that this approach has similar recognition performance as super-resolution methods for low-resolution face recognition. For instance, Hu et al. [41] conducted experiments using a video database of moving faces and people. Their experimental results show that applying super-resolution methods and then comparing to high-resolution galleries have similar performance as low-resolution to low-resolution comparison at a far range (5-10 pixel eye-to-eye distance). Xu et al. [42] showed that when image resolution is low enough, low-resolution to low-resolution comparison is superior to using
super-resolution methods. In their experiments conducted on Yale B and AR databases with down-sampled images, super-resolution methods perform much poorer than low-resolution to low-resolution comparison when the image size is $8 \times 8$ pixels. These results suggest that down-sampling gallery images and comparing to low-resolution probes has at least as good face recognition performance as applying super-resolution on low-resolution probes and comparing to galleries in the high-resolution domain.

Direct comparison of low-resolution probes and high-resolution galleries is a new area that has drawn researchers’ attention in recent years. Most methods of this approach find transformations for both low-resolution and high-resolution images and compare their features in a common space. This approach avoids losing information as a result of down-sampling high-resolution or adding artefacts by super-resolution. The mappings between high-resolution gallery and low-resolution probe data can also be learnt in such a way that different variations are modelled. Li et al. [34] proposed a method that projects both high-resolution galleries and low-resolution probes to a unified feature space for classification using coupled mappings. The mappings are learnt by optimizing the objective function that minimizes the difference between corresponding high-resolution and low-resolution images. Huang and He [38] proposed a method that uses canonical correlation analysis to project the PCA features of high-resolution and low-resolution image pairs to a coherent feature space. Radial based functions are then applied to find the mapping between the high-resolution and low-resolution pairs. A multidimensional scaling based method is proposed by Biswas et al. [35]. Both high-resolution and low-resolution images are transformed to a common space where the distance between them approximates the distance when they are both high-resolution. The transformations are learnt using an iterative majorization algorithm. Ren et al. [14] proposed a method called coupled kernel embedding. It projects the original high-resolution and low-resolution images onto reproducible kernel space using coupled nonlinear functions. The dissimilarities captured by their kernel Gram matrices are minimized in this space. Lei et al. [2] proposed a coupled discriminant analysis method. They find coupled transformations to project high-resolution and low-resolution images to a common space in which the low-dimensional embedding is well classified. The locality information in kernel space is also used as a constraint for the discriminant analysis process. This method is also suitable for images of different modalities, for example, visible and infrared faces. Moutafis and Kakadiaris [15] proposed a method that learns semi-coupled mappings for high-resolution and low-resolution images for optimized representations. The mappings aim at increasing class-separation for high-resolution images and projecting low-resolution images to
their corresponding class-separated high-resolution data.

### 2.4 Low-resolution face alignment

The low-resolution face recognition methods, including what we have mentioned above, usually assume both the gallery and probe images are perfectly aligned. However, unlike high-resolution images, alignment on low-resolution images especially very low-resolution images are very difficult.

The most commonly used face and facial landmark detection algorithm is the Viola-Jones face detector \[43\]. However, when the IPD of the facial images is below 10 pixels, the detected face area or eye areas are no longer accurate. Without accurate landmarks, the facial images cannot align well.

One solution is to promote face alignment method that can be applied to low-resolution facial images. For example, Ban et al. \[44\] proposed a method which assumes that face alignment is a classification problem to separate well-aligned faces from nonaligned faces. They use the confidence value of Real-AdaBoost to assess how well the face is aligned. They search spaces with ten different sizes and five different angles and find one with the highest confidence value. In \[45\], Zhang et al. proposed a alignment method using supervised descent method model. They apply super-resolution first when the image is of very low resolution, and then find the landmarks on the reconstructed high-resolution images.

It is also possible to design face recognition methods which is robust again poor alignment. Huang et al. \[46\] proposed a method which take care of both alignment and recognition at the same time. They use sparse representation prior that if the alignments of video faces are accurate, they can be represented as good linear combinations of well-aligned gallery still faces. They seek an optimal set of deformations for the low resolution video sequence simultaneously with their sparse representations over the gallery dictionary.

Although some researches have been conducted considering the alignment of low-resolution facial images, the techniques are yet not mature. In most of the works, manual landmarks are used for alignment. In this thesis, we will discuss the accuracy of manual landmarking and the influence to low-resolution face recognition.
2.5 Discussion

Low-resolution face recognition is still an challenging task. Despite lots of work has been done on low-resolution face recognition, the performance of the state-of-the-art on low-resolution facial images is still far below the performance on high-resolution facial images, and is not satisfactory for real-life applications. Besides, there is a lack of common protocol for research. The definition of low-resolution is not clear, different sizes of images are mentioned as low-resolution in different papers. Methods are tested under experimental protocol which varies in the resolutions of gallery and probe image, size of training sets, and etc. Therefore, it is very difficult to compare between different methods. Another problem is that most of the methods are tested on down-sampled images instead of real low-resolution images. It is questionable if the results represent the effectiveness of the methods in real-life applications. Face recognition on down-sampled images has advantages for exampled that the images can be perfectly aligned at high-resolution. In this thesis, we will explore these problems.
Chapter 3

Applying super-resolution for low-resolution face recognition

3.1 Introduction

One way to perform face recognition of low-resolution to high-resolution comparison, is to apply super-resolution on the low-resolution probe images and then use a high-resolution face classifier to compare them to the high-resolution gallery images. In this chapter, we follow this approach and evaluate how effective it is. This chapter contains two publications [47,48] which were published in the 33rd and 34th WIC Symposium on Information Theory in the Benelux. In Section 3.2, we evaluates the performance of three face recognition methods and the effect on recognition of two super-resolution methods. The experiments are conducted on two data sets. One of them is a commonly used high-resolution face data set, and we down-sample the images to generate low-resolution inputs. The other data set contains surveillance quality facial images. This section focuses on the face recognition performance differences between different methods. Section 3.3 continues the work of Section 3.2 and focuses on comparing super-resolution benefits on real and down-sampled low-resolution face images. A data set contains real low-resolution facial images from video sequences is used in this section.
3.2 An evaluation of super-resolution for face recognition

3.2.1 Abstract

We evaluate the performance of face recognition algorithms on images at various resolutions. Then we show to what extent super-resolution (SR) methods can improve the recognition performance when comparing low-resolution (LR) to high-resolution (HR) facial images. Our experiments use both synthetic data (from the FRGC v1.0 database) and surveillance images (from the SC-face database). Three face recognition methods are used, namely Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Local Binary Patterns (LBP). Two SR methods are evaluated. The first method learns the mapping between LR images and the corresponding HR images using a regression model. As a result, the reconstructed SR images are close to the HR images that belong to the same subject and far away from others. The second method compares LR and HR facial images without explicitly constructing SR images. It finds a coherent feature space where the correlation of LR and HR is maximum, and then compute the mapping from LR to HR in this feature space. The performance of the two SR methods are compared to that delivered by the standard face recognition without SR. The results show that LDA is mostly robust to resolution changes while LBP is not suitable for the recognition of LR images. SR methods improve the recognition accuracy when downsampled images are used and the first method provides better results than the second one. However, the improvement for realistic LR surveillance images remains limited.

3.2.2 Introduction

Face recognition has gained much attention in recent decades [49]. Face recognition systems deliver promising results when using high-resolution (HR) frontal images, but face recognition at a distance remains challenging. The face regions of images acquired at a distance are usually small and of low quality. To deal with the low-resolution (LR) problem of face images, super-resolution (SR) methods can be applied to increase the resolution of an image.

---

SR was initially intended to construct HR images for visual enhancement. These methods have achieved great success, but the objective of most SR methods is to construct high-frequency details that is insufficient for the recognition of LR images. Recently, some SR methods have been developed specially for face recognition problem. Hennings-Yeomans et al. [30] built a model for SR based on Tikhonov regularization and a linear feature extraction stage. This model can be applied when images from training, gallery and probe sets have varying resolutions. This approach was extended in [12] by adding a face prior to the model and using relative residuals as measures of fit. In [33], Biswas et al. proposed an approach using multidimensional scaling to improve the matching performance of LR images. Their method finds a transformation matrix so that the distance between transformed features of LR images can be as close as possible to the corresponding HR images. Identity information about subjects is also used to make sure that the distance is small between data from the same class. Li et al. [34] proposed a method to obtain coupled mappings that project both HR and LR image features to a unified feature space in which direct comparison of HR and LR is possible. The objective function is built to cluster the projections of LR images and their corresponding HR images in the new feature space. A face recognition system for long video sequences is presented by Nasrollahi and Moeslund [50]. In [50], key-frames are first selected and then a hybrid SR method is applied. The images that are closest to full-frontal and with higher quality score are chosen as the key-frames. Multiple images from the key-frames are used to construct HR images.

In this paper, we first evaluate the performance of three standard face recognition algorithms (PCA [51], LDA [52] and LBP [53]) at various image resolutions and then apply two SR methods to LR images in order to observe their contribution to the identification performance. In our experiments, two face databases are used: the first face database (FRGC v1.0 [8]) contains high-quality images captured at controlled situations. The second database, SCface [10], contains surveillance quality facial images captured at three different distances. The performance of two SR methods (DSR method [13] and Huang and He’s method [38]) are compared with standard face recognition without SR. The remainder of this paper is organized as follows. In Section 3.2.3, the two SR methods are introduced. Experimental setup and identification test results are presented in Section 3.2.4. Section 3.2.5 concludes the paper.
3.2.3 Super-resolution methods

Two state-of-the-art SR methods are chosen in our experiments. They are explained in detail in the following subsections.

3.2.3.1 DSR method

In [13], Zou and Yuen proposed a simple but effective SR method for very low resolution images which is compatible with various face recognition methods. This method is called discriminative super resolution (DSR). They introduce a data constraint which clusters the constructed SR images with the corresponding HR images. Identity information about the subject is also used to improve the recognition accuracy.

Given a set of HR and LR image pairs \( \{I^h_i, I^l_i\}_{i=1}^N \), the relation \( R \) is modeled as

\[
R = \underset{R'}{\arg \min} \frac{1}{N} \sum_{i=1}^N \left\| I^h_i - R'I_i^l \right\|^2 + \gamma d(R').
\]

(3.1)

where \( \gamma \) is a constant to balance the two terms. We set \( \gamma \) to 1 in our experiments. The first term of (3.1) minimizes the distance between HR and the space of LR projected by \( R \). The second term \( d(R') \) is represented as

\[
d(R') = \frac{1}{N(\lambda_i = \lambda_j)} \sum_{\lambda_i = \lambda_j} \left\| I^h_i - R'I_j^l \right\|^2 - \frac{1}{N(\lambda_i \neq \lambda_j)} \sum_{\lambda_i \neq \lambda_j} \left\| I^h_i - R'I_j^l \right\|^2
\]

(3.2)

where \( \lambda_i \) is the class label of \( I_i \). This makes sure the reconstructed HR images are clustered with the images from the same class and far away from those from other classes.

Thus, for a given LR image \( I_{input} \), we first apply \( I_{SR} = RI_{input} \) to obtain SR image \( I_{SR} \), and then use \( I_{SR} \) for face recognition.

3.2.3.2 NMCF method

In [38], Huang and He proposed a SR method where canonical correlation analysis (CCA) is used to project the PCA features of HR and LR image pairs to a coherent feature space. Radial based functions (RBFs) are then applied to find the mapping between the HR/LR pairs. This method finds
3.2 An evaluation of super-resolution for face recognition

nonlinear mappings on coherent features. Thus, we will refer to this method as NMCF method in this paper.

Given a training set of HR and LR image pairs \( \{I^H, I^L\} = \{I^h_i, I^l_i\}_{i=1}^N \), firstly PCA features are extracted to reduce computational costs. Next, CCA is used to project PCA features to a coherent feature space. In this feature space, the correlation between HR and LR features is maximum. This provides better condition for finding the mappings in the next step. Let \( \hat{X}^H \) and \( \hat{X}^L \) be the PCA features of HR and LR subtracted by the mean. Define \( C_{11} = E[\hat{X}^H (\hat{X}^H)^T] \) and \( C_{22} = E[\hat{X}^L (\hat{X}^L)^T] \) as the within-set covariance matrices, and \( C_{12} = E[\hat{X}^H (\hat{X}^L)^T] \) and \( C_{21} = E[\hat{X}^L (\hat{X}^H)^T] \) as the between-set covariance matrices, where \( E[\cdot] \) stands for mathematical expectation. Compute \( R_1 = C_{11}^{-1}C_{12}C_{22}^{-1}C_{21} \) and \( R_2 = C_{22}^{-1}C_{21}C_{11}^{-1}C_{12} \). The base matrices \( V^H \) comprises eigenvectors of \( R_1 \) and \( V^L \) comprises eigenvectors of \( R_2 \) when their corresponding eigenvalues are sorted in descending order. The coherent features of HR and LR images are

\[
C^H = (V^H)^T \hat{X}^H, \quad C^L = (V^L)^T \hat{X}^L. \tag{3.3}
\]

Then RBFs are applied to approximate the mapping between HR and LR coherent features. The function approximation is represented as \( C^H = W \Phi \) where \( W \) is a weighting coefficient matrix and \( \Phi \) is a multiquadratic basis function (see [38] for details). As a result, the weight matrix \( W \) can be solved by \( W = C^H (\Phi + \tau I_{id})^{-1} \). \( \tau I_{id} \) is included because \( \Phi \) is not always invertible. \( \tau \) is a small positive value, such as \( 10^{-3} \), and \( I_{id} \) is the identity matrix.

In the testing stage, the coherent features of HR gallery images are first computed. For a LR probe image \( I^p \), we compute the coherent features \( c^p \) and then apply the learnt mapping to obtain the SR features of the probe image by

\[
c_{SR} = W \cdot [\varphi(\|c^l_1 - c^p\|) \ldots \varphi(\|c^l_N - c^p\|)]^T. \tag{3.4}
\]

where \( \varphi(\|c_i - c_j\|) = \sqrt{\|c_i - c_j\|^2 + 1} \). Finally, the above features are fed to the nearest neighbour classifier for recognition.

3.2.4 Experimental results

In this section, we present identification results of the selected face recognition algorithms at various image resolutions and evaluate the performance of SR methods.
3.2.4.1 Low-resolution face recognition performance

In this section we provide the recognition performance of PCA, LDA and LBP on face images at various resolutions. In our identification experiments, we use 2820 facial images from the FRGC v1.0 database [8]. Original high-quality face images are resized to $70 \times 60$, $56 \times 48$, $42 \times 36$, $28 \times 24$, $14 \times 12$ and $7 \times 6$ pixel resolutions using bicubic interpolation. Sample images are shown in Fig. 3.1a. 1228 images of 139 persons are selected as training images. For the remaining images, two images per person are used as gallery images (242 images) and the rest are used as probe images (1350 images). Distance measures employed in 1-nearest neighbour classifier for PCA, LDA and LBP are L1 norm, cosine angle and Chi square, respectively. We use uniform patterns and (8, 1) neighbourhood for the LBP approach where face images are divided into 49 regions ($7 \times 7$ grid); but for images at the resolutions $14 \times 12$ and $7 \times 6$, the images are divided into 16 and 4 regions. The identification rates are shown in Fig. 3.1b.

![Sample face images from the FRGC v1.0 database at various resolutions](image)

![Identification rates](image)

Fig. 3.1: (a) Sample face images from the FRGC v1.0 database at various resolutions, (b) identification rates

As Fig. 3.1b shows, LDA outperforms PCA and LBP at all image resolutions. Both PCA and LDA accuracies decrease sharply when the image resolution is lower than $14 \times 12$. On the other hand, LBP is sensitive to resolution differences: correct classification rate of LBP is only 12.8% for images at resolution $7 \times 6$ but the performance rises as the resolutions get higher and becomes stable when the resolution reaches $42 \times 36$.

Then we test the recognition performance on the Surveillance Cameras Face (SCface) Database [10]. The SCface database contains images from 130 sub-
objects taken by five surveillance cameras at three distances, namely 4.20 meters (distance1), 2.60 meters (distance2), and 1.00 meter (distance3). It also contains one frontal mug-shot image for each subject. We crop faces from the original images according to the eye coordinates (see Fig. 3.2).

Fig. 3.2: Sample face images from the SCface database captured at: (a) distance1, (b) distance2, (c) distance3 and (d) mug-shots.

In our recognition experiments, the frontal mug-shot images are used as gallery images. For PCA and LDA, we use images from four cameras for training at each distance and use the images from the remaining camera as probe. Using this leave-one-out methodology, we perform five recognition experiments at each distance (distance 1, 2, and 3) and provide the average correct classification rate. Identification rates are shown in Table 3.1. Additionally, we also utilize the FRGC v1.0 database to train PCA and LDA classifiers to test the generalization ability of these methods on the SCface database. The last three rows of Table 3.1 provide the identification rates of FRGC training. It should be noted that the LBP approach does not need a training set. Therefore LBP results for different training schemes are identical in Table 3.1.

Table 3.1: Identification rates [%] for PCA, LDA and LBP on the SCface database.

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Probe Set</th>
<th>PCA</th>
<th>LDA</th>
<th>LBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCface Dist1</td>
<td>SCface Dist1</td>
<td>12.3</td>
<td>24.6</td>
<td>8.3</td>
</tr>
<tr>
<td>SCface Dist2</td>
<td>SCface Dist2</td>
<td>18.0</td>
<td>33.5</td>
<td>21.7</td>
</tr>
<tr>
<td>SCface Dist3</td>
<td>SCface Dist3</td>
<td>12.3</td>
<td>24.5</td>
<td>19.1</td>
</tr>
<tr>
<td>FRGC</td>
<td>SCface Dist1</td>
<td>9.9</td>
<td>10.3</td>
<td>8.3</td>
</tr>
<tr>
<td>FRGC</td>
<td>SCface Dist2</td>
<td>18.0</td>
<td>18.0</td>
<td>21.7</td>
</tr>
<tr>
<td>FRGC</td>
<td>SCface Dist3</td>
<td>11.1</td>
<td>10.7</td>
<td>19.1</td>
</tr>
</tbody>
</table>

The identification rates of PCA and LDA using FRGC training set are at most 18.0%. Though images captured at distance3 have a higher resolution than those of the other two, their results are worse than those of distance2.
as a consequence of the pose variation problem: because the subjects get close to the cameras, the images contain mostly the top part of the faces. Moreover, using SCface training sets instead of FRGC training set provides better results and it benefits LDA more than PCA. LBP is more sensitive to resolution changes. The LBP result is worse than that from both PCA and LDA at distance1. However, at distance2 and distance3, LBP outperforms LDA using FRGC training and PCA using both training configurations.

### 3.2.4.2 Super-resolution face recognition performance

Under real surveillance situations, HR training and gallery images are usually pre-stored and LR probe images are captured later. To simulate this situation, our experiments are conducted using the images from the FRGC database at resolution $70 \times 60$ as training and gallery, and images at lower resolutions as probe sets. The probe images are resized with bicubic interpolation to resolution $70 \times 60$. As shown in Fig. 3.3, the identification rates of images at resolutions $7 \times 6$ and $14 \times 12$ are worse than those of images at higher resolutions for all methods. The accuracies do not vary a lot when image resolutions are higher than $28 \times 24$. Thus, we will apply SR methods on images at resolution $7 \times 6$ and $14 \times 12$ to see their contribution in the following experiments. The LR images chosen from SCface are images captured at distance1.

![Fig. 3.3: Identification rates for various resolutions with bicubic interpolation on the FRGC database](image)

In SR experiments, we have two training phases: one is to train the SR method (e.g., learning the LR-HR mapping) and the other one is to train the face
classifier (e.g., PCA training). In FRGC experiment, the FRGC training set, which contains 1228 images, is used for both the two training phases. Images at a resolution of $70 \times 60$ are used as HR. LR images have resolutions $7 \times 6$ and $14 \times 12$. To train the SR system, both HR and LR images are needed, while only HR images are used to train the classifiers. In addition to comparing the performance using SR images and original HR images, we employ basic bicubic interpolation as a baseline SR method. Comparative identification rates obtained by both SR method are shown in Fig. 3.4.

![Fig. 3.4: Comparison of DSR method with NMCF method on the FRGC database: (a) LR $7 \times 6$, (b) LR $14 \times 12$.](image)

DSR method improves the correct classification rates for all face recognition methods at both $7 \times 6$ and $14 \times 12$ pixel resolutions: it consistently achieves higher identification rates than the bicubic interpolation method. It is also seen that the relative gain of using a SR method is more visible at lower resolutions. For instance, with LDA classifier, relative performance increase is higher at the $7 \times 6$ pixel resolution than $14 \times 12$. At both resolutions, LDA outperforms the other methods. The second SR method, NMCF approach, also attains better identification rates than the LR results but they are not as good as DSR approach when PCA and LDA are used as classifiers.

For the experiment using the SCface database (Fig. 3.5), images from FRGC are used to train the classifiers. Frontal mug-shots are gallery images and images captured at distance1 are used as probe images. DSR method requires that the number of training images must be larger than the number of pixels in the LR image. Since images captured at one distance are not enough, we use images captured at both distance2 and distance3 to train the SR system. The HR training images are resized to the same resolution as images at distance2, and the LR training images are downsampled from the HR images to the same
Fig. 3.5: Comparison of DSR method with NMCF method on the SCface database.

resolution as images at distance 1.

In the SCface database experiments, PCA outperforms LDA and LBP. DSR method can improve the results of PCA and LDA but not for LBP. NMCF approach provides better results than DSR methods with LBP but worse than the others.

In addition, some SR images constructed by the DSR method are shown in Fig. 3.6. The SR images of FRGC have much better quality than the SR images of SCface. The SR images of SCface have much more artifacts that lead to the poor quality of local features, thus, reduce the accuracy of LBP.

As a summary, we provide a comparison of three main schemes described above on the FRGC database (see Table 2): 1) matching LR probe to LR (downsampled) gallery, 2) matching upsampled (using bicubic interpolation) probe to HR gallery image and 3) matching SR probe to HR gallery image. The classifiers are trained with LR images in the first scheme while they are trained with HR images in the last two schemes. The upsampling scheme
provides the worst results for both LR resolutions. The SR scheme provides better results than the LR scheme when PCA and LBP are used. However, all the SR results are not as good as the LDA results of the first scheme.

