Using 3D morphable models for 3D photo-realistic personalized avatars and 2D face recognition

Dianle Zhou

Thèse de doctorat de Télécom SudParis dans le cadre de l'école doctorale S&I en co-accréditation avec l' Universite d'Evry-Val d'Essonne

Spécialité :
Informatique

Par
M. Dianle ZHOU

Thèse présentée pour l'obtention du diplôme de Docteur
de Télécom SudParis

Les modèles déformables 3D (3DMM) pour des avatars personnalisables photo-réalistes et la reconnaissance de visages 2D

Soutenue le : 5 Juillet 2011 devant le jury composé de :

Prof. DAOUDI Mohamed	Telecom Lille1	Rapporteur
Prof. GARCIA Christophe	Institut National des Sciences Appliquées (INSA) de Lyon	Rapporteur
Prof. LELANDAIS Sylvie	Université d'Evry-Val d'Essonne	Examinateur
Prof. CHARBIT Maurice	TELECOM ParisTech	Examinateur
Prof. BAILLY Gérard	CNRS/Universite de Grenoble	Examinateur
Dr. Dijana Petrovska-Delacrétaz	TELECOM SudParis	Encadrant de thèse
Prof. Bernadette Dorizzi	TELECOM SudParis	Directeur de thèse

Thèse n° 2011TELE0017

Using 3D Morphable Models for 3D Photo-realistic Personalized Avatars and 2D Face Recognition

by

Dianle ZHOU

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Informatique

in the

GRADUATE DIVISION
of the
Universit Evry Val d'Essonne, TELECOM & MANAGEMENT SUDPARIS
AND L'UNIVERSITÉ D'ÉVRY-VAL D'ESSONNE

Abstract

Using 3D Morphable Models for 3D Photo-realistic Personalized Avatars and 2D Face Recognition

by

Dianle ZHOU

Doctor of Philosophy in Informatique

TELECOM & Management SudParis and l'Université d'Évry-Val d'Essonne

Dr. Dijana Petrovska-Delacrétaz and Prof. Bernadette Dorizzi

In the past decade, 3D statistical face model has received much attention by both the commercial and public sectors. It can be used for face modelling for photo-realistic personalized 3D avatars and for the application 2D face recognition technique in biometrics. This thesis describes how to achieve an automatic 3D face reconstruction system that could be helpful for building photo-realistic personalized 3D avatars and for 2D face recondition with pose variability.

The first systems we propose Combined Active Shape Model for 2D frontal facial landmark location and its application in 2D frontal face recognition in degraded condition. We extend the original Active Shape Model by using the SIFT descriptor as a new local texture model and split the facial landmarks in facial internal region and facial contour landmarks. The experimental results show that proposed Combined Active Shape Model algorithm is more robust for eyes and mouth center localization in more challenging lighting conditions, and also where some pose and expressions variabilities are presented.

The Second proposal is 3D Active Shape Model (3D-ASM) algorithm which is presented to automatically locate facial landmarks from different views. By taking advantage of 3D scans of face as training data, we propose to exploit 3D statical shape models and projective geometry across different views. The experimental results show that our proposed algorithm based on automatically generated training landmarks gives better performances than Combined Active Shape Model when large pose variation is

presented .

The third contribution is to use biomatrix data (2D images and 3D scan ground truth) for quantitatively evaluating the 3D face reconstruction. During the experiment, the proposed two automatic facial landmark location algorithms are used to initialize our automatic 3D face reconstruction on the IV2 Multimodal Biometric Database and the results are compared with manual landmarks.

Finally, we address the issue of automatic 2D face recognition across pose. We follow the strategy proposed by Blanz et al.(Blanz 2003), which based on 3D face reconstruction, but using the 3D Active Shape Model landmark detector to automatic initialize the system. The 3D Morphable Model was used as a tool for correcting the pose of 2D images prior to presenting them to a face recognition algorithm. Experiments on the PIE database showed that the approaches proposed for pose correction improved the performance of a state of the art 2D face recognition algorithm when non frontal images were used on a system trained with near frontal images only. Although the experiment results are not out performance of the state of the art algorithm, but we have demonstrated in this chapter that we have studied in detail a version of an automated 3D Morphable Model based face recognition algorithm and discussed the issues related to its success and failure.