<table>
<thead>
<tr>
<th>LR resolution</th>
<th>Probe</th>
<th>Gallery</th>
<th>PCA</th>
<th>LDA</th>
<th>LBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 × 6</td>
<td>LR</td>
<td>LR</td>
<td>38.3</td>
<td>75.1</td>
<td>12.8</td>
</tr>
<tr>
<td>7 × 6</td>
<td>Bicubic</td>
<td>HR</td>
<td>9.6</td>
<td>22.8</td>
<td>6.8</td>
</tr>
<tr>
<td>7 × 6</td>
<td>DSR</td>
<td>HR</td>
<td>38.6</td>
<td>56.4</td>
<td>20.9</td>
</tr>
<tr>
<td>14 × 12</td>
<td>LR</td>
<td>LR</td>
<td>59.9</td>
<td>90.5</td>
<td>39.5</td>
</tr>
<tr>
<td>14 × 12</td>
<td>Bicubic</td>
<td>HR</td>
<td>36.1</td>
<td>81.2</td>
<td>31.0</td>
</tr>
<tr>
<td>14 × 12</td>
<td>DSR</td>
<td>HR</td>
<td>62.4</td>
<td>85.6</td>
<td>54.2</td>
</tr>
</tbody>
</table>

### 3.2.5 Conclusion

In this paper, the performance of several face recognition algorithms, namely PCA, LDA and LBP, is evaluated for low-resolution face images. Our results show that LDA outperforms the others when down-sampled images are used. LBP is found to be not suitable for very low resolution images. The overall recognition accuracy is improved if super-resolution methods are applied. DSR method, as a reconstruction-based super-resolution approach, can be easily paired with standard face recognition algorithms and outperforms NMCF approach. However, the improvement of super-resolution methods on surveillance images is limited. Finally, using downsampled gallery and training images instead of using super-resolved probe images obtains better results for LDA classifier.
3.3 Comparison of super-resolution benefits for downscaled images and real low-resolution data

3.3.1 Abstract

Recently, more and more researchers are exploring the benefits of super-resolution methods on low-resolution face recognition. However, often results presented are obtained on downscaled high-resolution face images. Because downscaled images are different from real images taken at low resolution, it is important to include real surveillance data. In this paper, we investigate the difference between downscaled images and real surveillance data in two aspects: (1) the influence of resolution on face recognition accuracy, and (2) the improvement of accuracy that can be achieved by super-resolution on these images. Specifically, we will test the following hypotheses: (1) face recognition performance on real images is much worse than on downscaled images, and (2) super-resolution improves the performance of downscaled images more than real images.

Our experiments are conducted using videos from the HumanID database. In each video, the target person’s face is captured while he is walking towards the surveillance camera. We detect the faces in the video frames using a Viola-Jones face detector. Then we select face images of four different resolutions: two low-resolution and two high-resolution. The high-resolution images are used for gallery and generating downscaled images. We perform two types of face recognition experiments. In the first type of experiments, three face recognition methods are evaluated for images with different resolution. The three methods are (1) Principal Component Analysis, (2) Linear Discriminant Analysis, and (3) Local Binary Patterns. In the second type, we apply two super-resolution methods: (1) a model based method and (2) a feature based method on the low-resolution (both real and downscaled) images and then compute the face recognition accuracy.

3.3 Comparison of super-resolution benefits for downsampled images and real low-resolution data

3.3.2 Introduction

Face recognition at a distance is a challenging topic in the face recognition domain. The face images captured at a distance usually are small and of low quality, thus, they are not suitable for most of the face recognition systems which are developed for high quality images. To improve the face recognition performance of these low quality images, one way is to use super-resolution (SR) to enhance the resolution. Some SR methods have been developed specially for face recognition.

In [30], a model for SR is built based on Tikhonov regularization and a linear feature extraction stage. It includes the face features which would be extracted for face recognition as a prior information. In [38], canonical correlation analysis is used to project high-resolution (HR) and low-resolution (LR) image pairs to a coherent feature space and then radial based functions is applied to find the mapping between them. A data constraint is developed in [13] to minimize both the distances between the constructed SR images and the corresponding HR images and the distances between SR images from the same class. In [35], multidimensional scaling is used to transform the HR gallery and LR probe images to a common space so that the distance between them approximate the distance between their corresponding HR images. Another SR method based on morphable model is proposed in [32].

The LR images used in most SR papers are downsampled from HR images. Recently, a few researchers also use data from real surveillance cameras. In [13] and [35], the proposed methods are also evaluated using images from the SCface database [10] which contains surveillance quality facial images captured at three different distances. In [50], videos are captured and enhanced with SR methods for face recognition. A real-world outdoor video dataset captured by a PTZ camera is used in [54].

In this paper, we investigate the difference between downsampled and real LR images for face recognition. The experiments are conducted using images from the HumanID database [11]. This database contains videos in which the target person’s face is captured while he/she is walking towards the camera. We select images with four different resolutions from the videos and we also generate LR images by downsampling. Then we conduct both standard face recognition and super-resolution experiments. Three face recognition algorithms (PCA [51], LDA [52] and LBP [53]) and two SR methods (RL/DSR method [13] and NMCF method [38]) are investigated.

The remainder of this paper is organized as follows. In Section 3.3.3, hypothe-
ses are proposed regarding the difference between downsampled and real LR images. Section 3.3.4 gives a brief introduction to the two SR methods that are used in our experiments. Experimental setup and verification test results are presented in Section 3.3.5. Section 3.3.6 concludes the paper.

3.3.3 Downsampled vs. real low-resolution image

We expect downsampled images have better performance for face recognition than real LR images intuitively. One of the main reasons is that the registration is much poorer for real LR images. Finding the landmarks on HR faces is relatively easy. Downsampled images can make use of these landmarks for registration. But for LR images, the landmarks usually cannot be located precisely. Thus, we make following hypotheses:

(1) Face recognition performance on real LR images is much worse than on downsampled images;

(2) Super-resolution improves the performance of downsampled images more than real LR images.

An additional hypothesis is made for the training approach. Three different training configurations are possible. In the first configuration, the images from the training set are downsampled to the same resolution as the probe images. The second one upsamples the images of the probe set to the same resolution as the training images. In the third configuration, images from the probe set are also upsampled, but the training images are first downsampled to the same resolution as the probe images and then upsampled to the original resolution. We would expect the first and third approaches to perform better than the second one.

3.3.4 Super-resolution methods

3.3.4.1 RL and DSR method

In [13], Zou and Yuen proposed a SR method for very low-resolution images which is compatible with various face recognition methods. They introduce a data constraint which clusters the constructed SR images with the corresponding HR images.
Given a set of HR and LR image pairs \(\{\{I^h_i, I^l_i\}_{i=1}^N\}\), the relation \(R\) is modeled as

\[
R = \arg\min_{R'} \sum_{i=1}^N \left\| I^h_i - R'I^l_i \right\|^2.
\]  

(3.5)

This is called relationship learning (RL). It minimize the distance between HR and the space of LR projected by \(R\).

To acquire better results in face recognition, identity information about the subject is used based on RL. This is called discriminative super-resolution (DSR). A second term is added to (3.5), see (3.6).

\[
R = \arg\min_{R'} \frac{1}{N} \sum_{i=1}^N \left\| I^h_i - R'I^l_i \right\|^2 + \gamma d(R').
\]  

(3.6)

where \(\gamma\) is a constant to balance the two terms. We set \(\gamma\) to 1 in our experiments. \(d(R')\) is represented as

\[
d(R') = \frac{1}{N(\lambda_i = \lambda_j)} \sum_{\lambda_i=\lambda_j} \left\| I^h_i - R'I^l_j \right\|^2 - \frac{1}{N(\lambda_i \neq \lambda_j)} \sum_{\lambda_i \neq \lambda_j} \left\| I^h_i - R'I^l_j \right\|^2
\]  

(3.7)

where \(\lambda_i\) is the class label of \(I_i\). This makes sure the reconstructed HR images are clustered with the images from the same class and far away from those from other classes.

In the testing phase, for both RL and DSR, we first apply \(I_{SR} = RI_{input}\) to a given LR image \(I_{input}\) to obtain SR image \(I_{SR}\), and then use \(I_{SR}\) for face recognition.

### 3.3.4.2 NMCF method

In [38], Huang and He proposed a SR method where canonical correlation analysis (CCA) is used to project the PCA features of HR and LR image pairs to a coherent feature space. Radial based functions (RBFs) are then applied to find the mapping between the HR/LR pairs. This method finds nonlinear mappings on coherent features. Thus, we will refer to this method as NMCF method in this paper.

In the training stage, firstly PCA features of HR and LR image pairs are extracted to reduce computational costs. Next, these PCA features are projected to a coherent feature space by CCA where the correlation between HR
and LR features is maximum. This provides better condition for finding the mappings in the next step. Let \( \hat{X}_H \) and \( \hat{X}_L \) be the PCA features of HR and LR subtracted by the mean, and \( C_{11} \) and \( C_{22} \) be the within-set covariance matrices, and \( C_{12} \) and \( C_{21} \) be the between-set covariance matrices. Compute \( R_1 = C_{11}^{-1}C_{12}C_{21}^{-1}C_{22}^{-1}C_{12} \) and \( R_2 = C_{22}^{-1}C_{21}C_{11}^{-1}C_{12} \). The coherent features of HR and LR images are

\[
C^H = (V^H)^T \hat{X}_H, \quad C^L = (V^L)^T \hat{X}_L.
\]  

(3.8)

\( V^H \) and \( V^L \) comprises eigenvectors of \( R_1 \) and \( R_2 \) when their corresponding eigenvalues are sorted in descending order.

Then RBFs are applied to approximate the mapping between HR and LR coherent features. The function approximation is represented as \( C^H = W \Phi \) where \( W \) is a weighting coefficient matrix and \( \Phi \) is a multiquadratic basis function (see [38] for details). Thus, \( W \) can be solved by \( W = C^H (\Phi + \tau I_{id})^{-1} \). \( \tau \) is a small positive value and \( I_{id} \) is the identity matrix.

In the testing stage, for a given LR probe image \( I_p \), first compute the coherent features \( c_p \) and then apply the learnt mapping to obtain the SR features by

\[
c_{SR} = W \cdot [\varphi(||c_1 - c_p||) \ldots \varphi(||c_N - c_p||)]^T. \tag{3.9}
\]

where \( \varphi(||c_i - c_j||) = \sqrt{||c_i - c_j||^2 + 1} \). The SR features are fed to the nearest neighbor classifier together with the coherent features of HR gallery images for recognition.

### 3.3.5 Experimental results

Our experiments are conducted on the HumanID database [11]. We use the videos in which the target person’s face is captured while he is walking towards the surveillance camera, see Fig. 3.7. The faces from the videos are detected by the Viola-Jones face detector [43]. For each video, we select four face images with different resolution: 70×70, 50×50, 30×30 and 25×25. Thus, we have about 400 images for each resolution. The images with resolution 70×70 are used as gallery images. The images with the remaining three resolutions are used as probe sets. We also generate images with different resolutions by downsampling images of resolution 50×50. Images with six different resolutions are generated: 30×30, 25×25, 20×20, 15×15, 10×10 and 5×5. To train the face recognition classifier and the super-resolution system, we use down-
3.3 Comparison of super-resolution benefits for downsampled images and real low-resolution data

We use the eye coordinates for registration of the face images. The FRGC database provides eye coordinates of its images while we manually click on the eyes for the images from the HumanID database.

Fig. 3.7: Sample image frames from the HumanID database.

3.3.5.1 Face recognition experiments

In this section we provide the face recognition performance of PCA, LDA and LBP on both downsampled and real LR face images. Distance measures employed in 1-nearest neighbor classifier for PCA, LDA and LBP are L1 norm, cosine angle and Chi square, respectively. Three different training configurations are applied for our face recognition experiments:

Config. 1: the images of the training set are downsampled to the same resolution of the probe images;

Config. 2: the training images are with resolution 70×70, and the probe images are upsampled to resolution 70×70 using bicubic interpolation.

Config. 3: the images of the training set are first downsampled to the same resolution of the probe images, then upsampled back to the resolution of 70×70 using bicubic interpolation. The probe images are also upsampled to the resolution of 70×70.

Config. 2 and 3 are also the baseline configurations for SR. Because LBP does not require a training set, we only apply Config. 1 and 2 for LBP.

We conducted verification experiments. The ROC curves of PCA is shown in Fig. 3.8. We also compute the genuine acceptance rates (GAR) when false
Fig. 3.8: ROC curves of PCA results, downsampled vs. real: (a) downsampled, Config. 1; (b) downsampled, Config. 2; (c) downsampled, Config. 3; (d) real, Config. 1; (e) real, Config. 2; (f) real, Config. 3.

Fig. 3.9: Face recognition results, GAR@FAR=0.1: (a) PCA, (b) LDA, (c) LBP.

acceptance rates (FAR) equal to 0.1 for PCA, LDA and LBP, see Fig. 3.9.

As we can see, resolution changes do not have much influence for downsampled images, but for real LR images, the recognition results decrease significantly as the resolution becomes lower. At resolution 30×30 and 25×25, the GARs of downsampled images are almost the same as for resolution 50×50, but the GARs of real LR images are much lower. The performances of different face recognition classifiers are of similar trend while LBP is more sensitive
3.3 Comparison of super-resolution benefits for downsampled images and real low-resolution data

...to resolution changes than PCA and LDA. Different training configurations influence little when the image resolution is higher than $10 \times 10$.

3.3.5.2 Super-resolution experiments

In this section, we apply the SR methods that were introduced in Section 3.3.4 to explore the benefits of SR. Images from the FRGC database are used to train the SR system. The training configurations are similar to Config. 2 and 3, but RL and DSR methods are used for upsampling LR images instead of bicubic interpolation. The NMCF method is only designed for Config. 2. The ROC curves of RL and DSR methods with PCA in Config. 2 are shown in Fig. 3.10. The GARs when FARs equal to 0.1 are presented in Fig. 3.11.

![ROC curves of SR results using RL/DSR methods with PCA, Config. 2: (a) dowsampled, RL; (b) dowsampled, DSR; (c) real, RL; (d) real, DSR.]

As we can see from these results, all of the SR methods tested in the experiments have no benefits for verification. LBP results drop dramatically after SR. For PCA and LDA, the RL and NMCF methods keep the verification performance at the same level of bicubic interpolation. But the DSR method makes the results worse. The two training configurations also give similar results. To better explain these results, we shown some reconstructed SR images in Fig. 3.12.
Chapter 3. Applying super-resolution for low-resolution face recognition

Fig. 3.11: Super-resolution RL/DSR/NMCF results, GAR@FAR=0.1: (a) PCA, Config. 2; (b) PCA, Config. 3; (c) LDA, Config. 2; (d) LDA, Config. 3; (e) LBP results; (f) NMCF method. (ds = downsample)

Fig. 3.12: Reconstructed SR images by RL method from (a) real LR images, resolution 50×50, 30×30 and 25×25; (b) downsampled images, resolution 30×30, 25×25 and 20×20; (c) downsampled images, resolution 15×15, 10×10 and 5×5.
Firstly, comparing the real LR images with downsampled images at the same resolution, it is easy to find that the downsampled images are clearer and contain more details than the real images. This explains why downsampled images have better performance in LR face recognition. Secondly, the SR images show more details about the faces but also add some artifacts on them. These artifacts are also a problem for face recognition and it may cause the failure of LBP for recognition. Thirdly, as the resolution become lower, the identity information in the images become less, so the SR images become more look like an average face other than the person himself.

### 3.3.6 Conclusion

We evaluated the difference between downsampled and real LR face images for both standard face recognition and super-resolution. Our results show that face recognition on downsampled images performs much better than real LR images. The face recognition accuracy hardly decreases when the resolution change is within a certain range. But for real LR images, the verification results drop significantly as the resolution become lower. Moreover, the single image super-resolution methods do not have benefits for face recognition in our experimental configurations. One possible solution is to apply SR method which can make use of the information of multiple images.

### 3.4 Discussion

In this chapter, we have addressed research question RQ.1(a) about super-resolution. Our experiments show that although super-resolution can increase the resolution of images and enhance the image quality visually, it provides limited improvement in face recognition performance on real low-resolution facial images. Therefore, we will search for other solutions and focus on direct low-resolution comparisons in our later work. Another research question that we have addressed is RQ.2 about the difference between down-sampled and real low-resolution images. In this chapter, we start to realize the difference between down-sampled and real low-resolution images. The performance of the three basic face recognition methods are much better on down-sampled images than on real low-resolution images. Super-resolution methods, which have been proved to be effective on down-sampled images, have no benefit for the recognition performance on real low-resolution images. We will dig
into more details about the differences between down-sampled and real low
resolution images in the following chapters.
Chapter 4

Low-resolution face recognition and the MixRes classifier

4.1 Introduction

In this chapter, we propose a novel face recognition method to cope with the problem of comparing low-resolution probe facial images with high-resolution gallery facial images. Two publications are included. The first one (Section 4.2) [55] was published in 3rd International Workshop on Biometrics and Forensics, where we proposed the method at the first time. Most of the content of the first paper are repeated in the second one. The second paper (Section 4.3) [56] was published in IET Biometrics. In this paper, we improved the introduction of the new method to be more understandable for the readers. In addition, we gave an overview of the three approaches for low-resolution to high-resolution comparison. We also discussed about the problems in conducting research on low-resolution face recognition, for example, lack of common protocol and using down-sampled images instead of real low-resolution images for testing. Therefore, the first paper (Section 4.2) can be skipped while reading.
4.2 Likelihood ratio based mixed resolution facial comparison

4.2.1 Abstract

In this paper, we propose a novel method for low-resolution face recognition. It is especially useful for a common situation in forensic search where faces of low resolution, e.g. on surveillance footage or in a crowd, must be compared to a high-resolution reference. This method is based on the likelihood ratio of a pair of mixed-resolution input images. The effectiveness of our method is tested on the SCface database which contains face images taken by surveillance cameras. The results show that our method outperforms recently published state-of-the-art.

4.2.2 Introduction

Biometric face recognition systems based on high-resolution (HR) images have achieved great success. However, the problem of low-resolution (LR) face recognition remains challenging. LR face recognition is the case if at least the probe images are in LR. The face regions are small and mostly with different variations in realistic situations. This problem is relevant for forensic search, in particular the case where LR face in a crowd or on surveillance videos need to be recognised.

The most common case for LR face recognition is given gallery images of a list of suspects, to verify whether a person in the surveillance scene is in the list using a pre-trained classifier. Usually gallery images are HR and probe images are LR. However, most classifiers are designed for HR images and can only work properly for images with the same resolution. Three approaches for LR face recognition can be distinguished.

The first one is to down-sample the gallery HR images and compare them to the probe images in the LR space. Since the LR feature sets are smaller than HR feature sets, this approach requires less computational cost. However, there is a loss of high-frequency information in the down-sampling process. Little research has been done in this direction.

The second one is to acquire higher resolution probe images using super-resolution (SR) technique then conduct comparison in the HR space. The simplest SR approach is interpolation. It up-samples the LR images but does not bring in more information. Thus, the resulting images usually not only have little visual enhancement but also have worse recognition performance because the difference between the gallery and probe images becomes even bigger. More advanced SR methods have been investigated for decades. Baker and Kanade [28] proposed a method that learns a prior on the spatial distribution of images gradient for frontal face images. Then this prior is incorporated with MAP framework. Hennings-Yeomans et al. [57] built a model for SR based on Tikhonov regularization and a linear feature extraction stage. This model can be applied when images from training, gallery and probe sets have varying resolutions. The authors extended this method in [12] by adding a face prior to the model and using relative residuals as measures of fit. A data constraint was developed by Zou and Yuen [40] to minimize both the distances between the constructed SR images and the corresponding HR images and the distances between SR images from the same class. Zhang et al. [32] proposed a SR method in morphable model space which tries to construct HR information required by both reconstruction and recognition.

The third approach is to compare LR probe images to the HR gallery images directly. Most methods of this approach find mappings to project both LR and HR images to a common space so that direct comparison between them could be possible. This has been drawing the attention of researchers during the last years. Li et al. [34] proposed a method that project both HR and LR images to a unified feature space for classification using coupled mappings. The mappings are learnt by optimizing the objective function to minimize the difference between corresponding HR and LR images. Huang and He [38] proposed a method where canonical correlation analysis is used to project the PCA features of HR and LR image pairs to a coherent feature space. Radial based functions are then applied to find the mapping between the HR/LR pairs. This method finds nonlinear mappings on coherent features. A multidimensional scaling based method was proposed by Biswas et al [35]. Both HR and LR images are transformed to a common space where the distance between them approximates the distance when they both are HR. The transformations are learnt using an iterative majorization algorithm. Ren et al. [14] proposed a methods called coupled kernel embedding. It projects the original HR and LR images onto reproducible kernel space using coupled nonlinear functions. The dissimilarities captured by their kernel Gram matrices are minimized in this space. Lei et al. [2] proposed a coupled discriminant analysis method. They find coupled transformations to project HR and LR
images to a common space in which the low-dimensional embedding are well classified. The locality information in kernel space is also used as a constraint for the discriminant analysis process. Moutafis and Kakadiaris [15] proposed a method that learns semi-coupled mappings for HR and LR images for optimized representations. The mappings are aim to increase class-separation for HR images and map LR images to their corresponding class-separated HR data.

Some of the existing methods were demonstrated using real LR image databases [2, 14, 35, 40], while most methods were tested using probe images generated by down-sampling and smoothing HR images. However, down-sampled images differ from real LR images. We see that the face recognition performance on down-sampled images is much better than on real LR images from these publications. Besides, obtaining good results using down-sampled data does not necessarily mean the method also works on real LR images. In [48], it is shown that methods that can improve performance using down-sampled images may not be able to benefit face recognition using real LR images.

This paper follows the approach of [15] and focuses on LR to HR comparison. We propose a novel method for direct comparison of different resolution images which is called Mixed Resolution Biometric Comparison. We also demonstrate that this method outperforms the state-of-the-art methods for LR to HR comparison on real LR images.

The remaining parts of this paper are organized as follows: in Section 4.2.3, our proposed method, Mixed-Resolution Biometric Comparison (MRBC), is introduced. In Section 4.2.4, we demonstrate that our proposed method outperforms the state-of-the-art methods by conducting experiments on a surveillance face database. Section 4.2.5 concludes the paper.

### 4.2.3 Mixed-resolution Likelihood ratio based similarity score

Given two biometric feature vectors $\mathbf{x} \in \mathbb{R}^M$ and $\mathbf{y} \in \mathbb{R}^N$ obtained from multi resolution acquisition devices we look for support for the hypothesis $H_s$: the samples originate from the same individual versus $H_d$: the samples originate
from different individuals, quantified by the likelihood ratio

\[ l(x, y) = \frac{p\left( \begin{pmatrix} x \\ y \end{pmatrix} \mid H_s \right)}{p\left( \begin{pmatrix} x \\ y \end{pmatrix} \mid H_d \right)}. \quad (4.1) \]

It is well-known that an optimal classifier in Neyman-Pearson sense is obtained by thresholding the likelihood ratio, cf. for example [58].

We take \( M \geq N \), i.e. \( x \) is of higher resolution than \( y \), and assume that \( x = \mu_x + w_x \) and \( y = \mu_y + w_y \), with \( \mu_x = \mathbb{E}\{x\} \) and \( \mu_y = \mathbb{E}\{y\} \) the subject-specific class means of the features, and with \( w_x \) and \( w_y \) the statistically independent, zero-mean within-subject variations. Furthermore, we assume normal, zero-mean probability densities for \( w_x \) and \( w_y \), and for \( \mu_x \) and \( \mu_y \). If \( x \) and \( y \) are not zero-mean, estimated means have to be subtracted prior to comparison. Such a simple model cannot be expected to work well for HR face recognition, but when LR faces with fewer details are involved it still works. The matrices \( \Sigma_{xx} = \mathbb{E}\{xx^T\} \in \mathbb{R}^{M \times M} \) and \( \Sigma_{yy} = \mathbb{E}\{yy^T\} \in \mathbb{R}^{N \times N} \) are the covariance matrices of \( x \) and \( y \), respectively. The matrices \( \Sigma_{xy} = \mathbb{E}\{xy^T\} \in \mathbb{R}^{M \times N} \) and \( \Sigma_{yx} = \Sigma_{xy}^T \) are the cross-covariance matrices. Then \( \Sigma_{xy} = \mathbb{E}\{\mu_x\mu_y^T\} \). For the probability densities of the pairs of feature vectors we then have, respectively

\[ \left( \begin{array}{c} x \\ y \end{array} \right) \mid H_s \sim \mathcal{N}\left( \begin{array}{c} \Sigma_{xx} \\ \Sigma_{yx} \end{array} \begin{array}{c} \Sigma_{xy} \\ \Sigma_{yy} \end{array} \right), \quad (4.2) \]

\[ \left( \begin{array}{c} x \\ y \end{array} \right) \mid H_d \sim \mathcal{N}\left( \begin{array}{c} \Sigma_{xx} \\ \Sigma_{yx} \end{array} \begin{array}{c} 0 \\ \Sigma_{yy} \end{array} \right). \quad (4.3) \]

Covariance and cross-covariance matrices need to be estimated in a training process. The cross-covariance matrix \( \Sigma_{xy} \) is estimated as \( \hat{\Sigma}_{xy} = \frac{1}{K} \sum_{i=1}^{K} \hat{\mu}_{x,i} \hat{\mu}_{y,i}^T \), with \( K \) the number of individuals involved in training and \( \hat{\mu}_{x,i} \) and \( \hat{\mu}_{y,i} \) the estimated sample means of subject \( i \). The rank of \( \hat{\Sigma}_{xy} \) can be at most \( \min(N, K - 1) \). The \(-1\) is included because the sample means are zero-mean.

Under the above assumptions the following similarity score can be derived from the likelihood ratio (4.1):

\[ s(x, y) = \left( x^T y^T \right) \quad \left( \begin{array}{c} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{array} \right)^{-1} \left( \begin{array}{c} \Sigma_{xx} \\ \Sigma_{yx} \end{array} \begin{array}{c} \Sigma_{xy} \\ \Sigma_{yy} \end{array} \right)^{-1} \left( \begin{array}{c} x \\ y \end{array} \right). \quad (4.4) \]
46 Chapter 4. Low-resolution face recognition and the MixRes classifier

This score is optimal because it monotonically increases with the (log-likelihood ratio). In order to simplify (4.4) and to assure that the estimated covariance matrices have full rank and can be inverted we reduce the dimensionality and apply whitening transforms to \( \mathbf{x} \) and \( \mathbf{y} \), resulting in \( \mathbf{x}_w = \mathbf{W}_H \mathbf{x} \in \mathbb{R}^{M_w} \) and \( \mathbf{y}_w = \mathbf{W}_L \mathbf{y} \in \mathbb{R}^{N_w} \), respectively. Usually \( M_w < M \), \( N_w < N \), and \( M_w \geq N_w \). As a result we have that \( \Sigma_{xx}^w = \mathbb{E}\{\mathbf{x}_w \mathbf{x}_w^T\} = \mathbf{I} \) and \( \Sigma_{yy}^w = \mathbb{E}\{\mathbf{y}_w \mathbf{y}_w^T\} = \mathbf{I} \), with \( \mathbf{I} \) an identity matrix of appropriate size. The similarity score then becomes

\[
\begin{align*}
\mathbf{s}(\mathbf{x}_w, \mathbf{y}_w) &= \langle \mathbf{x}_w^T \mathbf{y}_w^T \rangle \\
&= \left( \begin{array}{cc} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{array} \right)^{-1} - \left( \begin{array}{cc} \mathbf{I} & \Sigma_{xy}^w \\ \Sigma_{yx}^w & \mathbf{I} \end{array} \right)^{-1} \left( \begin{array}{c} \mathbf{x}_w \\ \mathbf{y}_w \end{array} \right).
\end{align*}
\]

We will further simplify (4.5). First we apply a singular value decomposition to \( \Sigma_{xy}^w \), such that \( \Sigma_{xy}^w = \mathbf{U} \mathbf{D} \mathbf{V}^T \), with \( \mathbf{U} \in \mathbb{R}^{M_w \times M_w} \), and orthonormal, \( \mathbf{V} \in \mathbb{R}^{N_w \times N_w} \), and orthonormal, and \( \mathbf{D} \in \mathbb{R}^{M_w \times N_w} \). The first \( N_w \) rows of \( \mathbf{D} \) form a diagonal matrix consisting of singular values \( \nu_i, i = 1, \ldots, N_w \) in decreasing order. The last \( M_w - N_w \) rows of \( \mathbf{D} \) are an all-0 matrix. In a trained classifier the rank of \( \mathbf{D} \) can be at most \( D = \min(N_w, K-1) \), with \( K \) the number of individuals in the training set. If a smaller feature vector is desired, \( D \) can be chosen to be less than \( \min(N_w, K-1) \). We now transform the feature vectors again, such that \( \mathbf{x}_c = (\mathbf{U}_*, 1:D)^T \mathbf{x}_w \in \mathbb{R}^D \) and \( \mathbf{y}_c = (\mathbf{V}_*, 1:D)^T \mathbf{y}_w \in \mathbb{R}^D \), where the subscript *,1 : D denotes that only the first \( D \) columns of matrix are taken. The subscript \( c \) indicates that these transformations map the feature vectors to a common subspace. It can be shown that these transformations, which reduce the feature dimensionality to \( D \), will result in the same similarity score as transformations using the full matrices \( \mathbf{U} \) and \( \mathbf{V} \). For the similarity score we now have

\[
\begin{align*}
\mathbf{s}(\mathbf{x}_c, \mathbf{y}_c) &= \langle \mathbf{x}_c^T \mathbf{y}_c^T \rangle \\
&= \left( \begin{array}{cc} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{array} \right)^{-1} - \left( \begin{array}{cc} \mathbf{I} & \mathbf{D} \\ \mathbf{D} & \mathbf{I} \end{array} \right)^{-1} \left( \begin{array}{c} \mathbf{x}_c \\ \mathbf{y}_c \end{array} \right),
\end{align*}
\]

with \( \mathbf{D} \in \mathbb{R}^{D \times D} \) redefined as a diagonal matrix with the \( D \) largest singular values \( \nu_i \) of \( \Sigma_{xy}^w \) on the diagonal. By using

\[
\begin{align*}
\left( \begin{array}{c} \mathbf{x}_c \\ \mathbf{y}_c \end{array} \right) &= \frac{1}{2} \left( \mathbf{I} - \mathbf{I} \right)^T \mathbf{I} \left( \mathbf{I} - \mathbf{I} \right) \left( \begin{array}{c} \mathbf{x}_c \\ \mathbf{y}_c \end{array} \right) \\
&= \left( \begin{array}{c} \mathbf{x}_c \\ \mathbf{y}_c \end{array} \right) - \left( \begin{array}{c} \mathbf{x}_c \\ \mathbf{y}_c \end{array} \right) \mathbf{I} \mathbf{x}_c + \mathbf{I} \mathbf{y}_c,
\end{align*}
\]

\[
\begin{align*}
\mathbf{s}(\mathbf{x}_c, \mathbf{y}_c) &= \langle \mathbf{x}_c^T \mathbf{y}_c^T \rangle \\
&= \left( \begin{array}{cc} \mathbf{I} & \mathbf{D} \\ \mathbf{D} & \mathbf{I} \end{array} \right)^{-1} \left( \begin{array}{c} \mathbf{x}_c \\ \mathbf{y}_c \end{array} \right),
\end{align*}
\]
and expanding the matrix multiplications in (4.6) we can show that

\[
s(x_c, y_c) =
\]

\[- \sum_{i=1}^{D} \frac{\nu_i}{1 - \nu_i} (x_{c,i} - y_{c,i})^2 + \sum_{i=1}^{D} \frac{\nu_i}{1 + \nu_i} (x_{c,i} + y_{c,i})^2.
\]

In (4.8) a factor of 1/2 has been left out. A full expression for the log-likelihood ratio, that includes all the constants that have been ignored is

\[
\log(l(x_c, y_c)) = -\frac{1}{2} \sum_{i=1}^{D} \log \left(1 - \nu_i^2\right) + \frac{1}{4} s(x_c, y_c).
\]  

(4.9)

Because the \( \nu_i \) depend on training data, the use of this full expression is recommended in \( n \)-fold cross-validation experiments, since then the first term may differ slightly per validation step. Fig. 4.1 shows a block diagram of the classifier according to (4.8). The blocks perform matrix multiplications, except the rightmost ones, which compute a squared vector norm. The vectors \( \mathbf{x} \) and \( \mathbf{y} \) are the average HR and LR facial images, respectively. The matrices \( \Delta_{\text{DIF}} \) and \( \Delta_{\text{SUM}} \) are diagonal matrixes, defined by \( \Delta_{\text{DIF},ii} = \sqrt{\frac{\nu_i}{1 - \nu_i}}, \ i = 1, \ldots, D \) and \( \Delta_{\text{DIF},ii} = \sqrt{\frac{\nu_i}{1 + \nu_i}}, \ i = 1, \ldots, D \), respectively.

![Block diagram of the classifier according to (4.8).](image)

In a similar way a likelihood-ratio based classifier can be derived for heterogeneous features, e.g. for visual light and near infrared facial images, and for the case that feature sets of possibly different numbers if multiple captures must be compared.
4.2.4 Experiments

The experiments are conducted on the Surveillance Camera Face (SCface) database [10]. The SCface database contains images from 130 subjects taken by five surveillance cameras at three distances, namely 4.20 m (distance1), 2.60 m (distance2), and 1.00 m (distance3). There are $5 \times 130 = 650$ images for each distance. It also contains one frontal mug-shot image for each subject.

In order to compare our method MRBC to a state-of-the-art method presented in [15] called CBD, that protocol is used in our experiments. The region-of-interest is obtained by cropping the face regions of the images according to the eye coordinates provided in the database. Images from distance2 are used as LR and images from distance3 are HR. The resolution for HR and LR images after process are $30 \times 24$ and $15 \times 12$ respectively. Sample images are shown in Fig. 4.2. 100 subjects are randomly selected and 4 images from these subjects are also selected randomly for training. The remaining 30 subjects are used for testing. We randomly select 4 images per subject for gallery and the others are used as probe. Thus, we have 400 training images, 120 gallery images and 40 probe images each time. We have three types of experiments: LR vs. LR, HR vs. HR and HR vs. LR. The HR vs. LR indicates that gallery images are HR and probe images are LR. Each experiment is repeated 100 times.

The likelihood ratio of all experiments are calculated using Eq. (4.9). The parameters of MRBC method in all experiments are shown in Table 4.1.

In Table 4.2 and Fig. 4.3 we compare our experimental results with the results from CBD [15]. The best CBD results are chosen for comparison. The average

Fig. 4.2: Sampled images from SCface database in our experiments. First row: HR, second row: LR.
4.2 Likelihood ratio based mixed resolution facial comparison

<table>
<thead>
<tr>
<th>Setting</th>
<th>(K)</th>
<th>(M)</th>
<th>(N)</th>
<th>(M_w)</th>
<th>(N_w)</th>
<th>(D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR vs. LR</td>
<td>100</td>
<td>180</td>
<td>180</td>
<td>80</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>HR vs. HR</td>
<td>100</td>
<td>720</td>
<td>720</td>
<td>80</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>HR vs. LR</td>
<td>100</td>
<td>720</td>
<td>180</td>
<td>80</td>
<td>70</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 4.1: Parameters of MRBC in each experiment. The meaning of the parameters can be found in Section 4.2.3.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>AUC</th>
<th>Rank-1 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR vs. LR</td>
<td>CBD</td>
<td>0.78 (0.03)</td>
<td>57.17 (2.40)</td>
</tr>
<tr>
<td></td>
<td>MRBC</td>
<td>0.99 (0.01)</td>
<td>95.53 (3.68)</td>
</tr>
<tr>
<td>HR vs. HR</td>
<td>CBD</td>
<td>0.77 (0.03)</td>
<td>56.40 (9.50)</td>
</tr>
<tr>
<td></td>
<td>MRBC</td>
<td>1.00 (0.00)</td>
<td>99.13 (1.68)</td>
</tr>
<tr>
<td>HR vs. LR</td>
<td>CBD</td>
<td>0.77 (0.03)</td>
<td>52.67 (9.90)</td>
</tr>
<tr>
<td></td>
<td>MRBC</td>
<td>0.88 (0.03)</td>
<td>57.33 (9.52)</td>
</tr>
</tbody>
</table>

Table 4.2: Comparison of MRBC to CBD. The values are in the format: average value (standard deviation).

Fig. 4.3: ROC curves for comparing MRBC to CBD for HR vs. LR setting.

value and standard deviation of AUC and rank-1 identification rates are shown in Table 4.2. We also collect all the genuine and imposter scores from the 100
times to plot ROC curves. Since [15] only provided ROC for HR vs. LR setting, we compare our ROC curve with theirs in Fig. 4.3. The ROC curve of CBD is reproduced by selecting points on the original figure. Our ROC curves for all the three settings are also provided in Fig. 4.4. In all cases, our method significantly outperforms CBD.

![ROC curves using MRBC for all the three settings.](image)

The above experiments use multiple gallery images per subject. This is not very common in real cases. We repeated the experiments using single gallery image per subject, the results are shown in Table 4.3. Although there is some drop of performance in both verification and identification, MRBC still provides promising results.

<table>
<thead>
<tr>
<th>Setting</th>
<th>AUC</th>
<th>Verification %</th>
<th>Rank-1 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR vs. LR</td>
<td>0.96 (0.02)</td>
<td>89.65 (5.25)</td>
<td>83.33 (6.43)</td>
</tr>
<tr>
<td>HR vs. HR</td>
<td>0.98 (0.01)</td>
<td>94.94 (3.43)</td>
<td>94.27 (4.08)</td>
</tr>
<tr>
<td>HR vs. LR</td>
<td>0.84 (0.03)</td>
<td>57.42 (7.51)</td>
<td>47.90 (8.86)</td>
</tr>
</tbody>
</table>

Table 4.3: MRBC results using single gallery image per subject. The values are in the format: average value (standard deviation). The verification rates are obtained at false acceptance rate 0.1.

Our experiments followed the protocol by Moutafis and Kakadiaris [15] to compare with their method. However, this protocol has some limitations. Training data are taken at the same day of the testing data, it is also the same with gallery and probe images. Those images are taken in the same situation and with very similar illumination. Thus, the protocol is similar to a
within session comparison. The results would be worse if mug-shots are used as gallery images as in real-world scenarios.

4.2.5 Conclusion

In this paper we propose a novel method for mixed resolution biometric comparison, i.e. comparison of two facial images with different resolutions. The method is based on the likelihood ratio framework where in the derivation of the expression for the likelihood ratio, the combined statistics of the low and high resolution images is taken into account. The resulting method is, therefore, especially suitable for the case that gallery and probe are of different resolutions, but it can be extended for other heterogeneous features. The experiments on surveillance quality images demonstrate that this method significantly outperforms the state-of-the-art. We also remark that the protocols used in the comparison and that are used in other publications are limited in the fact that for training, gallery and probe images are used that are recorded within a very narrow time frame (usually 1 day) and constant conditions. This simplifies the task of facial recognition. A more realistic protocol should use images recorded at different times and with larger variations in conditions.

4.3 Low-resolution face alignment and recognition using mixed-resolution classifiers

4.3.1 Abstract

A very common case for law enforcement is recognition of suspects from a long distance or in a crowd. This is an important application for low-resolution face recognition (in our case, face region below 40 × 40 pixels in size). Normally, high-resolution images of the suspects are used as references, which will lead to a resolution mismatch of the target and reference images since the target images are usually taken at a long distance and are of low resolution. Most existing methods that are designed to match high-resolution images, cannot handle low-resolution probes well. In this paper, we propose a novel method especially designed to compare low-resolution images to high-resolution ones,

\[\text{The contents of this section are published in [56] “Low-Resolution Face Alignment and Recognition using Mixed-Resolution Classifiers”, in IET Biometrics, vol. 6, no. 6, pp. 418-428, November 2017}\]
which is based on the log-likelihood ratio. In addition, we demonstrate the difference in recognition performance between real low-resolution images and images down-sampled from high-resolution ones. Misalignment is one of the most important issues in low-resolution face recognition. Two approaches – matching-score based registration and extended training of images with various alignments – are introduced to handle the alignment problem. Our experiments on real low-resolution face databases show that our methods outperform the state-of-the-art.

4.3.2 Introduction

Biometric face recognition for high-resolution images has been highly successful. However, low-resolution face recognition, which refers to the case where at least the probe images are of low resolution (in our case, a face region below $40 \times 40$ pixels), is a challenging task because low-resolution face images contain less discriminative information than higher-resolution face images.

In low-resolution face recognition, a common surveillance task is, given a list of suspects, to try to verify whether a person in the surveillance scene is on this list (in face recognition known as gallery). Usually the gallery images are high-resolution frontal images of high quality. The probe images are taken at a distance and without user cooperation. We call this sort of images ‘real low-resolution images’ as opposed to low-resolution images obtained by down-sampling. The images are not only of low resolution but also have a higher noise level and deviate in other ways from the the high-resolution gallery images. However, most classifiers are designed to work properly for images of the same high resolution and cannot handle the resolution and quality mismatch. There are three approaches to deal with the resolution mismatch in low-resolution face recognition. The first one is to reconstruct higher-resolution probes using super-resolution techniques, and perform comparison to the gallery images in the high-resolution space [13,32]. The second approach is to down-sample the high-resolution galleries and compare them to the probes in the low-resolution space. Since the low-resolution images are smaller than the high-resolution ones, this approach involves lower computational costs. The third approach is to compare the low-resolution probe images to the high-resolution gallery images directly. Most methods following this approach find mappings to project both low-resolution and high-resolution images to a common space in which a direct comparison is performed.

It is hard to compare the existing methods directly due to the lack of a common
4.3 Low-resolution face alignment and recognition using mixed-resolution classifiers

protocol: everyone uses a different experimental setting. Another issue is that most researchers only present closed-set identification results such as the rank-1 recognition rate. These results depend highly on the size of the test sets, the number of subjects and on whether the classifier was trained on the galleries. Verification results are better suited for direct comparison and they directly address a relevant biometric question: “Are the two images of the same person?”

The evaluation of most existing methods for low-resolution face recognition is based on images that were aligned at high resolution and then smoothed and down-sampled to low resolution. However, we observe that in [2,14,35,40] the face recognition performance on down-sampled images is much better than on real low-resolution images. In this paper we will confirm these observations and conclude that, in order to produce realistic results, experiments should be based on real low-resolution images.

In this paper, instead of proposing a general face recognition method, we deal with problems in the specific scenario as mentioned above: comparison of low-resolution probes to high-resolution galleries. We extend our work in [55] concerning a method especially designed for comparing images captured at different distances, called Mixed-Resolution Biometric Comparison. This method not only works for different resolutions, but also captures other differences between images recorded at various distances. Since proper alignment proves to be crucial for a good recognition performance, and precise alignment (e.g. via accurately detected facial landmarks) is difficult for low-resolution images, we provide two methods to deal with the alignment problem: Matching-Score Based Registration and Extended Training with images with slightly varying alignment. We demonstrate that the combination of the three methods outperforms the state-of-the-art for real low-resolution face recognition. In addition, we address the differences in face recognition performance between down-sampled and real low-resolution images and pay special attention to proper evaluation protocols.

We conduct experiments of different settings using images of varying resolutions. First we duplicate the experimental protocol of a state-of-the-art method so that we can directly compare our methods to it. Then we set up a more realistic experiment to demonstrate that our methods are effective in realistic situations.

In the remainder of this paper, we use the terms HiRes, LoRes and SupRes for high-resolution, low-resolution and super-resolution.

This paper is organised as follows: Section 4.3.3 is a literature review of exist-
ing low-resolution face recognition methods. In Section 4.3.4, we explain why researchers should present results from real low-resolution images by demonstrating the difference between using down-sampled and real low-resolution probes. Our proposed methods – Mixed-Resolution Biometric Comparison, Matching-Score Based Registration and Extended Training are introduced in Section 4.3.5. In Section 4.3.6 we report results and Section 4.3.7 presents the conclusions.

### 4.3.3 Approaches to LoRes-HiRes comparison

Face recognition for low resolution is different from face recognition for high resolution. Firstly, it is much harder to detect landmarks reliably and accurately in LoRes images. In addition, LoRes images contain far less discriminative information than HiRes images. There is a small but growing body of literature that specifically addresses the problem of LoRes face recognition. Important references of LoRes face recognition are listed in Table 4.4 and Table 4.5. Table 4.4 lists papers that present experiments on down-sampled probe images and Table 4.5 lists papers using real LoRes probe images. We separate the papers in two tables because the evaluation of the two types of data is different. We will discuss the differences in Section 4.3.4. In the tables, we include the methods, the experimental settings and rank-1 recognition rates. As we can see, each paper presents experiments in a different setting even when they use the same database. For example, there are three papers which conducted experiments on the SCface database in Table 4.5, but they have different numbers of subjects for testing, different numbers of gallery and probe images per subject. This makes it impossible to compare the methods objectively.

As mentioned in Section 4.3.2, existing methods solve the resolution mismatch problem mainly by following the three approaches: applying SupRes to LoRes probes, down-sampling HiRes gallery images and direct LoRes–HiRes comparison. We will discuss the existing methods based on the three approaches.

#### 4.3.3.1 SupRes versus HiRes comparison

Since most face recognition systems are designed for HiRes images, many researchers reconstruct HiRes versions from the LoRes probes using SupRes techniques to make use of the information contained in the HiRes gallery images and conduct comparison in the HiRes space.
4.3 Low-resolution face alignment and recognition using mixed-resolution classifiers


<table>
<thead>
<tr>
<th>Method</th>
<th>Database</th>
<th>( N_C )</th>
<th>( N_G )</th>
<th>( N_P )</th>
<th>IMsize</th>
<th>Rank-1</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKE [14]</td>
<td>Multi-PIE</td>
<td>229</td>
<td>7</td>
<td>13</td>
<td>6 × 6</td>
<td>88%</td>
<td>L–H</td>
</tr>
<tr>
<td>MDS [35]</td>
<td>Multi-PIE</td>
<td>237</td>
<td>1</td>
<td>1</td>
<td>12 × 10</td>
<td>81%</td>
<td>L–H</td>
</tr>
<tr>
<td>DTCWT [31]</td>
<td>Multi-PIE</td>
<td>202</td>
<td>1</td>
<td>20</td>
<td>20 × 20</td>
<td>99%</td>
<td>S–H</td>
</tr>
<tr>
<td>SDA [36]</td>
<td>Multi-PIE</td>
<td>149</td>
<td>1</td>
<td>10+</td>
<td>12 × 12</td>
<td>70%</td>
<td>L–H</td>
</tr>
<tr>
<td>S2R2 [12]</td>
<td>Multi-PIE</td>
<td>224</td>
<td>1</td>
<td>13</td>
<td>6 × 6</td>
<td>73%</td>
<td>S–H</td>
</tr>
<tr>
<td>MFF [37]</td>
<td>FERET</td>
<td>200</td>
<td>1</td>
<td>1</td>
<td>12 × 12</td>
<td>84%</td>
<td>S–H</td>
</tr>
<tr>
<td>NMCF [38]</td>
<td>FERET</td>
<td>1195</td>
<td>1</td>
<td>1</td>
<td>12 × 12</td>
<td>84%</td>
<td>L–H</td>
</tr>
<tr>
<td>CLPM [34]</td>
<td>FERET</td>
<td>1195</td>
<td>1</td>
<td>1</td>
<td>12 × 12</td>
<td>90%</td>
<td>L–H</td>
</tr>
<tr>
<td>GDAMM [39]</td>
<td>AR</td>
<td>126</td>
<td>7</td>
<td>7</td>
<td>8 × 7</td>
<td>72%</td>
<td>S–H</td>
</tr>
<tr>
<td>MM [32]</td>
<td>CMU video</td>
<td>68</td>
<td>1</td>
<td>16</td>
<td>23 × 23</td>
<td>81%</td>
<td>S–H</td>
</tr>
<tr>
<td>EigenSR [3]</td>
<td>CMU video</td>
<td>68</td>
<td>1</td>
<td>16</td>
<td>10 × 10</td>
<td>74%</td>
<td>S–H</td>
</tr>
<tr>
<td>EigenTr [29]</td>
<td>XM2VTS</td>
<td>295</td>
<td>1</td>
<td>1</td>
<td>10 × 10</td>
<td>59%</td>
<td>S–H</td>
</tr>
<tr>
<td>DSR [40]</td>
<td>FRGC v2.0</td>
<td>311</td>
<td>8</td>
<td>2</td>
<td>7 × 6</td>
<td>78%</td>
<td>S–H</td>
</tr>
</tbody>
</table>

The simplest SupRes approach is interpolation. It up-samples the LoRes images but does not use additional information about the images, e.g. that they are faces. Thus, the resulting images usually have poor recognition performance because there is still a large difference between the gallery and probe images.

Face hallucination refers to SupRes techniques that were specially designed to improve face image quality. Baker and Kanade [28] proposed a method that learns a prior on the spatial distribution of image gradients for frontal face images. Then this prior is incorporated in the MAP framework. Wang and Tang [29] proposed a face hallucination method using eigen-transformation. The input LoRes image is represented as a linear combination of the LoRes images in the training set by Principal Component Analysis. The SupRes image is reconstructed using the corresponding HiRes training images with
Table 4.5: Papers using real LoRes data as probe. \( N_C \): number of subjects for testing. \( N_G \): number of images per subject in the gallery set. \( N_P \): number of images per subject in the probe set. L–H: LoRes–HiRes; S–H: SupRes–HiRes. As for the SCface database, the images from dist1, dist2 and dist3 are captured at a distance of 4.2 m, 2.6 m, 1.0 m.

<table>
<thead>
<tr>
<th>Method</th>
<th>Database</th>
<th>Gallery</th>
<th>Probe</th>
<th>( N_C )</th>
<th>( N_G )</th>
<th>( N_P )</th>
<th>Rank-1</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKE [14]</td>
<td>SCface</td>
<td>mugshot</td>
<td>dist1</td>
<td>130</td>
<td>1</td>
<td>5</td>
<td>8%</td>
<td>L–H</td>
</tr>
<tr>
<td>DSR [40]</td>
<td>SCface</td>
<td>dist2</td>
<td>dist1</td>
<td>130</td>
<td>5</td>
<td>5</td>
<td>22%</td>
<td>S–H</td>
</tr>
<tr>
<td>CBD [15]</td>
<td>SCface</td>
<td>dist3</td>
<td>dist2</td>
<td>100</td>
<td>4</td>
<td>1</td>
<td>53%</td>
<td>L–H</td>
</tr>
<tr>
<td>CDA [2]</td>
<td>local video</td>
<td>photo</td>
<td>video</td>
<td>161</td>
<td>5</td>
<td>5</td>
<td>53%</td>
<td>L–H</td>
</tr>
<tr>
<td>DTCWt [31]</td>
<td>local video</td>
<td>photo</td>
<td>video</td>
<td>34</td>
<td>1?</td>
<td>1</td>
<td>56%</td>
<td>S–H</td>
</tr>
</tbody>
</table>

the same coefficients.

Though the face hallucination methods can enhance visual image quality, the restored information may not contribute to better face recognition. Therefore, researchers started to work on SupRes methods that aim to improve face recognition performance. Gunturk et al. [3] proposed to apply SupRes in an eigen-domain that reconstructs only the necessary information for recognition. Hennings-Yeomans et al. [57] built a model for SupRes based on Tikhonov regularization and a linear feature extraction stage. This model can be applied when images from training, gallery and probe sets have varying resolutions. This approach is extended in [12] by adding a face prior to the model and using relative residuals as measures of fit. Zou and Yuen [13] developed a data constraint to minimize both the distances between the constructed SupRes images and the corresponding HiRes images as well as the distances between SupRes images from the same class. This method is extended in [40] with a linearity cluster. Bilgazyev et al. [31] proposed a method that uses dual-tree complex wavelet transform to extract high frequency components of training images. Then the SupRes features are represented as a weighted combination of the HiRes training images and the weights are the same as the ones that represent the input LoRes probe using corresponding LoRes training images. Zhang et al. [32] proposed a SupRes method in morphable model space, which provides HiRes information required by both reconstruction and recognition.

Though applying SupRes to LoRes probes makes the comparison to HiRes galleries possible in the HiRes space, the process of SupRes usually brings in artefacts or noise which might influence the face recognition performance.
4.3 Low-resolution face alignment and recognition using mixed-resolution classifiers

4.3.3.2 LoRes versus down-sampled HiRes comparison

Down-sampling HiRes gallery images and conducting comparison in the LoRes space is a very simple way for LoRes to HiRes comparison. It requires lower computational costs than using HiRes images. Although some information in the HiRes images will be lost in the down-sampling process, it has been reported by some researchers that this approach has similar recognition performance as SupRes methods for LoRes face recognition. For instance, Hu et al. [41] conducted experiments using a video database of moving faces and people. Their experimental results show that applying SupRes methods and then comparing to HiRes galleries have similar performance as LoRes to LoRes comparison at a far range (5-10 pixel eye-to-eye distance). Xu et al. [42] showed that when image resolution is low enough, LoRes to LoRes comparison is superior to using SupRes methods. In their experiments conducted on Yale B and AR databases with down-sampled images, SupRes methods perform much poorer than LoRes to LoRes comparison when the image size is 8 × 8 pixels. These results suggest that down-sampling gallery images and comparing to LoRes probes has at least as good face recognition performance as applying SupRes on LoRes probes and comparing to galleries in the HiRes domain.

4.3.3.3 LoRes versus HiRes comparison

Direct comparison of LoRes probes and HiRes galleries is a new area that has drawn researchers’ attention in recent years. Most methods of this approach find transformations for both LoRes and HiRes images and compare their features in a common space. This approach avoids losing information as a result of down-sampling HiRes or adding artefacts by SupRes. The mappings between HiRes gallery and LoRes probe data can also be learnt in such a way that different variations are modelled.

Li et al. [34] proposed a method that projects both HiRes galleries and LoRes probes to a unified feature space for classification using coupled mappings. The mappings are learnt by optimizing the objective function that minimizes the difference between corresponding HiRes and LoRes images. Huang and He [38] proposed a method that uses canonical correlation analysis to project the PCA features of HiRes and LoRes image pairs to a coherent feature space. Radial based functions are then applied to find the mapping between the HiRes and LoRes pairs. A multidimensional scaling based method is proposed by Biswas et al. [35]. Both HiRes and LoRes images are transformed to a common space where the distance between them approximates the distance
58 Chapter 4. Low-resolution face recognition and the MixRes classifier

when they are both HiRes. The transformations are learnt using an iterative majorization algorithm. Ren et al. [14] proposed a method called coupled kernel embedding. It projects the original HiRes and LoRes images onto reproducible kernel space using coupled nonlinear functions. The dissimilarities captured by their kernel Gram matrices are minimized in this space. Lei et al. [2] proposed a coupled discriminant analysis method. They find coupled transformations to project HiRes and LoRes images to a common space in which the low-dimensional embedding is well classified. The locality information in kernel space is also used as a constraint for the discriminant analysis process. This method is also suitable for images of different modalities, for example, visible and infrared faces. Moutafis and Kakadiaris [15] proposed a method that learns semi-coupled mappings for HiRes and LoRes images for optimized representations. The mappings aim at increasing class-separation for HiRes images and projecting LoRes images to their corresponding class-separated HiRes data.

4.3.4 Evaluation using down-sampled or real LoRes images

4.3.4.1 Problem statement

As we stated in the previous section, most papers test the recognition performance of their proposed methods on down-sampled probe images. Some researchers conducted experiments on real LoRes databases as well as on down-sampled probe images. However, we observe that face recognition methods perform much worse on real LoRes images than down-sampled images. In [40] for instance, the proposed method achieves 78% rank-1 recognition rate on images down-sampled to $7 \times 6$ pixels from the FRGC v2.0 database, while the result is only 22% on dist1 images (about $30 \times 30$ pixels) from the SCface database. Also in [14], the rank-1 recognition rate on down-sampled images of $6 \times 6$ pixels from the Multi-PIE databases is 88%, however it is only 6% on dist1 images from the SCface database. In [2], the rank-1 recognition rate of the proposed method is 97% on images down-sampled to $16 \times 16$ pixels from the Multi-PIE database, while it is only 52% on a self-collected database, where the image size is $35 \times 35$ pixels.

4.3.4.2 Analysis experiment

Our hypothesis is that there are differences between down-sampled images and real LoRes images that result in poorer face recognition results for the latter.
4.3 Low-resolution face alignment and recognition using mixed-resolution classifiers

To investigate this and exclude other influences, we set up a face recognition experiment where the down-sampled images and the real LoRes images are from the same source and pose, and there are no pose and illumination variations. Because there is no database available that meets this requirement, we collected a database ourselves.

We used a commercial camera CASIO EX-FC100. The images were recorded in the following way: the subject sat still and faced the camera. The first photograph of each subject was taken at a distance of 2 m, and then we moved the camera one metre away and took a photograph each time until we had nine photographs of this subject. Thus, the face images (probes) were taken at nine distances in total from 2 m to 10 m. Two weeks later, gallery images of each person were captured at a distance of 1 m with the same setup. The faces were always frontal and with the same illumination. Twenty-five subjects are included in our database. The original images are around 1 MB and of resolution 1600 × 1200. The image file format is JPEG. During pre-processing, all images were aligned using manually annotated eye-coordinates. The size of cropped face regions is 243 × 243 pixels for the galleries, and 131 × 131, 87 × 87, 64 × 64, 51 × 51, 44 × 44, 36 × 36, 33 × 33, 28 × 28 and 23 × 23 pixels for distance 2 m to 10 m, respectively. An elliptic mask is applied to select the region of interest. Histogram equalization is used to normalize the illumination. Unfortunately we are not able to make this database available, but we are willing to evaluate algorithms on request. Sample images are shown in Fig. 4.5.

![Sample images from our own database.](image)

To test face recognition performance on our database, we use three different probe sets: the first one contains images taken at distances from 2 m to 10 m (Real); the second one contains images taken at 2 m down-sampled to the same
resolution as the first probe set and then aligned using the eye-coordinates annotated at the highest resolution (Downsample); the third one also contains down-sampled images, but they are aligned using eye-coordinates annotated after the down-sampling process (Align after ds). The type of images in the second set, which are pre-aligned before down-sampling, is commonly used for evaluation of LoRes face recognition methods. From now on, we shall refer to pre-aligned down-sampled images when we mention down-sampled images. The gallery images are down-sampled to the same resolution as the probe images for comparison. To train the face classifiers, 4471 high-quality images of 275 subjects from the FRGC database [8] are used. We use four face recognition methods. One of them is the state-of-the-art LoRes face recognition method CLPM [34] (detailed explanation of CLPM is in Section 4.3.6). The other three methods are classical face classifiers, namely Principal Component Analysis (PCA) [51], Linear Discriminant Analysis Log Likelihood Ratio (LDALLR) [59], and Local Binary Patterns (LBP) [53]. The LDALLR is based on the LDA classifier [52] but computes log-likelihood ratio similarity scores instead of distance measurements. Distance measures employed for PCA and LBP are L1 norm and Chi square, respectively. In order to obtain overall good performance, 80 PCA and 50 LDA vectors are chosen. The images are divided into $6 \times 6$ regions for LBP. The feature dimension selected for the CLPM method is 170 (the authors of [34] chose 80, but 170 gives better performance).

We present verification results because, as argued in Section 4.3.2, they allow better comparison and address a more realistic biometric question than rank-1 recognition rates. It must be remarked in advance that the verification performance for low-resolution images is poorer than for high-resolution images. This is also illustrated in [55], where the performance of the state-of-the-art low-resolution face recognition is presented. Besides, LoRes face recognition is commonly used for surveillance, by which good verification performance is more important than limiting False Acceptance Rate (FAR). For the above reasons, verification rates (also known as genuine acceptance rate) are presented at FAR 10% rather than the more common FAR equals to 1% or 0.1%. See Fig. 4.6.

As we can see in the figure, the performance of all the four classifiers has a similar trend. When the images are pre-aligned and then down-sampled, the face recognition performance remains stable for every resolution. The performance on both the real LoRes images and the post-aligned down-sampled images becomes worse when the distances increase from 2 m to 10 m. Although the post-aligned down-sampled images generally have better results than the real
Fig. 4.6: Verification rate (VR) at FAR 10% on images from various distances using different classifiers: (a) PCA (b) LDALLR (c) LBP (d) CLPM.

LoRes images, the differences are much smaller than when compared to the results from the pre-aligned images. This trend demonstrates that face recognition performance on the pre-aligned down-sampled images is different from the performance on real LoRes images and alignment plays a very important role.

There is also different behaviour of different face recognition methods on images from different distances. The CLPM method, as it was designed to operate optimally for LoRes images, could not handle images from 3 m and 4 m and it performs stably for the rest of the resolutions. We have a memory problem for the 2 m images from the available code from the authors of this method, but it does not affect the trend. The other three methods all perform worse on lower resolution images than higher resolution ones, while LBP is more sensitive to resolution changes than PCA and LDALLR. In addition, LDALLR has the best performance among the four face recognition methods.
4.3.4.3 Conclusion

In this section we have demonstrated that down-sampled images are not fully representative of realistic low-resolution images and hence should not be used as probes when testing the effectiveness of LoRes face recognition methods. Face recognition methods perform much better on down-sampled images than on real LoRes images. The fact that down-sampled images are usually pre-aligned at HiRes while real LoRes images are aligned at LoRes plays an important role. Other factors like viewing angle, Bayer mask, noise level and compression artefacts may influence the recognition performance as well. We identify alignment as one of the most important factors. Besides, in real LoRes face recognition applications, pose, illumination and facial expression may differ significantly between gallery and probe, though we do not deal with them in this paper. Thus, real LoRes data should be used for the evaluation of LoRes face recognition methods to ensure their feasibility in reality.

4.3.5 Proposed methods

4.3.5.1 Mixed-resolution Biometrics Comparison

Here we present a method especially designed for comparing images captured at different distances, called Mixed-resolution Biometric Comparison. This method not only works for different resolutions, but also learns variations in the image quality. It was first derived in [55], but is repeated here for completeness. A similar method for homogeneous cases was proposed in [60], but we consider heterogeneous cases where the gallery and probe are from different scenarios.

Given two biometric feature vectors $x \in \mathbb{R}^M$ and $y \in \mathbb{R}^N$ obtained from multi-resolution acquisition devices we look for support for the hypothesis $H_s$: the samples originate from the same individual versus $H_d$: the samples originate from different individuals, quantified by the likelihood ratio

$$l(x, y) = \frac{p\left(\left(\begin{array}{c} x \\ y \end{array}\right) | H_s\right)}{p\left(\left(\begin{array}{c} x \\ y \end{array}\right) | H_d\right)}.$$  \hspace{1cm} (4.10)

It is well-known that an optimal classifier in the Neyman-Pearson sense is obtained by thresholding the likelihood ratio, cf. for example [58]. This means
the likelihood ratio will give the highest verification rate at a given FAR.

Note that \( \mathbf{x} \), with dimension \( M \), is a realization of a feature vector of a random individual, characterised by its feature mean, which is therefore also random. Similarly, \( \mathbf{y} \) is also a realization of a feature vector of a random individual, but with dimension \( N \). We take \( M \geq N \), i.e. \( \mathbf{x} \) is of higher resolution than \( \mathbf{y} \). Let \( \omega \) and \( \theta \) denote the identity of \( \mathbf{x} \) and \( \mathbf{y} \), we assume that \( \mathbf{x} = \mu_\omega + \mathbf{w}_\omega \) and \( \mathbf{y} = \mu'_\theta + \mathbf{w}'_\theta \), with \( \mu_\omega = \mathbb{E}\{\mathbf{x}|\omega\} \in \mathbb{R}^M \) and \( \mu'_\theta = \mathbb{E}\{\mathbf{y}|\theta\} \in \mathbb{R}^N \) the subject-specific mean, modelling the between-subject variations, and with \( \mathbf{w}_\omega \) and \( \mathbf{w}'_\theta \) the statistically independent, zero-mean within-subject variations. Furthermore, we assume Normal, zero-mean probability densities for \( \mu_\omega \) and \( \mu'_\theta \) for unknown \( \omega \) and \( \theta \), and for \( \mathbf{w}_\omega \) and \( \mathbf{w}'_\theta \). If \( \mathbf{x} \) and \( \mathbf{y} \) are not zero-mean, estimated means have to be subtracted prior to comparison. Such a simple model cannot be expected to work well for HiRes face recognition, but when LoRes faces with fewer details are involved it can still be applied successfully.

The covariance matrices of \( \mathbf{x} \) and \( \mathbf{y} \) are

\[
\Sigma_{xx} = \mathbb{E}\{\mathbf{x}\mathbf{x}^T\} \in \mathbb{R}^{M \times M} \quad \text{and} \quad \Sigma_{yy} = \mathbb{E}\{\mathbf{y}\mathbf{y}^T\} \in \mathbb{R}^{N \times N},
\]

respectively, and the cross-covariance matrices are

\[
\Sigma_{xy} = \mathbb{E}\{\mathbf{x}\mathbf{y}^T\} \in \mathbb{R}^{M \times N} \quad \text{and} \quad \Sigma_{yx} = \Sigma_{xy},
\]

respectively. Then \( \Sigma_{xy} = \mathbb{E}\{\mu_\omega\mu'_\theta^T|\omega = \theta\} \). If \( \omega \neq \theta \), \( \Sigma_{xy} = 0 \). For the probability densities of the pairs of feature vectors we then have, respectively

\[
\begin{align*}
\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} | H_s \sim \mathcal{N}
\begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}
\begin{pmatrix} 0 \\ \Sigma_{yy} \end{pmatrix},
\end{align*}
\]

\[
\begin{align*}
\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} | H_d \sim \mathcal{N}
\begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{pmatrix}
\begin{pmatrix} 0 \\ \Sigma_{yy} \end{pmatrix}.
\end{align*}
\]

Covariance and cross-covariance matrices need to be estimated in a training process. The cross-covariance matrix \( \Sigma_{xy} \) is estimated as

\[
\hat{\Sigma}_{xy} = \frac{1}{K} \sum_{i=1}^{K} \hat{\mu}_i \hat{\mu}'_i^T,
\]

with \( K \) the number of individuals involved in training and \( \hat{\mu}_i \) and \( \hat{\mu}'_i \) the estimated sample means of subject \( i \). It can be shown that estimating \( \Sigma_{xy} \) in this way is equivalent to estimating \( \Sigma_{xy} \) from all the possible combinations of pairs of feature vectors \( \mathbf{x} \) and \( \mathbf{y} \) in a training set. The rank of \( \hat{\Sigma}_{xy} \) can be
at most min\((N, K - 1)\). The \(-1\) is included because the sample means are zero-mean.

By substituting the Normal PDFs corresponding to (4.13) and (4.14) into (4.10), taking the log and ignoring some constants, we arrive at the following similarity score:

\[
s(x, y) = (4.16) \left( x^T y^T \right) \left( \begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{pmatrix}^{-1} - \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}^{-1} \right) \begin{pmatrix} x \\ y \end{pmatrix}.
\]

This score is optimal because it monotonically increases with the (log-)likelihood ratio. In order to simplify (4.16) and to assure that the estimated covariance matrices have full rank and can be inverted we simultaneously reduce the dimensionality and apply whitening transforms to \(x\) and \(y\), resulting in

\[
x_w = W_H x \in \mathbb{R}^{M_w} \quad \text{and} \quad y_w = W_L y \in \mathbb{R}^{N_w},
\]

with dimensionalities \(M_w\) and \(N_w\) respectively. Usually \(M_w < M\), \(N_w < N\), and \(M_w \geq N_w\). As a result we have that

\[
\Sigma_{xx}^w = E\{x_w x_w^T\} = I, \quad \Sigma_{yy}^w = E\{y_w y_w^T\} = I \quad \text{and} \quad \Sigma_{xy}^w = W_H \Sigma_{xy} W_L^T
\]

with \(I\) an identity matrix of appropriate size. The similarity score then becomes

\[
s(x_w, y_w) = (4.19) \left( x_w^T y_w^T \right) \left( \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}^{-1} - \begin{pmatrix} I & \Sigma_{xy}^w \\ \Sigma_{yx}^w & I \end{pmatrix}^{-1} \right) \begin{pmatrix} x_w \\ y_w \end{pmatrix}.
\]

We will further simplify (4.19). First we apply a singular value decomposition to \(\Sigma_{xy}^w\), such that

\[
\Sigma_{xy}^w = UDV^T,
\]

with \(U \in \mathbb{R}^{M_w \times M_w}\), and orthonormal, \(V \in \mathbb{R}^{N_w \times N_w}\), and orthonormal, and \(D \in \mathbb{R}^{M_w \times N_w}\). The first \(N_w\) rows of \(D\) form a diagonal matrix consisting of singular values \(\nu_i, i = 1, \ldots, N_w\) in decreasing order. The last \(M_w - N_w\) rows of \(D\) are an all-0 matrix. In a trained classifier the rank of \(D\) can be at most \(D = \min(N_w, K - 1)\), with \(K\) the number of individuals in the training set. If a smaller feature vector is desired, \(D\) can be chosen to be less than
4.3 Low-resolution face alignment and recognition using mixed-resolution classifiers

\[ \min(N_w, K - 1). \]

We now transform the feature vectors again, such that

\[ x_c = (U_{*,1:D})^T x_w \in \mathbb{R}^D \quad \text{and} \quad y_c = (V_{*,1:D})^T y_w \in \mathbb{R}^D, \quad (4.21) \]

where the subscript \(*, 1:D\) denotes that only the first \(D\) columns of matrix are taken. The subscript \(c\) indicates that these transformations map the feature vectors to a common subspace. It can be shown that these transformations, which reduce the feature dimensionality to \(D\), will result in the same similarity score as transformations using the full matrices \(U\) and \(V\). For the similarity score we now have

\[ s(x_c, y_c) = (4.22) \]

\[ (x_c^T y_c^T) \left( \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} - \begin{pmatrix} 1 & D \\ D & 1 \end{pmatrix}^{-1} \right) \begin{pmatrix} x_c \\ y_c \end{pmatrix}, \]

with \(D \in \mathbb{R}^{D \times D}\) redefined as a diagonal matrix with the \(D\) largest singular values \(\nu_i\) of \(\Sigma_{xy}^w\) on the diagonal. After some manipulations, we obtain

\[ s(x_c, y_c) = (4.23) \]

\[ -\sum_{i=1}^{D} \frac{\nu_i}{1 - \nu_i} (x_{c,i} - y_{c,i})^2 + \sum_{i=1}^{D} \frac{\nu_i}{1 + \nu_i} (x_{c,i} + y_{c,i})^2. \]

In (4.23) a factor of 1/2 has been left out. A full expression for the log-likelihood ratio, that includes all the constants that have been ignored is

\[ \log(l(x_c, y_c)) = -\frac{1}{2} \sum_{i=1}^{D} \log \left(1 - \nu_i^2 \right) + \frac{1}{4} s(x_c, y_c). \quad (4.24) \]

Because the \(\nu_i\) depend on training data, the use of this full expression is recommended in \(n\)-fold cross-validation experiments, since then the first term may differ slightly per validation step. Fig. 4.7 shows a block diagram of the classifier according to (4.23). The blocks perform matrix multiplications, except the rightmost ones, which compute a squared vector norm. The vectors \(\bar{x}\) and \(\bar{y}\) are the average HiRes and LoRes facial images, respectively. The matrices \(\Delta_{\text{DIF}}\) and \(\Delta_{\text{SUM}}\) are diagonal matrices, defined by \(\Delta_{\text{DIF},ii} = \sqrt{\frac{\nu_i}{1 - \nu_i}}, i = 1, \ldots, D\) and \(\Delta_{\text{SUM},ii} = \sqrt{\frac{\nu_i}{1 + \nu_i}}, i = 1, \ldots, D\), respectively.

In a similar way a likelihood-ratio based classifier can be derived for other types of heterogeneous features, e.g. for visual light and near infrared facial images,
and for the case that feature sets of possibly different numbers if multiple captures must be compared. In the following part of this paper we use the term *MixRes* for this method. The MixRes method has been implemented in MATLAB. The code is available on request.

### 4.3.5.2 Matching-score based registration

In the LoRes face recognition field, manually marked eye-coordinates are often still used, because landmarks cannot reliably be detected automatically. However, even the manual landmarks are usually not accurate enough because the eyes are not clear when the images are too small [61]. On the other hand, the variations between the true eye-coordinates and the manually marked ones are small in pixels, for example, the variations are within [-2, 2] pixels if the distance between the eyes is 10 pixels. Thus, we propose to use matching-score based registration to benefit LoRes face recognition. This method was proposed for inaccurately aligned faces in [62], and its effectiveness on HiRes images has already been demonstrated in [63,64].

In a normal face recognition system, the probe and reference images are registered using facial landmarks, and then the aligned images are compared. For a probe image $x_p$ and a reference image $x_r$, given the eye-coordinates of the probe image $\rho$ (reference images are assumed to be pre-aligned), the similarity score is written as $s(x_p(\rho), x_r)$. For matching-score based registration, the alignment of the probe image is varied resulting in several aligned images. All those images are compared with each gallery image using the chosen classifier. The best result for each gallery image is stored as the genuine or imposter
score which is used for the subsequent verification or identification process. Matching-score based registration tries to find the eye-coordinates \( \rho^* \), that maximizes the similarity between \( x_p(\rho) \) and \( x_r \), (4.25).

\[
\rho^* = \arg \max_\rho s(x_p(\rho), x_r)
\] (4.25)

The output similarity score is then \( s^*(x_p(\rho^*), x_r) \). We call this method MSBR for short.

### 4.3.5.3 Extended training

An alternative way to compensate for misalignment is to include misaligned images in the training set [65]. Unlike in MSBR, different aligned images are generated in the training set instead of on the probe images. The training set contains HiRes and LoRes image pairs. HiRes images are assumed to be perfectly aligned. But for each LoRes image, the eye-coordinates are varied to generate a set of (mis-)aligned images. The possible misalignment in the probe is thus modelled using this extended training set. The given eye-coordinates are noted as \( \rho \) which can be obtained by manual marking. We vary \( \rho \) to obtain \( m \) different coordinates \( [\rho_1, \rho_2, ..., \rho_m] \). Thus, for one LoRes training image \( x_L \), \( m \) images are generated with different alignment \( [x_L(\rho_1), x_L(\rho_2), ..., x_L(\rho_m)] \). This extended LoRes set together with the original HiRes set will form the new training set. This method will be denoted as ET.

### 4.3.6 Experiments

In this section, we demonstrate the effectiveness of our proposed methods using real LoRes data. The SCface database is chosen in our experiments because it contains surveillance quality face images. We did not choose other commonly used databases like FERET, FRGC or LFW because their image resolutions are much higher.

As shown in Table 4.4 and Table 4.5, many publications use different experimental settings and there is no straightforward way to select the best method. We chose to compare our method with three state-of-the-art LoRes face recognition methods CBD [15], CLPM [34] and DSR [13] and the state-of-the-art HiRes face recognition method FaceVACS [66]. The CBD method is proposed in a recent publication and we are able to duplicate their protocol of evaluation using surveillance quality images from the SCface database. The CLPM
method was not evaluated using real LoRes data in [34], but the source code is available so we can test it in our experiments. The correctness of the CLPM code was evaluated using the same setting on FERET database as in [34]. The rank-1 recognition rate with 80 features is 90%, which is the same as reported in [34]. For the DSR method we used our own implementation of the method described in [13]. We repeated the experiment described in [13] on SCface data to verify our implementation and found that our implementation had a 24% rank-1 recognition rate, which was slightly better than the result in [13]. FaceVACS is a commercial face recognition software developed by Cognitec Systems GmbH. We use FaceVACS for comparison because commercial systems are also used in real surveillance cases. Note that FaceVACS was not designed for very LoRes facial images and thus is used here outside its normal specifications. For HiRes images it outperforms our method by a large margin.

In our experiments, the parameters of each method are chosen to ensure a good performance. The feature dimension selected for CLPM is 170. The DSR method is employed with the LDALLR classifier, for which 50 PCA and 40 LDA feature vectors are chosen. The likelihood ratio of the MixRes method in all experiments is calculated using equation (4.24). The parameters of the MixRes method are shown in Table 4.6. Larger numbers are chosen when MixRes is used in combination with MSBR. The experiments using FaceVACS are conducted with the eye-coordinates provided. Otherwise, eight images from dist3, 30 images from dist2 and 151 images from dist1 could not be processed successfully because FaceVACS could not detect the eyes in those images.

Table 4.6: Parameters of MixRes and its combination with MSBR and ET. \( M_w \) and \( N_w \) are the number of feature vectors of HiRes and LoRes training after the first dimensionality reduction. \( D \) is the number of feature vectors after the second dimensionality reduction.

<table>
<thead>
<tr>
<th>Method</th>
<th>( M_w )</th>
<th>( N_w )</th>
<th>( D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MixRes, +ET</td>
<td>70</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>+MSBR, +Both</td>
<td>100</td>
<td>100</td>
<td>60</td>
</tr>
</tbody>
</table>

In Section 4.3.5.1, we assume a Normal distribution for the facial features. We illustrate the appropriateness of this assumption using images from the FRGC database which were also used for training in Section 4.3.4. We choose two resolutions, \( 131 \times 131 \) and \( 23 \times 23 \) as HiRes and LoRes, respectively. We use 3464 images of the first 185 subjects to train the MixRes classifier. Then
we use the remaining 1007 images of 90 subjects for testing. We apply the transformation matrices $W_H$ and $U_{*,1:D}$ on the HiRes testing images, and $W_L$ and $V_{*,1:D}$ on the LoRes testing images. This results in a feature vector of dimensionality 40 for each image. We randomly select two HiRes feature elements $x_{c,i}$ and $x_{c,j}$ and corresponding LoRes feature elements $y_{c,i}$ and $y_{c,j}$ and plot the histograms $h_{x,i}$, $h_{x,j}$, $h_{y,i}$ and $h_{y,j}$ in Fig. 4.8. As we can see, all of the four seem to follow a Normal distribution, which illustrates that our assumption for MixRes is reasonable.

![Histograms](image)

Fig. 4.8: Histograms illustrating the Normal distribution for facial features for MixRes. Top row: histograms of HiRes feature elements; Bottom row: histograms of LoRes feature elements.

In Subsection 4.3.6.1, we compare our methods with the four methods described above, following the protocol that was used to evaluate the CBD method in [15]. Thanks to the authors, who had provided us with the necessary details, we were able to duplicate their protocol, which will be used in this subsection. The images of the second longest distances from the SCface database are down-sampled and used as probe in these experiments.

In Subsection 4.3.6.2, we further demonstrate the effectiveness of our methods on real LoRes data following a more realistic protocol. The experiments are more challenging because images from the SCface database captured at the greatest distance are used as probe and single mug-shot images are used as gallery. The CLPM method and FaceVACS are used for comparison.

In Subsection 4.3.6.3, we show results obtained on our own database to demonstrate that our methods are not optimized only for the SCface database.
In Table 4.7 we list the original resolutions of probe and gallery images and the resolutions actually used in each subsection. In subsection 4.3.6.1 the testing protocol defined in [15] describes that the input images are rescaled to a lower resolution. In sections 4.3.6.2 and 4.3.6.3 we also down-sampled the HiRes gallery images, but the LoRes probe images preserve their original resolution.

Table 4.7: Resolutions (pixels) of gallery and probe images in each subsection.

<table>
<thead>
<tr>
<th>Subsection</th>
<th>HiRes Gallery</th>
<th>LoRes Probe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>Actual</td>
</tr>
<tr>
<td>4.3.6.1</td>
<td>68 x 55</td>
<td>30 x 24</td>
</tr>
<tr>
<td>4.3.6.2</td>
<td>800 x 800</td>
<td>80 x 80</td>
</tr>
<tr>
<td>4.3.6.3</td>
<td>243 x 243</td>
<td>80 x 80</td>
</tr>
</tbody>
</table>

4.3.6.1 Comparison with the state-of-the-art

The experiments are conducted on the Surveillance Camera Face (SCface) database [10]. The SCface database contains images of 130 subjects taken by five surveillance cameras at three distances, namely 4.20 m (dist1), 2.60 m (dist2), and 1.00 m (dist3). There are 5 x 130 = 650 images for each distance. It also contains one frontal mug-shot image for each subject.

Fig. 4.9: Sample images from the SCface database in the experiments in Section 4.3.6.1. First row: HiRes, second row: LoRes.

The protocol of [15] is duplicated so that we can compare the reported results of the CBD method directly. The region of interest is obtained by cropping
the face region of the images based on the eye-coordinates provided in the
database. Images from dist2 are used as LoRes and images from dist3 are
used as (relatively) HiRes. The sizes of the HiRes and LoRes images after align
and rescaling using bicubic interpolation are $30 \times 24$ pixels and $15 \times 12$ pixels
respectively. Histogram equalization is used to normalize the illumination.
Sample images are shown in Fig. 4.9. We randomly select 100 subjects and
four images of these subjects for training. The remaining 30 subjects are used
for testing, of which four images per subject from dist3 are randomly selected
for the gallery and one image per subject from dist2 is used for the probe. The
gallery and probe images of the same subject are taken by different cameras.
Thus, we have 400 training images for both HiRes and LoRes, 120 gallery
images and 30 probe images each time. This is the same setting as in [15]. In
addition, we repeated the experiments using a single gallery image per subject
as this is a more realistic setting. Each experiment is repeated 100 times.

Table 4.8: Comparison of MixRes to CBD, DSR, CLPM and FaceVACS. The values
are in the format: average value (standard deviation). The verification rates are
obtained at FAR = 10%. Probe: dist2, $15 \times 12$ pixels. Gallery: dist3, $30 \times 24$ pixels.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>AUC</th>
<th>Verification %</th>
<th>Rank-1 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Gallery</td>
<td>CBD</td>
<td>0.77 (0.03)</td>
<td>-</td>
<td>52.7 (9.9)</td>
</tr>
<tr>
<td></td>
<td>DSR</td>
<td>0.69 (0.03)</td>
<td>29.3 (6.1)</td>
<td>31.7 (7.3)</td>
</tr>
<tr>
<td></td>
<td>CLPM</td>
<td>0.64 (0.05)</td>
<td>23.9 (8.4)</td>
<td>9.0 (5.4)</td>
</tr>
<tr>
<td></td>
<td>MixRes</td>
<td><strong>0.88 (0.03)</strong></td>
<td><strong>64.6 (8.1)</strong></td>
<td><strong>57.3 (9.5)</strong></td>
</tr>
<tr>
<td></td>
<td>FaceVACS</td>
<td>0.76 (0.02)</td>
<td>47.0 (3.5)</td>
<td>19.3 (3.2)</td>
</tr>
</tbody>
</table>

| Single Gallery | DSR    | 0.69 (0.03) | 28.9 (6.0)    | 30.2 (8.8)  |
|                | CLPM   | 0.62 (0.05) | 19.1 (7.8)    | 7.0 (5.1)   |
|                | MixRes | **0.84 (0.03)** | **57.4 (7.5)** | **47.9 (8.9)** |
|                | FaceVACS | 0.76 (0.02) | 46.5 (1.6)    | 19.1 (1.5)  |

The Area Under the Curve (AUC) and rank-1 identification rates using the
CBD method are reported in [15]. We choose their best results and compare
with the results from our experiments in Table 4.8. In addition, we provide
verification rates at FAR 10%. We also collect all the genuine and imposter
scores from the 100 repetitions of each experiment to plot Receiver Operating
Characteristic (ROC) curves (Fig. 4.10).
As we can see, the MixRes method performs best of the five. The CBD method, which was also designed for real LoRes face recognition, performs better than the rest but still worse than our MixRes method. DSR performs the second worst. Although CLPM was demonstrated to perform well on down-sampled images (see Table 4.4), it gives the worst results in our real LoRes experiments. FaceVACS outperforms DSR and CLPM despite the fact that it was designed for HiRes face recognition. When a single gallery image per subject is used, all the methods decrease in performance, but the MixRes method remains the best.

To further improve the face recognition performance, MSBR and ET are employed. Rigid transformation and isotropic scaling are used for alignment in our experiments. The variation of each eye-coordinate for probe images is [-2,2] pixels based on the manually marked eye-coordinates. The training procedure of ET uses 20 randomly selected (mis-)aligned images for each original image. The results are shown in Table 4.9 and Fig. 4.11.

As we can see, the verification results are significantly improved in all cases, especially when we combine MSBR and ET the improvement is more than 10%. The rank-1 recognition rate only improved slightly using ET and not in the other settings. It is due to the fact that closed-set recognition only picks the best score, whereas MSBR optimizes all the similarity scores (imposter scores could benefit more from MSBR than genuine scores in some cases).
4.3 Low-resolution face alignment and recognition using mixed-resolution classifiers

Table 4.9: MixRes combined with MSBR and ET. The values are in the format: average value (standard deviation). The verification rates are obtained at FAR = 10%. MG: multi-gallery, SG: single-gallery. Probe: dist2, 15 × 12 pixels. Gallery: dist3, 30 × 24 pixels.

<table>
<thead>
<tr>
<th>Settings</th>
<th>Verification %</th>
<th>Rank-1 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>MG: MixRes</td>
<td>64.6 (8.1)</td>
<td>57.3 (9.5)</td>
</tr>
<tr>
<td>MG: +MSBR</td>
<td>74.5 (6.2)</td>
<td>55.6 (7.2)</td>
</tr>
<tr>
<td>MG: +ET</td>
<td>75.4 (6.5)</td>
<td>58.2 (9.4)</td>
</tr>
<tr>
<td>MG: +both</td>
<td>81.4 (6.1)</td>
<td>51.8 (7.6)</td>
</tr>
<tr>
<td>SG: MixRes</td>
<td>57.4 (7.5)</td>
<td>47.9 (8.9)</td>
</tr>
<tr>
<td>SG: +MSBR</td>
<td>65.2 (7.4)</td>
<td>46.7 (8.9)</td>
</tr>
<tr>
<td>SG: +ET</td>
<td>66.8 (7.4)</td>
<td>50.3 (8.1)</td>
</tr>
<tr>
<td>SG: +both</td>
<td>72.5 (9.4)</td>
<td>47.5 (9.1)</td>
</tr>
</tbody>
</table>

Fig. 4.11: ROC curves of MixRes and its combination with MSBR and ET: (a) multi-gallery (b) single gallery. Probe: dist2, 15 × 12 pixels. Gallery: dist3, 30 × 24 pixels.

4.3.6.2 Real LoRes experiments

In the experiment of the previous section, the gallery images were not separately captured mug-shots, but simply higher resolution images from the sequence. A more realistic setting is to use separately captured mug-shots as a gallery.

We follow the guidelines below to make sure the experiments are representative of realistic applications.
1. Report verification results.
2. Only use real LoRes images as probe.
3. Gallery images should be HiRes, preferably mug-shots.
4. No training on a gallery, nor on users represented in the gallery.
5. Use only one HiRes gallery image per subject.
6. For small databases, use cross-validation in order to produce statistically more significant results. Use as many probe images as possible to reduce the standard deviation of the results.

![Sample images from the SCface database in the experiments in Section 4.3.6.2. First row: mug-shots, second row: dist3, third row: dist2, last row: dist1.](image)

The images are processed in such a way that they preserve their original resolution. The sizes of the cropped images for dist3, dist2, dist1 are $80 \times 80$, $56 \times 56$, $32 \times 32$ respectively (note that the cropping is different than in sub-section 4.3.6.1). The images are aligned using eye-coordinates provided in the database. An elliptic mask is applied to select the region of interest. Thus the selected regions are smaller than the above resolutions. Histogram equalization is used to normalize the illumination. Some samples of processed images are shown in Fig. 4.12. The original size of mug-shot face images is around $800 \times 800$ pixels. To make the comparison more efficient, mug-shot images are down-sampled to the same resolution as dist3 images in our experiments. The probe images (dist1), are always of the same resolution as the original.
Dist1 (the largest distance) images are used as probes in all the following experiments. Different combinations of training and gallery sets are employed to test the performance of our methods in different situations. In all experiments, 100 subjects are randomly selected for training and the rest are for testing. The HiRes training images are of the same size as dist3 (80 × 80 pixels), and the LoRes training images are of the same size as dist1 (32 × 32 pixels). There are five images per subject from dist1, dist2 and dist3, and one image per subject from mug-shots. Our implementation of the MixRes method requires that the HiRes and LoRes training sets have the same number of images. Therefore, we replicate the mug-shot images so that each subject has five identical mug-shots.

Three different training settings are used. The first one is to have mug-shot images as HiRes training and dist1 images as LoRes training. Because there is only one mug-shot image per subject, we add images from dist3 or dist2 to the training set to obtain better training results in the other two settings. All the three training settings are used for MixRes and CLPM. Then we select the best setting to conduct the combinations of MixRes, MSBR and ET.

All the mug-shots are used as gallery and all 650 images from dist1 are used as probe for FaceVACS, because it does not require training images. Thus, no standard deviation is presented for FaceVACS because no cross-validation was used. MSBR and ET are conducted in the same way as in Section 4.3.6.1. In addition, MSBR is also conducted with automatic initialization (denoted as MSBRa). That is, instead of manually marked eye-coordinates, the starting points are determined by the region of interest detected by the Viola-Jones face detector [43]. The scale factor is set to 1.002 to detect the small face regions in the images. The false positives are then manually discarded. We compute verification rates at FAR = 10% and rank-1 recognition rates, see Table 4.10.

MixRes significantly outperforms CLPM and FaceVACS in all three settings. FaceVACS performing the worst shows that this is a difficult experiment for state-of-the-art algorithms which are designed for HiRes images. The CLPM method still performs poorly as in the experiments in the previous section, while it had very good performance on down-sampled data as shown in Table 4.4. This further supports our conclusion made in Section 4.3.4 that methods designed for LoRes face recognition should be evaluated using real LoRes data instead of images down-sampled from HiRes.

A sufficient number of training images is required for MixRes to achieve good results. If we only use the mug-shot images as HiRes training and dist1 as
Table 4.10: Results on the SCface database with different experiment settings. Gallery: mug-shots, 80 × 80 pixels. Probe: dist1, 32 × 32 pixels. The values are in the format: average value (standard deviation). The verification rates (%) are obtained at FAR = 10%. The abbreviation ds stands for down-sampled images. d1, d2 and d3 stand for dist1, dist2 and dist3, respectively.

<table>
<thead>
<tr>
<th>Principle</th>
<th>Train (HiRes)</th>
<th>Train (LoRes)</th>
<th>Verific. (%)</th>
<th>Rank-1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MixRes</td>
<td>mug-shot</td>
<td>d1</td>
<td>47.8 (4.8)</td>
<td>34.9 (5.0)</td>
</tr>
<tr>
<td></td>
<td>mug-shot + d3</td>
<td>d1 + d2 ds</td>
<td>63.1 (5.8)</td>
<td>43.4 (5.6)</td>
</tr>
<tr>
<td></td>
<td>mug-shot + d2</td>
<td>d1 + d2 ds</td>
<td>67.2 (6.0)</td>
<td>48.3 (5.3)</td>
</tr>
<tr>
<td>CLPM</td>
<td>mug-shot</td>
<td>d1</td>
<td>10.2 (2.7)</td>
<td>3.4 (1.7)</td>
</tr>
<tr>
<td></td>
<td>mug-shot + d3</td>
<td>d1 + d2 ds</td>
<td>15.9 (3.5)</td>
<td>5.9 (2.0)</td>
</tr>
<tr>
<td></td>
<td>mug-shot + d2</td>
<td>d1 + d2 ds</td>
<td>18.3 (3.6)</td>
<td>7.2 (2.2)</td>
</tr>
<tr>
<td>FaceVACS</td>
<td>-</td>
<td>-</td>
<td>21.7</td>
<td>2</td>
</tr>
<tr>
<td>MixRes +MSBR</td>
<td>mug-shot + d2</td>
<td>d1 + d2 ds</td>
<td>73.4 (5.8)</td>
<td>48.4 (5.2)</td>
</tr>
<tr>
<td>MixRes +ET</td>
<td>mug-shot + d2</td>
<td>d1 + d2 ds</td>
<td>61.0 (6.9)</td>
<td>39.0 (6.7)</td>
</tr>
<tr>
<td>MixRes +Both</td>
<td>mug-shot + d2</td>
<td>d1 + d2 ds</td>
<td>53.2 (7.2)</td>
<td>25.6 (4.5)</td>
</tr>
<tr>
<td>MixRes +MSBRa</td>
<td>mug-shot + d2</td>
<td>d1 + d2 ds</td>
<td>72.0 (5.7)</td>
<td>48.0 (6.1)</td>
</tr>
</tbody>
</table>

LoRes training, the verification result at FAR = 10% is less than 50%. When we enlarge the training sets by combining two sets of images, the verification rates increase significantly up to 67.2%. The rank-1 recognition rates also show an improvement of around 10%. If we have more realistic data (mug-shot images and the corresponding LoRes images) to add in the training set, we would expect even more improvement. With the help of MSBR, MixRes can reach 73.4% verification rate. However, ET does not improve the results as it did in Section 4.3.6.1. A possible explanation is that the added training data may not be fully representative of the testing data, causing the results to become worse in this experiment. This could also be the reason that MixRes in combination with both MSBR and ET has worse performance. It is promising that the results from MSBR with automatic initialization (MSBRa) show only marginal degradation as compared to the results from MSBR with initialization based on manual eye-coordinates. This means that the LoRes face recognition system can be fully automatic if the detection is correct.
4.3 Low-resolution face alignment and recognition using mixed-resolution classifiers

4.3.6.3 Evaluation using other data

To demonstrate that our methods are not optimized only for the SCface database, we conduct another experiment using SCface images for training but using our own database, which is described in Section 4.3.4, for testing. The images taken at a distance of 8 m are selected as probe because their original resolution is similar to that of images at dist1 from the SCface database. The gallery images are the ones that were taken at a distance of 1 m in a separate session (same as in Section 4.3.4). The training setting that provided the best results from the previous SCface experiments is selected, that is, mug-shots and dist2 images are combined for HiRes training while dist1 and down-sampled dist2 images are used for LoRes training. MSBR is conducted with automatic initialization. FaceVACS is used with manually marked eye-coordinates. The results are shown in Table 4.11. Because this is an easier experiment on a small database with only 25 subjects and controlled environment, MixRes gives very good results. The fully automatic method +MSBRa performs close to MixRes with manually marked eye-coordinates. On the other hand, FaceVACS performs much poorer than MixRes (about a 50% difference in verification rate at FAR 10%).

Table 4.11: Results on our own database. Gallery: 1m, 80 × 80 pixels. Probe: 8m, 32 × 32 pixels. The verification rates are obtained at FAR = 10%.

<table>
<thead>
<tr>
<th></th>
<th>MixRes</th>
<th>+MSBRa</th>
<th>FaceVACS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verification</td>
<td>91%</td>
<td>84%</td>
<td>43%</td>
</tr>
<tr>
<td>Rank-1</td>
<td>80%</td>
<td>64%</td>
<td>20%</td>
</tr>
</tbody>
</table>

4.3.7 Conclusion

In a common case of low-resolution face recognition, recognition of suspects from a long distance, there is usually a resolution mismatch between low-resolution probe and high-resolution gallery. Most existing methods, which are designed to match high-resolution images, cannot handle low-resolution probes well. In this paper, we present a classifier, mixed-resolution biometric comparison, specifically designed for heterogeneous cases, which allows direct comparison of different resolution images. We identify that proper alignment is one of the major challenges in low-resolution face recognition. We also show that there is a large difference in face recognition performance
between down-sampled and real low-resolution images and only the results from real low-resolution can be trusted. To cope with the alignment problem, we investigate two methods, matching-score based registration and extended training. Extended training is less effective but it helps when there is not enough training data. Matching-score based registration is useful in combination with our mixed-resolution classifier. It can also be initialized using a region obtained from face detection, which results in a fully automatic low-resolution face recognition method. Our experiments using real low-resolution data show that our methods outperform the state-of-the-art. In our experiment on the SCface database, the combination of our mixed-resolution classifier with matching-score based registration and extended training outperforms the state-of-the-art by 20% of verification rate at false acceptance rate 10% in the multi-gallery setting.

4.3.8 Acknowledgements

We would like to thank Cognitec Systems GmbH for supporting our research by providing the FaceVACS software. Results obtained for FaceVACS were produced in experiments conducted by the University of Twente, and should therefore not be construed as a vendor’s maximum effort full capability results.

We would also like to thank the authors of [15] for providing us with the necessary details to duplicate their protocol and the authors of [34] for availing software.

4.4 Discussion

In this chapter, we discussed about RQ.1(b) regarding a method for direct comparison of low-resolution and high-resolution facial images and RQ.2(a)(b) regarding the problem of using down-sampled images instead of real low-resolution images for testing. To answer RQ.1(b), we proposed the classifier, mixed-resolution biometric comparison, which can directly compare low-resolution probe images with high-resolution gallery images. This method is based on the log-likelihood ratio of both high-resolution and low-resolution image pairs. Our experiments on surveillance quality images demonstrate that this method significantly outperforms the state-of-the-art. For RQ.2(a), we identified that alignment is the main cause of the differences in face recognition performance on down-sampled images and real low-resolution images taken at controlled situations. With this knowledge, we can improve the performance
of face recognition on real low-resolution facial images by improving alignment (RQ.2(b)). In our experiments on the SCface data set, matching-score based registration was helpful for recognition. We will further explore the importance of alignment for real low-resolution facial images and how well matching-score based registration can improve the face recognition performance in the next chapter.
Chapter 5

The importance of proper alignment in low-resolution face recognition

5.1 Introduction

We have identified that alignment plays an important role in the recognition performance on down-sampled and real low-resolution images using a self-collected data set and three basic face recognition methods in the previous chapter. In this chapter, we explore further the importance of alignment for low-resolution face recognition. We design a complete set of experiments using different face data sets which are collected in controlled or uncontrolled situations. We also use various types of face recognition methods: basic face recognition methods, a deep-network based face recognition method, methods specially designed for low-resolution face recognition and two commercial face recognition systems. Therefore, we not only test the influence of alignment on these methods, but also show the performance of these methods on different resolutions. The detailed explanation of the resolution scale, which we have proposed in Fig. 1.2 in Chapter 1, is also presented in the following section. Section 5.2 contains our paper which is submitted to IET Biometrics.
5.2 Low-resolution face recognition and the importance of proper alignment

5.2.1 Abstract

Face recognition methods for low-resolution are often developed and tested on down-sampled images instead of on real low-resolution images. Although there is a growing awareness that down-sampled and real low-resolution images are different, few efforts have been made to analyse the differences in recognition performance. In this paper, we explore the differences and demonstrate that alignment is a major cause especially in absence of pose and illumination variations. We found that the recognition performances on down-sampled images are flattered mostly due to the fact that the images are perfectly aligned before down-sampling using high-resolution landmarks while the real low-resolution images have much poorer alignment. To obtain better alignment for real low-resolution images, we apply matching-score based registration which does not rely on accurate landmarks. We propose to divide low resolution into three ranges to harmonize the terminology: upper low resolution, moderately low resolution and very low resolution. Most face recognition methods perform well on upper low resolution. Moderately low resolution is a challenge for commercial systems, but a low-resolution deep-learning method can handle it very well. The performance of most methods degrades significantly for very low resolution, except for simple holistic methods which perform the best.

5.2.2 Introduction

Face recognition is one of the most popular biometric modalities [67,68]. Comparing to fingerprint and iris, facial images can be captured at a distance and thus they are natural and non-intrusive. Face recognition for high-resolution facial images has achieved great success in the past decades. However, common applications for face recognition such as surveillance cannot always capture facial images at a close range. The images captured at a distance suffer from various problems which make them more difficult to recognize than those captured at close range. One of the problems is that the images are of low-resolution so that they contain less information. Another problem is that the images are usually captured in uncontrolled situations, which results in illumi-

\[1\]The contents of this section are in “Low-resolution face recognition and the importance of proper alignment”, submitted to IET Biometrics, 2017.
nation and pose variations. In addition, the images are often noisy because of low light and suffering from compression artifacts. Thus, low-resolution face recognition is still challenging and receiving substantial attention nowadays.

One approach to improve low-resolution face recognition is by using super-resolution. Hennings-Yeomans et al. [57] proposed an algorithm that combines the underlying assumptions of super-resolution methods with subspace distance metrics used for classification. Zhang et al. [32] proposed a super-resolution method in morphable model space, which provides high-resolution information required for both reconstruction and recognition. Zou and Yuen [40] developed a data constraint for reconstructing super-resolution image features so that both the distances between the reconstructed images and the corresponding high-resolution images and the distances between super-resolution images from the same class are minimized.

Researchers also developed face recognition methods that perform face recognition directly on low-resolution images. Li et al. [34] proposed a method that projects both high-resolution gallery and low-resolution probe to a common feature space for classification using coupled mappings which minimize the difference between corresponding images. Moutafis and Kakadiaris [15] proposed a method that learns semi-coupled mappings for optimized representations. The mappings aim at increasing class-separation for high-resolution images and mapping low-resolution images to their corresponding class-separated high-resolution data. Peng et al. [55] proposed a likelihood ratio based method for direct comparison between images of different resolutions.

Due to the lack of appropriate data sets of real life surveillance situations, most of the research is conducted on down-sampled images, for example [32,34,57]. Researches like [2, 14, 40] include experiments on real low-resolution images but most of their experiments and analysis are conducted using down-sampled images. It is shown in [48] and [56] that face recognition and super-resolution methods both perform much better on down-sampled images than on real low-resolution images. In [56], it was mentioned that alignment is important for the performance differences but the experiments were conducted on a small data set and only simple basic face recognition methods are used.

In this paper we analyse the performance differences between down-sampled and real low-resolution facial images in more depth using various face recognition methods. We demonstrate that in absence of pose and illumination variations, the loss of recognition performance for real low-resolution images compared to down-sampled images can largely be attributed to poor alignment (or registration). When images are captured at uncontrolled situations
Chapter 5. The importance of proper alignment in low-resolution face recognition

(with illumination and pose variations), alignment still plays an important role in recognition performance. Proper alignment is a problem for low-resolution face recognition, because the landmarks, such as eye coordinates, that are used for high-resolution face recognition are much less reliable or even unavailable. We particularly demonstrate that: 1) If down-sampled images are aligned based on landmarks estimated after down-sampling instead of before, which is common but unrealistic, then the recognition performance becomes similar/closer to that obtained with real low-resolution images. 2) Conversely, if real low-resolution images are better registered by means of matching-score based registration, then the recognition performance increases towards that obtained with down-sampled images.

In literature on low-resolution face recognition, what in some papers is considered as low-resolution, is still considered as high-resolution in other papers. For example, the Interpupillary Distances (IPD) of low-resolution images described in [2] are from 2 pixels to 8 pixels, and the IPD of high-resolution images is 16 pixels. While in [3], the IPD of low-resolution images are 20 pixels. To harmonize the terminology in low-resolution face recognition, we propose the resolution scale as shown in Fig 5.1. We use IPD as the measure of resolution.

There are four biometric standards in this graph. Two of them are ISO/IEC 19794-5:2005 [4] and ANSI/INCITS 385-2004 [5]. They describe an example of proper face position in an image [69], where the size of the face is approximately 106 × 96 pixels and the corresponding IPD is about 50 pixels. We use this point to separate high resolution and low resolution. Another standard, the European norm EN 50132-7 [6], describes the recommended minimum size of the object for CCTV system. We calculate the IPD, which is approximately 25 pixels according to the description. ICAO Doc 9303 [7] mentions that standardized size portrait results in a facial image with IPD approximately 90 pixels. We divide low resolution into Upper Low Resolution (ULR), Moderately Low Resolution (MLR) and Very Low Resolution (VLR), and we use IPD 25 pixels (EN 50132-7) and 13 pixels as the separating points. ULR is relative higher resolution and thus is not a difficult task for most existing face recognition methods. Images of MLR are harder to recognize than ULR. Methods that are designed for low-resolution face recognition should outperform methods for high-resolution face recognition for MLR. VLR is extremely difficult for face recognition and we expect poor results for most methods. In addition, we added in the graph the resolution of several popular face data sets: FRGC v2.0 [8], LFW [9] and SCface [10], and the data sets used in our experiments: UT-FAD [70], Human ID [11] and COX [71]. The latter three
5.2 Low-resolution face recognition and proper alignment

Fig. 5.1: Resolution scale.
Chapter 5. The importance of proper alignment in low-resolution face recognition

contain data over a range of resolutions and are presented as bars in Fig 5.1. We focus on low resolution so we divide it into three ranges, but high resolution can also be divided.

The body of this paper is divided into two parts. The first part analyses and demonstrates that alignment is an important factor which causes the loss in recognition performance of real low-resolution images. The second part is about improving face recognition performance by compensating alignment problem on real low-resolution images using matching-score based registration.

5.2.3 The difference between down-sampled and real low-resolution images

5.2.3.1 Analysis

We have mentioned in the previous section that down-sampled images are commonly used as a substitute for real low-resolution images, but face recognition performance on down-sampled images is always much better than on real low-resolution images [48]. In this section, we will analyse the differences between down-sampled and real low-resolution facial images.

Firstly, we compare down-sampled images with real low-resolution images from the visual aspect. In Fig 5.2 we show some example images. Fig 5.2(a) is a high-resolution facial image, Fig 5.2(b) is obtained by down-sampling Fig 5.2(a) and Fig 5.2(c) is a real low-resolution image of the same subject captured at a distance. Although Fig 5.2(b) and Fig 5.2(c) are of the same resolution, the down-sampled image seems to have much better quality than the real low-resolution one. Firstly, the down-sampled image is much sharper. We can still see a lot of details from the down-sampled image, for example, the eyebrows are very clear, the white of eyes and the colour of the lips are visible. The real low-resolution image is much more blurred: the eyes become two blurred dark dots, the lips become gray, and the nose and eyebrows are almost invisible. Secondly, the real low-resolution image contains more noise than the down-sampled image, which mostly can be seen on the cheeks. Despite the above differences we can see from the pictures, there are other factors could play a role. For example, the perspective is different between images taken at different distances; there is more air between the subject and the camera in large distance which may affect the shape of the image [72]; lens distortion [73], etc. Pose and illumination are well-known factors that have a large influence
on face recognition for both high- or low-resolution facial images. In low-resolution face recognition, for example surveillance, the images are more likely to be captured under uncontrolled situations thus pose and illumination can be a serious problem (for example, in Fig 5.2(d)).

![Fig. 5.2: Down-sampled and real low-resolution image examples. (a) High-resolution, IPD 96 pixels (b) down-sampled, IPD 10 pixels (c) real, IPD 10 pixels (d) real with pose, IPD 12 pixels (subject from the Human ID data set)](image)

Secondly, we compare the differences in processing real and down-sampled low-resolution images for face recognition. The face recognition processing flow graphs are presented in Fig 5.3. These flow graphs are derived from [68]. Fig 5.3(a) is the recognition process for real low-resolution images. Fig 5.3(b) is the procedure how down-sampled images are normally processed in order to test low-resolution face recognition methods. We can see that a significant difference between the two is where the landmarks are detected. For real low-resolution images, the landmarks are detected on the low-resolution images which are very likely to be inaccurate. While for down-sampled images, the landmarks are detected on the high-resolution images which are usually very accurate. This could be essential because inaccurate landmarks will result in poor alignment and subsequently poor recognition results. In this case, Fig 5.3(c) is a more proper way of processing down-sampled images because it is closer to real low-resolution face recognition.

We choose to focus on the recognition process aspect and propose the following hypothesis: in absence of pose and illumination variations, alignment is the most important cause of the performance differences in face recognition between down-sampled and real low-resolution images.

### 5.2.3.2 Controlled situation

In this subsection, we set up experiments to test our hypothesis that alignment is the most important cause of the recognition performance differences between down-sampled and real low-resolution images.
Chapter 5. The importance of proper alignment in low-resolution face recognition

Fig. 5.3: Face recognition processing flow of (a) real low-resolution facial images (b) down-sampled facial images (commonly) (c) down-sampled images (proposed).

To test the hypothesis we first need both down-sampled images and real low-resolution images captured at the same moment of the corresponding high-resolution images: the former ones can be achieved by down-sampling the high-resolution images and the later ones should be captured at a farther distance. It is preferred that images captured at different distances are available so that we can see if the results are valid for all distances. Ideally, the images should be frontal with uniform illumination because we only want to test the influence of alignment.

The University of Twente - Faces At Distances (UT-FAD) data set [70] is used in our experiment. Although it is a small data set of only 22 subjects, it is the only data set available that meets the above requirements. Besides, noise is allowed in the experimental results as long as the trend is clear. The images of the UT-FAD data set were captured using a commercial camera CASIO EX-FC100. To minimize the changes of the condition on the subjects’ face, images were recorded in the following way: subjects remained a fixed position and the photos were taken at different distances. The first photo of each subject was taken at a distance of 2 m, and then we moved the camera one metre away and took a photo each time until we had 9 photos of this subject. Thus, the photos were taken at nine distances from 2 m to 10 m. Those images are used as probe in the following experiments. The faces are always frontal and have the same illumination. In addition, a gallery image of each subject was captured at a distance of 1 m after two weeks using the same camera with the same illumination and frontal pose.

The size of the cropped face regions of gallery images is $243 \times 243$ pixels with
5.2 Low-resolution face recognition and proper alignment

96 pixels between the eyes. The sizes of the probes from distance 2 m to 10 m are $131 \times 131$, $87 \times 87$, $64 \times 64$, $51 \times 51$, $44 \times 44$, $36 \times 36$, $33 \times 33$, $28 \times 28$ and $23 \times 23$ pixels, respectively. The corresponding average IPDs are 51, 34, 25, 20, 17, 14, 13, 11 and 9 pixels, respectively. The first three resolutions are ULR, the middle four resolutions are MLR and the last two resolutions are VLR according to our divisions of low resolutions. The facial images are aligned using manually annotated eye-coordinates. An ellipse mask is added to define the region of interest. Sample images are shown in Fig 5.4.

![Sample Images](image-url)

Fig. 5.4: Sample images from the UT-FAD data set (with IPD in the brackets).

We use eight different face recognition methods. Two basic face classifiers, Principal Component Analysis (PCA) [51] and Local Binary Patterns (LBP) [53], are used. We also use four state-of-the-art low-resolution face recognition methods, namely RIDN [25], MixRes [55], CLPM [34] and DSR [13]. RIDN, MixRes and CLPM are specially designed for comparing low-resolution probes with high-resolution galleries. RIDN is based on a deep neural network and developed for low-resolution face recognition. DSR is a super-resolution method which reconstructs high-resolution features of the probes. We use DSR in combination with PCA for recognition, which was also used in the original paper [13]. PCA, MixRes, CLPM and DSR+PCA are trained using 4064 high-quality images of 254 subjects from the FRGC v2.0 data set [8]. Distance measures employed for PCA and LBP are L1 norm and Chi square, respectively. In addition, two commercial face recognition methods are used. We call them System A and System B.

The eight methods are designed for different situations and thus have different requirements for the image resolutions. PCA and LBP require probe and gallery images have the same resolution, so the training and gallery images
are down-sampled to the same resolution as the probe images. RIDN is a pre-trained system and all the testing images have to be resized to $55 \times 60$ pixels. MixRes, CLPM and DSR are capable of comparing gallery and probe with different resolutions, so gallery images of their original size are used. The commercial systems should be able to cope with different resolutions, thus the images of original size are used.

Four of the methods, PCA, MixRes, CLPM and DSR+PCA, have parameters that can be changed. We choose the following parameters as in Table 5.1 to ensure a good performance of all the face recognition methods. For LBP, the images from distance 2 m to 7 m are divided into $7 \times 7$ regions and the ones from 8 m to 10 m are divided into $5 \times 5$ regions. Although the commercial systems both have landmark-detecting functions, they do not function well for low-resolution images, therefore, System A is provided with manual landmarks. However, System B has to use its own detecting function because it does not allow manual input of landmarks.

Three experimental settings are conducted. A brief description is in Table 5.2 and detailed explanations are in the following.

1. **REAL**: REAL low-resolution images captured at different distances are used as probe, those are, the images captured at 2 m to 10 m in the UT-FAD data set. The images are aligned using manually marked eye-coordinates.

2. **ADS**: Align before Down-Sample. The images of the highest resolution of ‘REAL’ are down-sampled to the same resolution as the ones captured at other distances. In the UT-FAD data set, the images from 2 m (IPD 51 pixels) are down-sampled to the same resolution as images from 3 m (IPD 34 pixels) to 10 m (IPD 9 pixels). The images are aligned before

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>REAL</td>
<td>Real low-resolution images.</td>
</tr>
<tr>
<td>ADS</td>
<td>First Align then Down-Sample.</td>
</tr>
<tr>
<td>DSA</td>
<td>First Down-Sample then Align.</td>
</tr>
</tbody>
</table>

Table 5.1: Parameter settings.

<table>
<thead>
<tr>
<th>Method</th>
<th>PCA</th>
<th>MixRes</th>
<th>CLPM</th>
<th>DSR+PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>50</td>
<td>70/60/40</td>
<td>100</td>
<td>60</td>
</tr>
</tbody>
</table>
down-sampling using eye-coordinates manually marked on the images of the highest resolution.

3. **DSA**: Down-Sample then Align. the images of the highest resolution of ‘REAL’ are down-sampled to the same resolution as the images captured at other distances. The images are aligned after down-sampling process using eye-coordinates marked on the down-sampled images.

‘ADS’ is not applicable for System B because System B does not allow manual input of eye-coordinates.

We provide verification results obtained from our experiments. Verification means to compare two images and verify if they are from the same subject. It must be remarked in advance that the verification performance for low-resolution images is poorer than on high-resolution images. This is also illustrated in [55] where the performance of the state-of-the-art low-resolution face recognition is presented. For that reason, verification rates (or genuine acceptance rate) are presented at False Acceptance Rate (FAR) 0.1 rather than the more common FAR equals to 0.01 or 0.001. Also note that the results of the first two resolutions for CLPM and DSR+PCA are not shown because, to be able to perform correctly, they require the number of training images to be larger than the dimensionality of the low-resolution images in the training set.

The verification results on the UT-FAD data set are shown in Fig 5.5.

Although the UT-FAD data set is small, there is a clear trend over most of the methods shown in Fig 5.5. Firstly, the recognition performance on pre-aligned down-sampled images (‘ADS’) is the best and there is almost no performance degradation when resolution changes for the basic recognition methods. Secondly, the results of ‘REAL’ and ‘DSA’ settings are strikingly alike for all the eight methods. The results of ‘REAL’ and ‘DSA’ settings both decrease when the resolution decreases but they remain similar at the same resolution. The above two points support our hypothesis that the poorer alignment of low-resolution images is the key factor for the performance degradation.

While the above hypothesis holds for seven methods, System A is an outlier. Its results of ‘ADS’ are similar to those of ‘REAL’ and ‘DSA’. Apparently the resolution changes are more dominant than the alignment issue for System A. Because we do not have internal knowledge about this commercial system, we cannot explain why it performs so differently than other methods. We guess that it might also include alignment compensation algorithms.

The results also show the different capabilities for low-resolution face recognition between the methods. PCA, LBP and MixRes have similar verification
Chapter 5. The importance of proper alignment in low-resolution face recognition

Fig. 5.5: Comparing ‘REAL’, ‘ADS’ and ‘DSA’ settings on the UT-FAD data set. X axis: IPD (pixels); Y axis: verification rate (VR) at FAR 0.1.

rates and most results become worse when resolution decreases (MLR and VLR) but their verification rates are all above 0.6 at FAR 0.1. RIDN significantly outperforms other methods at higher resolutions (ULR and MLR) but the results at the lowest resolution are not better than the three former mentioned methods. CLPM does not perform well for ULR and the results for VLR are also worse than most of the other methods. DSR+PCA has a reasonable performance over all the resolutions, however, DSR is actually making the performance worse when compared to the results of PCA alone. System A and System B are much more sensitive to resolution changes. System A maintains good performance for images of ULR but its performance drops significantly when the resolution decreases. For the images of VLR, it completely collapses. It is an even harder task for System B; without the aid of manual eye-coordinates, the performance of System B decreases much faster than that of System A.
5.2.3.3 Uncontrolled situation

In the previous subsection, we show support for our hypothesis that alignment is the most important cause for the face recognition performance difference between down-sampled and real low-resolution images, in absence of pose and illumination variations. The previous experiments were conducted on a data set that was captured under strictly controlled conditions. Now we are going to use images which are captured at uncontrolled situations to explore whether alignment is still important in combination with pose and illumination changes.

The Human ID data set [11] is chosen for our experiment. It has parallel gait videos that were captured when the subjects were walking towards the camera. We obtain facial images captured at different distances from those videos. The images are near-frontal. There were maximum four sessions for each subject and the intervals between sessions were more than seven days. It is captured in less controlled conditions. The poses of the captured images were mostly frontal but could change when the subjects were getting close and turning away from the camera. The viewing angle was not consistent during the recording of the videos because the camera was placed lower than the subjects’ face. The illumination was affected by sunlight thus could vary within a video or between different videos.

We use the Viola-Jones face detector (from MATLAB function) [43] to detect the faces in the videos. From the detected faces we choose images suitable for our experiments. The facial images with nine different resolution are selected: $70 \times 70$, $60 \times 60$, $50 \times 50$, $45 \times 45$, $40 \times 40$, $35 \times 35$, $30 \times 30$, $25 \times 25$, and $23 \times 23$ pixels. The corresponding average IPDs are 27, 23, 20, 18, 16, 14, 12, 10 and 9 pixels, respectively. For each resolution, two images are randomly selected from each video. The number of images for each resolution is different because some of the videos do not have images of all the nine resolutions. Images of the highest resolution (IPD 27 pixels) are used as gallery and images of other resolutions are used as probe. Compared with the previous experiments on the UT-FAD data set, the probe images are of lower resolutions that are only in the range of MLR and VLR. Detailed information about the data are shown in Table 5.3.

All the images are aligned using manually marked eye-coordinates. The regions of interest are defined by an elliptic mask. Sample images are shown in Fig 5.6.

The eight face recognition methods from the previous experiments are also used. The configurations of the methods are similar to the previous experi-
Chapter 5. The importance of proper alignment in low-resolution face recognition

Fig. 5.6: Sample images after pre-processing from the Human ID data set.

ments, only that the number of regions the images are divided into for LBP is $6 \times 6$.

The three probe settings corresponding to previous experiments are also applied on this dataset as following:

1. **REAL**: the images of IPD 23 to 9 pixels in the Human ID data set. The images are aligned using manually marked eye-coordinates.
2. **ADS**: the images of IPD 23 pixels are down-sampled to IPD 20 to 9 pixels. The images are aligned before down-sampling.
3. **DSA**: the images of IPD 23 pixels are down-sampled to IPD 20 to 9 pixels, but they are aligned using eye-coordinates marked on the down-sampled images.

The verification results are shown in Fig 5.7. Note that the images are captured at uncontrolled situation so that illumination and pose variations are present in the experiments. The resolution of the gallery images is also much lower than that from the UT-FAD data set.

Table 5.3: Details of our experimental data from the Human ID data set. Ni: total number of images. Ns: number of subjects.

<table>
<thead>
<tr>
<th>IPD</th>
<th>27</th>
<th>23</th>
<th>20</th>
<th>18</th>
<th>16</th>
<th>14</th>
<th>12</th>
<th>10</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ni</td>
<td>707</td>
<td>664</td>
<td>755</td>
<td>837</td>
<td>768</td>
<td>811</td>
<td>827</td>
<td>877</td>
<td>873</td>
</tr>
<tr>
<td>Ns</td>
<td>251</td>
<td>259</td>
<td>279</td>
<td>282</td>
<td>281</td>
<td>276</td>
<td>276</td>
<td>276</td>
<td>272</td>
</tr>
</tbody>
</table>
We observe a similar trend as in the experiments on the UT-FAD data set. The resolution changes almost have no influence on the results of the basic methods for ‘ADS’ setting. The two settings for which the images are aligned at low-resolution, ‘REAL’ and ‘DSA’, have worse results than the ‘ADS’ setting. This demonstrates that alignment still is an important factor for the face recognition performance in this uncontrolled situation. The difference to the experiments on the UT-FAD data set is that the results of ‘DSA’ are better than ‘REAL’, which is as expected because, besides alignment, other factors like pose and illumination also lead to performance loss for the ‘REAL’ setting. System A is still the outlier, but its behaviour is consistent with the previous subsection except for that the pose and poor illumination cause the performance to drop for the ‘REAL’ setting.

The capability of handling images under uncontrolled situations is also different between the eight methods. RIDN outperforms other methods by a large margin for MLR. However, its performance decreases significantly when the probe resolution decreases to the range of VLR. Especially for ‘REAL’ images, RIDN performs worse than most of the other methods at the lowest resolu-
Chapter 5. The importance of proper alignment in low-resolution face recognition

IPD 9 pixels. PCA, LBP and MixRes have similar trend. They have a reasonable performance for MLR and perform better than other methods for VLR. LBP performs slightly better among the three. Although CLPM does not perform well on higher-resolutions, its results on VLR images are at a similar level as PCA, LBP and MixRes. DSR+PCA uses PCA for recognition and thus has a very similar trend as PCA, but the fact, that it performs worse than PCA, suggests that this super-resolution method does not improve face recognition. System A is the second best method for the highest probe resolution, however, it is very sensitive to resolution changes and thus not suitable for low-resolution face recognition. System B completely collapses for the Human ID data set.

5.2.3.4 Uncontrolled situation with uncontrolled training

In the above experiments, PCA, MixRes, CLPM and DSR+PCA are trained using images from the FRGC v2.0 data set. The training images are of much higher quality than the testing images from the Human ID data set. The results are not very promising since the best result on the lowest resolution real images is only about 0.3 verification rate at FAR 0.1. To maximize the recognition capability of these methods, we use another data set for training in this subsection, the COX face data set [71]. The COX data set contains video sequences that simulate video surveillance with three different video-based face recognition scenarios. We crop facial images from these videos. We randomly select five images per subject for training in our experiments. We use images of 996 subjects. The parameters for each method are chosen to be the same as in the previous subsection. The results are shown in Fig 5.8.

As we can see, the results of MixRes and CLPM are significantly improved using COX training, but PCA results remain similar as in previous experiments. At the lowest resolution, the verification rate of MixRes increased by 0.13 at FAR 0.1. The results at the highest resolution also increased by 0.07. This makes that MixRes significantly outperforms the other methods in the previous subsection for VLR images. CLPM performs overall worse than MixRes. It could not perform well on higher-resolutions and is still sightly worse than MixRes on the lowest resolution.
5.2 Low-resolution face recognition and proper alignment

Fig. 5.8: Comparing ‘REAL’, ‘ADS’ and ‘DSA’ settings on the Human ID data set using COX training. X axis: IPD (pixels); Y axis: verification rate (VR) at FAR 0.1.

5.2.3.5 Discussion

In this section, we demonstrate that poor alignment is a very important factor that cause the loss in the recognition performance for real low-resolution images compared to down-sampled images. Our experiments on the UT-FAD data set demonstrate this for under controlled situation in absence of pose and illumination variations. The face recognition performance on pre-aligned down-sampled images remains similar to the high-resolution images. However, when we align the images after down-sampling, the performance becomes much worse and strikingly alike the performance on real low-resolution images. Furthermore, we demonstrate, using the Human ID data set, that under uncontrolled situations with pose and illumination variations, poor alignment is still one of the more important reason for the performance loss.

Our experiments also show the capabilities of different face recognition methods on different ranges of low-resolution. ULR is not a difficult task for most of the recognition methods, and commercial face recognition systems that are developed for high-resolution face recognition can still perform well. Images of MLR are harder to recognize than ULR. The performance of most of the face recognition methods, especially the commercial systems, starts to become
worse. A state-of-the-art low-resolution deep-learning face recognition method (RIDN) outperforms the other methods. VLR is extremely difficult for face recognition and all the methods give poor results. However, a simple method based on holistic features (MixRes) performs the best of them all. It outperforms commercial systems and deep-learning methods by a large margin. It can also benefit from training images of real low resolution and provides the best results for VLR and the lower range of MLR.

5.2.4 Matching-score based registration

In the previous section we demonstrated that poor alignment is a major factor that cause the loss in face recognition performance for real low-resolution images. Thus, if we improve the alignment, we would expect improvement in the recognition performance. We use matching-score based registration for this purpose.

5.2.4.1 Method

In the previous experiments we show that align low-resolution images using manually marked landmarks are not good enough for face recognition. An intuitive way for a better alignment is to use better landmark detection methods. There are many good detection methods available for high-resolution facial images. However, because low-resolution images are much more blurry and lack of details, those landmark detection methods are very likely to fail on them. For example, in our experiments, the two commercial face recognition softwares both have their own landmark detection function, but they both fail to detect their required landmarks accurately for low-resolution images.

If we have a look at the low-resolution facial images, we notice that the size of the images are very small. This has two consequences. On one hand, facial landmarks cannot be detected accurately. Automatic face detectors which are commonly used for registration usually fail to detect the landmarks on these images. And even manually marked landmarks can be inaccurate. On the other hand, because the images are so small, the possible position of the landmarks will be within a range of only a few pixels. Thus, if we have the rough locations of the landmarks, we could try to find the locations for best recognition performance by varying a few pixels around them. Therefore, matching-score based registration, which scales and aligns the images to reach maximum matching score based on crude initial estimates landmarks, can
be effective way of improving face recognition performance on low-resolution images. Its effectiveness on high-resolution images has been demonstrated in [62–64].

The process of matching-score based registration is a variation of the general face recognition procedure. The method describe here is based on the eye-coordinates because they are the most commonly used landmarks for alignment. It can be extended to more landmarks with higher computational cost. In a standard face recognition system, the probe and reference images are registered using facial landmarks, and then compared. For a probe image \( x_p \) and a reference image \( x_r \), given the eye-coordinates \( \rho \) of the probe image (reference images are assumed to be pre-aligned), the similarity score is written as \( \Delta(x(x_p, \rho), x_r) \). For matching-score based registration, the alignment of the probe image is varied so that there are several aligned images. All those images are compared with each gallery image using a classifier. The best result for each gallery image is stored as the genuine or imposter score which can be used for the subsequent verification or identification process. Matching-score based registration tries to find the eye-coordinates \( \rho^* \), that maximize the similarity between \( x(x_p, \rho) \) and \( x_r \), as:

\[
\rho^* = \arg\max_{\rho} \Delta(x(x_p, \rho), x_r)
\] (5.1)

The output similarity score is then \( \Delta^*(x(x_p, \rho^*), x_r) \).

Because the identity of the probe image is unknown, when we apply matching-score based registration, we have to pick the best result when comparing to each gallery images. As a result, all the similarity scores will increase no matter they are from the genuine or imposter pairs. There is a risk that the overall face recognition performance will become worse when imposter scores benefit more from the process than the genuine scores. In the following experiments we will explore the effectiveness of matching-score based registration.

### 5.2.4.2 Experimental Setup

The goal of our experiments is to show matching-score based registration improves the face recognition performance on the real low-resolution images. The results of ‘REAL’ and ‘ADS’ from previous experiments are used here for comparison. In addition, one more setting, REALM, is included, where Matching-score based registration is applied on the REAL low-resolution probe images.

All the face recognition methods except system B are used with the same
Chapter 5. The importance of proper alignment in low-resolution face recognition

Table 5.4: Three probe settings with matching-score based registration.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>REAL</td>
<td>Real low-resolution images.</td>
</tr>
<tr>
<td>ADS</td>
<td>First Align then Down-Sample.</td>
</tr>
<tr>
<td>REALM</td>
<td>Matching-score based registration on REAL low-resolution images.</td>
</tr>
</tbody>
</table>

settings as in previous experiments. System B is not used in the following experiments because it does not allow input of eye-coordinates so that matching-score based registration is not applicable.

The manually marked eye-coordinates which we used to align the ‘REAL’ images are used as the initial points for matching-score based registration. The range of variation of the eye-coordinates is a $5 \times 5$ pixel-region around the manually marked eyes. Within this range we search for the maximum similarity scores. This region is well balanced between face recognition performance and computational cost and it works for all the images in our experiments. However, a region of $3 \times 3$ pixels is chosen for System A and RIDN on the Human ID data set because otherwise the computing time would be too long.

5.2.4.3 Experimental results

First, we compare the results of ‘REAL’, ‘ADS’ and ‘REALM’ settings on the UT-FAD data set in Fig 5.9.

Matching-score based registration has different impact on different recognition methods. It significantly improves the performance of PCA, DSR+PCA and MixRes: results of ‘REALM’ are much better than that of ‘REAL’ and are getting close to the results of ‘ADS’. This further confirms that alignment is the major difference between down-sampled and real low-resolution images. Matching-score based registration also improves the performance of LBP and CLPM, though the improvement is not as much. However, matching-score based registration could not benefit RIDN and System A. This is probably because the imposter scores benefit more as we have mentioned in the previous subsection.

Then we present the results on the Human ID data set, see Fig 5.10. Matching-score based registration has a similar effect for each method as on the UT-FAD data set. The difference is that, for most of the methods, the results of ‘REALM’ are even better than the results of ‘ADS’ on the higher resolutions.
This confirms that poor alignment is an important problem in low-resolution face recognition and also indicates that for the Human ID data set, even the highest resolution probes do not have perfect alignment.

Finally, we present the results on the Human ID data set with COX training, Fig 5.11. The results of PCA and DSR+PCA are very similar to the results from FRGC2 training. MixRes and CLPM significantly improved by both matching-score based registration and COX training. When comparing those two, MixRes has a better performance especially at the lowest resolution. MixRes obtains 0.55 verification rate at FAR 0.1 which is 0.1 more than the second best (PCA) at the lowest resolution.

5.2.4.4 Discussion

From the above experiments on two data sets, we show that improving the alignment by means of matching-score based registration can indeed improve the face recognition performance on real low-resolution images no matter under controlled or uncontrolled situations. However, the improvement varies
Chapter 5. The importance of proper alignment in low-resolution face recognition

Fig. 5.10: Comparing ‘REAL’, ‘ADS’ and ‘REALM’ settings on the Human ID data set. X axis: IPD (pixels); Y axis: verification rate (VR) at FAR 0.1.

Fig. 5.11: Comparing ‘REAL’, ‘ADS’ and ‘REALM’ settings on the Human ID data set with COX training. X axis: IPD (pixels); Y axis: verification rate (VR) at FAR 0.1.
for different methods. For face recognition methods using global features, the improvement of matching-score based registration is significant. In the experiments using the Human ID data set, because even the alignment of the highest-resolution images is not good enough, some results of matching-score based registration are even better than the pre-aligned down-sampled images. On the other hand, RIDN and the commercial face recognition system could not benefit from matching-score based registration: there is hardly any improvement on the face recognition performance. Commercial systems are designed strictly for high-resolution images and behave unpredictably for low-resolution. Deep-learning is less dependent on alignment accuracy and therefore does not profit from matching-score based registration. Furthermore, with the help from both real low-resolution training and matching-score based registration, MixRes significantly outperforms all the other methods for VLR.

5.2.5 Conclusion

In low-resolution face recognition, the effectiveness of a new method is usually tested on images that are down-sampled from high resolution. However, face recognition on down-sampled images is much easier than on real low-resolution images. In this paper, we provide an insight into the difference between down-sampled and real low-resolution images. Surprisingly, it turns out that the degradation of face recognition performance is not caused in the first place because low-resolution image would contain less information. We demonstrate that, under controlled situations (in absence of pose and illumination variations), poor alignment is the major problem that causes the poor recognition performance on real low-resolution images; when images are captured under uncontrolled situations, alignment still plays an important role, but performance also degrades due to pose and illumination variations. The results on down-sampled images are flattered because they are usually aligned using landmarks annotated on the high-resolution images while landmarks of real low-resolution images are much less accurate. If the down-sampled images are aligned based on landmarks estimated after down-sampling, the face recognition performance appears to be similar to that obtained with real low-resolution images. In addition, we propose to use matching-score based registration to achieve better alignment and hence better face recognition performance. Matching-score based registration allows for fine tuning the registration of the poorly aligned images, and does not need accurate landmarks. Our experiments show that matching-score based registration sig-
significantly improves the face recognition performance of most of the methods, though it did not improve the performance of a deep-learning method and commercial face recognition systems. Furthermore, we divide low resolution into three classes: upper low, moderately low and very low resolution. Most face recognition methods, including commercial systems that are developed for high-resolution face recognition methods, can perform well on upper-low-resolution images. The performance of commercial systems start to become worse for moderately-low-resolution images while methods that are aimed at low-resolution face recognition (especially a deep-learning based method) still perform well. Very-low-resolution images are very difficult to recognize and the performance of the deep-learning method degrades significantly while simple holistic methods perform the best. In our experiments, a simple method based on holistic features, which can benefit from real low-resolution training and matching-score based registration, outperforms other methods for the range of very low resolution and the lower range of moderately low resolution. Finally, we recommend that if not enough real low-resolution images are available, to use down-sampled images that are aligned based on landmarks estimated after down-sampling as substitutes.

5.3 Discussion

In this chapter, we focused on research question RQ.2 about the difference between down-sampled and real low-resolution facial images. We confirmed that alignment is an important factor for face recognition on images captured no matter at controlled or uncontrolled situations (RQ.2(a)). It is the main cause of the difference in recognition performance on down-sampled and real low-resolution images, when the images are captured at controlled situations. For images captured at uncontrolled situations, other problems like pose and illumination also affect the face recognition performance on real low-resolution images, but poor alignment is still one of the major factor that cause the poor performance. We proposed to use matching-score based registration to achieve better alignment (RQ.2(b)). It significantly improved the performance of most of the methods in our experiments, but it could not improve the performance of the deep-learning based method and the commercial systems. Furthermore, we used different training sets in our experiments. The results showed that the best performance on real low-resolution facial images is obtained when the classifiers are also trained using real low-resolution images (RQ.2(c)).
Chapter 6

Low-resolution face recognition for a range of resolutions

6.1 Introduction

This chapter contains our publication [74] in 2016 International Conference of the Biometrics Special Interest Group. In this chapter, we discuss the possibility of designing a face recognition system which is optimized for a range of resolutions on the perspective of the usage of training data. We use a set of experiments to explore the influence of training data of various resolutions on the face recognition performance of testing facial images of a range of resolutions.
6.2 Designing a low-resolution face recognition system for long-range surveillance

6.2.1 Abstract

Most face recognition systems deal well with high-resolution facial images, but perform much worse on low-resolution facial images. In low-resolution face recognition, there is a specific but realistic surveillance scenario: a surveillance camera monitoring a large area. In this scenario, usually the gallery images are of high-resolution and the probe images are of various low-resolutions depending on the distances between the subject and the camera. In this paper, we design a low-resolution face recognition system for this scenario. We use a state-of-the-art mixed-resolution classifier to deal with the resolution mismatch between the gallery and probe images. We also set up experiments to explore the best training configuration for probe images of various resolutions. Our experimental results show that one classifier which is trained on images of various resolutions covering the whole range has promising results in the long-range surveillance scenario. This system has at least as good performance as combining multiple face recognition systems that are optimised for different resolutions.

6.2.2 Introduction

Face recognition at a distance is a challenging subject. The face images captured at a distance are of low-resolution so that they contain less information. In addition, the most common application for face recognition at a distance is camera surveillance, where the images are usually captured in uncontrolled situations, which results in illumination and pose variations. The images are also often noisy due to low light or compression artifacts. Those various problems suffered by the images in low-resolution face recognition make them more difficult to recognize than those in high-resolution face recognition. In this paper, we will focus on the low-resolution problem.

There are many methods that have been developed for low-resolution face recognition. Some of the methods improve face recognition performance on

---

low-resolution images by applying super-resolution to increase the image resolution. Zhang et al. [32] proposed a super-resolution method in morphable model space, which provides high-resolution information required by both reconstruction and recognition. Zou and Yuen [40] developed a data constraint for reconstructing super-resolution image features so that both the distances between the reconstructed images and the corresponding high-resolution images and the distances between super-resolution images from the same class are minimized. There are also face recognition methods that perform face recognition directly on low-resolution images. Li et al. [34] proposed a method that projects both high-resolution gallery and low-resolution probe to a unified feature space for classification using coupled mappings which minimize the difference between corresponding images. Moutafis and Kakadiaris [15] proposed a method that learns semi-coupled mappings for optimized representations. The mappings aim at increasing class-separation for high-resolution images and mapping low-resolution images to their corresponding class-separated high-resolution data. Peng et al. [55] proposed a likelihood ratio based method for direct comparison between images of different resolutions. All these works consider a specific resolution.

In face recognition at a distance, if the camera is monitoring a large area, the face images captured at a long distance can have very different resolutions. There are two approaches to deal with this situation. The first one is to improve acquisition devices so that the images captured at the farthest distance have a high enough resolution for recognition. An acquisition system using wide-field-of-view cameras and near-field-of-view cameras is proposed by Wheeler et al. in [16]. Wide-field-of-view cameras monitor the large area which detect and locate the person. Then Near-field-of-view cameras are controlled automatically to capture high-resolution face images. A commercial face recognition software is used for recognition. This system can detect faces at distances of 25-50 m and recognize faces (first successful face recognition) at distances of 15-20 m. This type of camera systems is also used in [17,18]. The second approach is to design a face recognition system for images captured at various distances. However, there is only a limited number of publications on this topic. Moon and Pan [19] proposed an LDA-based long distance face recognition algorithm. They demonstrated that using images from multiple distances for training has better recognition performance on far distances than using single near distance images for training. Tome et al. [20] proposed an estimator of the acquisition distance based on the segmented face area and the full image area. In the face verification process, the scores of DCT-GMM-based system and PCA-SVM-based system are fused based on the distance estimator so that for far distance the DCT-GMM-based system has more weight.
We follow the second approach and explore what is important in designing a low-resolution face recognition system for long range surveillance. We use a state-of-the-art low-resolution face recognition method and explore how we should set up the classifiers to cover the whole range in this scenario.

The remainder of this paper is organised as follows: we describe the scenario of the problem and propose the hypothesis in Section 6.2.3. The face recognition method we use in the experiments is introduced in Section 6.2.4. Experimental results are reported in Section 6.2.5. Section 6.2.6 gives conclusions.

6.2.3 Scenario and hypothesis

We focus on this specific but common surveillance circumstance: a surveillance camera monitoring a large area. The camera is connected to a face recognition system. When a person appears in the camera view, the face recognition system detects the person’s face and compares it with the suspects’ faces in the database. The output of the system is the decision whether this person is one of the suspects or not. In this situation, the gallery images (from the suspects) that are of high resolution. The probe images, captured at a distance, are of low resolution. Because the person can show up at any point of the monitored area, the resolution of the probe images can vary a lot.

To design a face recognition system for this scenario, we address two problems. Firstly, there is a resolution mismatch between high-resolution gallery and low-resolution probe while most classifiers are designed for high-resolution images and can only work properly for images of the same resolution. Secondly, usually a single classifier is not able to achieve sufficient recognition performance for images captured at very different distances.

To solve the first problem, we use the MixRes classifier [55]. After training with high-resolution and low-resolution image pairs, the MixRes classifier can directly compare low-resolution probe to high-resolution gallery. It is shown in [55] that this method has promising performance on very low-resolution probes.

To address the second problem, we test the following hypothesis: it is beneficial for the recognition performance of long range face recognition to combine several classifiers that are tuned to images of different resolutions. Each of the classifiers gives the best face recognition performance on images of a certain resolution. The combination of those classifiers are supposed to give optimal results across the whole range of distances.
6.2 Designing a low-resolution face recognition system for long-range surveillance

6.2.4 MixRes classifier

Here we briefly describe the MixRes classifier, which is originally named Mixed-Resolution Biometric Comparison [55]. The MixRes classifier is especially designed for comparing images captured at different distances. It is based on the likelihood ratio of a pair of mixed-resolution input images. This method is similar to [59] (also described in [75]) which is derived for comparing images of the same resolution.

This classifier transforms the reference sample $x$ and test sample $y$ to a common lower dimensional space by

$$
x_c = Z_R(x - \bar{r}) \quad \text{and} \quad y_c = Z_T(y - \bar{t}),
$$

where $\bar{r}$ and $\bar{t}$ are the grand means of, respectively, the reference and probe training sets. The transformations $Z_R$ and $Z_T$ are computed in a training phase. They are built up of a PCA step, reducing the dimensionality of probe and reference to manageable proportions as well as whitening them, followed by an LDA step that aims for optimal discrimination after transformation to a common lower dimensional subspace. This method and the training procedure are described in [55]. The log-likelihood ratio is then computed as

$$
\log(l(x_c, y_c)) = -\frac{1}{2} \sum_{i=1}^{D} \log(1 - \nu_i^2) + \frac{1}{4} s(x_c, y_c),
$$

where $\nu_i$ is the between-class covariance of feature element $i$ after the LDA step and

$$
s(x_c, y_c) = -\sum_{i=1}^{D} \frac{\nu_i}{1 - \nu_i} (x_{c,i} - y_{c,i})^2 + \sum_{i=1}^{D} \frac{\nu_i}{1 + \nu_i} (x_{c,i} + y_{c,i})^2.
$$

6.2.5 Experiments

The goal of our experiment is to test the hypothesis formulated in Section 6.2.3 that combining multiple face recognition systems, optimised for different resolutions improves the face recognition performance in the long-range surveillance scenario.

The Human ID database [11] is chosen because it is suitable for simulating the scenario as described in Section 6.2.3. This database contains high-resolution
mug shots which we use as gallery. It also contains parallel gait videos which are the best source for the probe images. There are maximum four sessions recorded at different time. We have 588 mug shots from 312 subjects. There are at most four mug shots per person. Most people have one or two mug shots. The parallel gait videos are captured when a person was walking towards the camera. Because the videos are not taken under controlled condition, there are some pose and illumination variations. We use Viola-Jones face detector [43] (implemented in MATLAB) to detect the faces in the videos. From the detected faces we choose the images of near frontal pose and relatively good quality for our experiments. From the detected faces we selected images with nine different resolutions: 70 × 70, 60 × 60, 50 × 50, 45 × 45, 40 × 40, 35 × 35, 30 × 30, 25 × 25, and 23 × 23 pixels. The distance between the eyes goes down from 28 pixels for 70 × 70 to 9 pixels for 23 × 23. For each resolution, two images are randomly selected from each video. The number of images for each resolution is different because some of the videos do not have images with all the nine resolutions. Detailed information about the data we use is in Tab. 6.1. All the images are aligned using manually marked eye-coordinates. The regions of interest are cropped using an elliptic mask. Sample images are shown in Fig. 6.1.

<table>
<thead>
<tr>
<th>Res</th>
<th>70</th>
<th>60</th>
<th>50</th>
<th>45</th>
<th>40</th>
<th>35</th>
<th>30</th>
<th>25</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ni</td>
<td>707</td>
<td>664</td>
<td>755</td>
<td>837</td>
<td>768</td>
<td>811</td>
<td>827</td>
<td>877</td>
<td>873</td>
</tr>
<tr>
<td>Ns</td>
<td>251</td>
<td>259</td>
<td>279</td>
<td>282</td>
<td>281</td>
<td>276</td>
<td>276</td>
<td>276</td>
<td>272</td>
</tr>
</tbody>
</table>

Table 6.1: Number of images and subjects of each resolution used in our experiments. Res: resolution. Ni: total number of images. Ns: number of subjects.

In order to test our hypothesis, we performed experiments, which are discussed below.

First, we test on how the classifier performs on images of different resolutions when it is trained with images of different resolutions. This is not only to test our hypothesis is correct, but also to find out if the difference is significant between the classifiers trained with different resolution images.

Cross-validation is used in our experiments. During the training procedure, we randomly selected images of 200 subjects for training. The high-resolution training images are from the mug shots and the low-resolution training images are from the video images of each resolution. Because the high-resolution training images are always from the mug shots set, we always mean the low-resolution training sets when we refer to different training sets in the remain-
6.2 Designing a low-resolution face recognition system for long-range surveillance

Fig. 6.1: Sample images of each resolution after pre-processing.

der of the paper. The images of the rest of the subjects are used for testing. There is no overlap between subjects for training and testing. After training with images of each resolution, we test the classifier using images of all the nine resolutions. The cross-validation has 100 rounds and the average verification rates (also known as genuine acceptance rates) at False Acceptance Rate (FAR) equals to 0.1 are shown in Fig. 6.2. The standard deviations of the verification rates are around 0.04.

Fig. 6.2: Verification results of training and testing with images of different resolutions. X axis: probe image resolution, Y axis: Verification Rate (VR) at FAR 0.1.
As we can see, for each probe resolution, training with images of the same resolution gives the best results. Especially when comparing the classifiers trained with images of 70 × 70 and 23 × 23 pixel-resolutions, the performance differences are significant: the difference of the verification rates on 70 × 70 probe is 0.25 and the difference on 23 × 23 probe is 0.13. The classifiers, which are trained with resolutions different from the probe resolution, usually perform worse at the probe resolution. This supports our hypothesis. We can design a system with nine classifiers, each of them dedicated to a certain probe resolution. However, some of the classifiers, which were trained with images of neighbouring resolutions, have similar performance. For example, the classifiers 70 and 60 has similar performance on probe resolutions 70 × 70 and 60 × 60. This means it could be possible to reduce the number of classifiers needed in the system.

In the second experiment we explore how many classifiers are necessary in the system. There are four different configurations of classifiers trained with image sets of different resolution divisions, shown in Tab. 6.2.

<table>
<thead>
<tr>
<th>Division</th>
<th>Resolutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIV1</td>
<td>70, 60, 50, 45, 40, 35, 30, 25, 23</td>
</tr>
<tr>
<td>DIV2</td>
<td>70-50, 45-40, 35-30, 25-23</td>
</tr>
<tr>
<td>DIV3</td>
<td>60-40, 35-23</td>
</tr>
<tr>
<td>DIV4</td>
<td>70-23</td>
</tr>
</tbody>
</table>

Table 6.2: Division of resolutions for training in the second experiment.

DIV1 has nine classifiers trained on each resolution. DIV2 has combined two or three neighbouring resolutions in the training set which results in four classifiers. DIV3 has two classifiers, each of them are trained on images of four neighbouring resolutions. DIV4 only has one classifier, but the training images have all the nine resolutions. All images in each training set are up-sampled to the highest resolution in this training set. For example, the training set of the second classifier in DIV2 consists of images of original resolutions 45 × 45 and 40 × 40. The images of original resolution 40 × 40 are up-sampled to 45 × 45 before they are used to train the classifier. In the testing phase, if the input image has a resolution of 45 × 45 or 40 × 40, they will be scaled to 45 × 45 and this (the second) classifier is used to compute scores.

To ensure a fair comparison between the four settings, we randomly select five images per subject for training for all the classifiers in the four divisions even though much more images are available in the last three settings. The results
are shown in Fig. 6.3.

As we can see, the performance on each testing resolution is similar for all the four training divisions. The differences between the average values are within their standard deviations. This means that, for a long range surveillance system, it is not necessary to have a number of classifiers optimised for different resolutions. We can use only one classifier but trained on images captured at different distances. Although the first experiment confirmed our hypothesis and showed that it is indeed beneficial to combine classifiers tuned to various resolutions, the second experiment showed that the same performance can be achieved when one classifier is trained with images with varying resolutions. The latter solution has a lower computational complexity. In addition, because there is usually not that many real low-resolution training images available, a single classifier can make the best use of the training data.

Then we compare the results of our optimised system, which is the DIV4 from the second experiment, to the results of Train70 and Train23 from the first experiment and a commercial face recognition system. We compare with this commercial system because commercial face recognition systems are used in real-life surveillance cases. The results of the commercial system are obtained using all the images available at each resolution (no training image required),
and manually marked eye-coordinates of the images are provided to the system. To make best use of the training images in DIV4, 40 images per subjects are used to train the classifier. The other configurations are the same as in the previous experiments. The verification rates at FAR equals to 0.1 are shown in Fig. 6.4.

![Comparison of DIV4, Train70 and Train23 to the commercial face recognition system.](image)

**Fig. 6.4:** Comparison of DIV4, Train70 and Train23 to the commercial face recognition system. X axis: probe image resolution, Y axis: Verification Rate (VR) at FAR 0.1.

As we can see, DIV4 performs the best across all the testing resolutions. Train70 has similar performance as the performance of DIV4 at resolution $70 \times 70$, but the difference between them becomes significant when the resolution decreases. This is similar with Train23, which has the closest result to DIV4 at resolution $23 \times 23$. The commercial face recognition system performs differently as it is designed for high-resolution face recognition. At the highest resolution $70 \times 70$, it performs only slightly worse than DIV4, but the performance dropped quickly when the resolution decreases and from resolution $40 \times 40$ to lower, the results are very close to random guessing.

### 6.2.6 Conclusion

Most face recognition systems deal well with high-resolution facial images, but perform much worse on low-resolution facial images. We focus on a specific but realistic surveillance scenario: face recognition at a distance for long range
surveillance. Our aim is to design a face recognition system for a range of resolutions. We use an existing mixed-resolution face recognition method and investigated whether it is beneficial for the recognition performance of long range face recognition to combine several classifiers that are tuned to images of different resolutions. Our experimental results show that if a classifier is only trained on images captured at a single distance, it could not perform well on the images from a very different distance. However, if we combine the images captured at various distances for training, a single classifier can perform at least as good as a combination of different classifiers when each of them are trained on images captured at a single distance. We also show that this one classifier system outperforms a state-of-the-art commercial face recognition system.

6.3 Discussion

In this chapter, we answered RQ.3 considering the design of a face recognition system that operates well on a large range of resolutions. Facial images captured at various distances and at similar conditions as the testing scenario is required for training. With this training set, a low-resolution face classifier can have optimal performance over the target range and the performance is better or at least as good as using several classifiers which are trained using images of a certain resolution.
Chapter 7

Conclusion

In this thesis, we discussed low-resolution face recognition focus on the problem of low-resolution to high-resolution comparison. In Chapter 2 we reviewed related literatures. In Chapter 3 we studied the approach of using super-resolution techniques. Though applying super-resolution to low-resolution probes makes the comparison to high-resolution galleries possible in the high-resolution space, we have shown that the super-resolution methods have no benefits for face recognition on real low-resolution images that were captured at far distances. In chapter 4, we consider direct comparison of low-resolution and high-resolution facial images and proposed a method, mixed-resolution biometric comparison. We also identified problems when researching on low-resolution face recognition. For example, most new methods are tested using down-sampled images, but the performance on down-sampled images cannot represent the performance on real low-resolution images. In Chapter 5, we continued study on the performance difference of face recognition on down-sampled images and real low-resolution images and demonstrated that inaccurate alignment is the major problem that causes the poor recognition performance on real low-resolution images. In addition, we proposed a resolution scale to harmonize the terminology in low-resolution face recognition. We define the range of low-resolution and further divide it into Upper Low Resolution, Moderately Low Resolution and Very Low Resolution. In chapter 6, we focus on a specific scenario where the probe images are of various low-resolutions. We investigated whether it is beneficial for the recognition performance of long range face recognition to combine several classifiers that are tuned to images of different resolutions. We found out that the best training set is to combine the images captured at various distances together.
In Section 7.1 we review the research questions that have been presented in Chapter 1 and provide answers to the research questions as a conclusion of this thesis. Section 7.2 presents recommendations for future research.

7.1 Answers to the research questions

In this section, we repeat the research questions and summarize the answers.

RQ.1 Which approaches exist for comparison of low-resolution probe facial images to high-resolution gallery facial images? And what is the optimal approach?

(a) Does super-resolution benefit low-resolution face recognition?

(b) Can we design a method that allows direct comparison of facial data from different domains, in particular low-resolution and high-resolution facial images, that fully exploits the information from both domains without transforming images from the one domain to the other?

Existing methods solve the resolution mismatch problem mainly by following the three approaches: applying super-resolution to low-resolution probes, down-sampling high-resolution gallery images and direct low-resolution to high-resolution comparison. In Chapter 3 we have evaluated the effect of super-resolution. Although super-resolution can increase the resolution of images and enhance the image quality visually, it has no benefit in recognition performance on real low-resolution images. The second approach, down-sampling high-resolution gallery images, has been proved to be at least as good as the first approach. Its performance on very-low-resolution images is even better than methods using super-solution techniques. This approach also has an advantage that it requires lower computational costs than using high-resolution images. However, some information in the high-resolution images, which may be useful for recognition, will be lost in the down-sampling process. Therefore, we believe that the third approach, direct low-resolution to high-resolution comparison, is the optimal approach.

In answer to sub-question(a), no, super-resolution does not benefit low-resolution face recognition. In answer to sub-question(b), yes, we proposed a new method (Chapter 4), mixed-resolution biometric comparison, to directly compare low-resolution probe images to high-resolution gallery images. This method, when trained with appropriate data, is also applicable to other types of heterogeneous biometric recognition.
RQ.2 What are the limitations of using down-sampled images instead of real low-resolution images in testing and training classifiers?

(a) What is the main cause of the difference in face recognition performance when testing on real low-resolution images and down-sampled images?

(b) How can this knowledge be exploited to improve face recognition performance on real low-resolution images?

(c) What are the benefits of training a classifier with real low-resolution images as opposed to down-sampled images?

Down-sampled images are not fully representative for realistic low-resolution images. We demonstrated this with experiments using various databases and face recognition methods in Chapter 3, Chapter 4 and Chapter 5. When down-sampled images are used as probes, the recognition performance is flattered compared to testing on real low-resolution images. Therefore, methods that have been proved to be effective on down-sampled images may have very poor performance on real low-resolution images. Real low-resolution data should be used for the evaluation of low-resolution face recognition methods to ensure their feasibility in reality. When down-sampled images are used for training, the results are worse than using real low-resolution images for training when testing on real low-resolution images. Better classifiers result if the training images have similar image quality as the testing images.

In answer to sub-question(a), we identified that alignment plays an important role in the recognition performance difference on down-sampled and real low-resolution images. We demonstrated in Chapter 5 that, under controlled situations (in absence of pose and illumination variations), poor alignment is the major problem that causes the poor recognition performance on real low-resolution images; when images are captured under uncontrolled situations, alignment still plays an important role, but performance also degrades due to pose and illumination variations. We found that the recognition performances on down-sampled images are flattered mostly due to the fact that the images are perfectly aligned before down-sampling using high-resolution landmarks while the real low-resolution images have much poorer alignment.

In answer to sub-question(b), the recognition performance on real low-resolution images can be greatly improved if the alignment becomes better. We proposed to use matching-score based registration to achieve better alignment and hence better face recognition performance. Matching-score based registration allows for fine tuning the registration of the poorly aligned images, and does not need accurate landmarks. Our experiments have shown that
matching-score based registration significantly improves the face recognition performance of most of the face recognition methods.

In answer to sub-question (c), using real low-resolution training images significantly improves the face recognition performance over using down-sampled images. Therefore, it is recommended to use real low-resolution images for training when they are available.

**RQ.3 How to design a face recognition system that operates well on a large range of resolutions?**

Face recognition methods usually perform well on images of a certain resolution, but for images of a very different resolution, the performance will drop. Nowadays, commercial face recognition systems can handle high-resolution facial images very well, but not low-resolution images. In Chapter 6 we looked into the optimal way of using training images, and investigated whether it is beneficial for the recognition performance on facial images of various low-resolutions to combine several classifiers that are tuned to images of different resolutions. We demonstrated that if a classifier is only trained on images captured at a single distance, it could not perform well on the images from a very different distance. However, if we combine the images captured at various distances for training, a single classifier can perform at least as well as a combination of different classifiers when each of them is trained on images captured at a single distance. We also show that this one classifier system outperforms a state-of-the-art commercial face recognition system. In conclusion, facial images captured at various distances and at the similar condition as the testing scenario is required for training. With this training set, a low-resolution face classifier can have optimal performance over the target range.

**RQ.4 Can we improve the performance of face recognition on low resolution surveillance quality images to the extent that it contributes to forensic investigation?**

This research question was not addressed by a specific chapter, but we can extract an answer to it through our findings. In Chapter 1, we have introduced that face recognition can be used in the three types of forensic applications: *search, intelligence investigation* and *individualization*. *Individualization* requires high accuracy and at the moment low-resolution face recognition systems do not meet the requirement. *Intelligence investigation* is to compare surveillance images from different crime scenes. This is comparison between low-resolution trace materials and in this thesis we only investigated comparison of low-resolution traces to high quality reference images. However, current
low-resolution face recognition techniques can be helpful in search. Our proposed methods have 73% verification rate at false acceptance rate 10% on very-low-resolution facial images (Table 4.10). Taking into account that the facial images are very hard to recognize for human (see Fig.4.12), with our proposed methods we can increase the efficiency of human experts by a factor of ten.

7.2 Future work

We only investigated low-resolution face recognition for mostly frontal images that are taken under somehow controlled situations. In real-life low-resolution face recognition applications, pose, illumination and facial expression may differ significantly for captured facial images. Therefore, it is important to further explore low-resolution face recognition on non-frontal images and eventually images taken under completely uncontrolled situations.

There is not enough low-resolution facial images available for research. With only small datasets, new methods cannot be properly trained and the test results cannot represent large population in real-life applications. In our experiments, we combined different sets of images for training and use cross-validation to ensure the significance of the results. However, better performance can still be obtained by using a better training set. Therefore, it is important to collect datasets with sufficient low-resolution facial images.

Another way of obtaining more low-resolution facial images is to generate realistic data from high-resolution facial images. We have demonstrated that realistic alignment is important for generating realistic data. But the effect of other factors like geometric transformation, viewing angle, noise level and compression artefacts can still be explored.

As we have mentioned in Chapter 1, we did not investigate deep-learning in low-resolution face recognition because of the lack of data. However, deep-learning technique is worth investigating for very-low-resolution facial images, especially that nowadays more datasets are published.

Better alignment methods have great potential in improvement of low-resolution face recognition performance. We proposed matching-score based registration to cope with alignment problem. It works very well with many classifiers but it could not improve the performance of some other methods, for example deep-learning methods. Therefore, a good alignment method for low-resolution facial images is demanded.
Our work focused on the problem of low-resolution face recognition where the probe images are still images. It is very common in surveillance situations that video data are available. Video sequences provide more information than a single image and therefore it can be very helpful to use this information.


[50] K. Nasrollahi and T. Moeslund, “Finding and improving the key-frames of long video sequences for face recognition,” in Biometrics: Theory Applica-


List of Publications