To my parents

Acknowledgments

During the five year of my study in France, I have read a lot of thesis. Each time I read one, I dream when I can have my one and how will I write the acknowledgements. And I didn't understand why for each thesis there are the same word "To my parents.". Now I have my thesis and I have passed one third of my life. I begin to understand. They are the persons who bring me to the world, they are the person give me the life to understand.

I want to express my sincere thanks to my thesis advisor, Dr. Dijana Petrovska. Thanks for the guidance, for giving the chance to make my dream true, for the questions and for making me understand. We argued, we discussed and we tried to understand each other. That is the two different cultures and two different life. I want to say "we make the thesis" instead of "I make the thesis".

I thank my thesis director, Prof. Bernadette Dorizzi. Its been an honour and pleasure to work with her. Her continuous guidance, constant support, and invaluable advice was instrumental for the success of this work.

I would like to thank Prof. Daoudi Mohamed and Prof. GARCIA Christophe for accepting to be a reporter for this thesis. I also thank Prof. LELANDAIS Sylvie, Prof. CHARBIT Maurice, Prof. BAILLY Grard for honouring me by being a part of the jury.

I also thank Prof. Gerard Chollet for his interest in my work and the critical comments he gave which helped me improve this work.

I also thank Prof. Guangjun Zhang who is my professor in China, for his support all the time.

I thank my former colleague, Dr. Mohamed Anouar Mellakh, who was a Ph.D researcher in our group during the first year of my thesis. His comments and suggestions were quite helpful. I also thank my other former collegues, Dr. Emine Krichen, Mr. Aurelien Mayoue, for helping me during my initial days. I also thank Dr. Sanjay Kanade for his help related to kindly help for my bad English and the way to do research. I also thank my other colleagues, Nesma, Guillaume, Walid, Dr Zhenbo Li,Dr. Quoc-dinh, Dr. Patrick Horain, David Gomez, Yannick Allusse, Aurelien Mayoue and our secretary Patricia Fixot, for their help and support.

– Dianle ZHOU

Contents

List of Figures

List of Tables

List of Abbreviations

ANR	Agence Nationale de la Recherche
EER	Equal Error Rate
FAR	False Acceptance Rate
FeaLingECc	*Fea*ture *L*evel Fus*i*on through Wei*g*hted *E*rror *C*orre*c*tion
FRGC	Face Recognition Grand Challenge
FRR	False Rejection Rate
HTTPS	Hypertext Transfer Protocol Secure
NIST	National Institute of Standards and Technology
SudFROG	SudParis Face Recognition System
TLS	Transport Layer Security
ASM	Active Shape Model
AAM	Active Appearance Model
3DMM	3D Morphable Model
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
DLDA	Direct Linear Discriminant Analysis
LTM	Local Texture Model
PDM	Point Distribution Model
C-ASM	Combined Active Shape Model
3DASM	3D Active Shape Model
EBGM	Elastic Bunch Graph Matching
MES	Mean Squared Error
STD	Standard Deviation

Chapter 1

Introduction

Human faces play an important role for face recognition, video games and animated movies. Faces are associated to people, who are related to key events and key activities happening from all over the world. There are many applications using face information as the key ingredient, for example, video mining, video indexing and retrieval, person recognition and so on. However, face appearance in real environments exhibits many variations such as pose changes, facial expressions, aging, illumination changes, low resolution and occlusions, making it difficult for current state-of-the-art face processing techniques to obtain satisfactory results in all these various conditions.

Using the face for recognition has a crucial advantage, since in principle it requires no cooperation of the subject to be identified. Also, face recognition research and technology have become increasingly important for better security scenario. Face recognition systems are useful for access control in controlled applications; however significant improvements in the technology are still required before it finds its way into everyday activities, such as identity checks on automated teller machines (ATMs) or recognising offenders from public video surveillance. Furthermore, on the recently introduced biometric passports scheme by the International Civil Aviation Organization (ICAO)[1], face recognition was selected as the global interoperable biometrics for machine-assisted identity confirmation after rating highest in terms of compatibility with key operational considerations of the scheme.

Another application of face processing is the field of face modelling for photo-realistic personalized representations. Modelling the behaviour of human face in some situations, or the effects on human face within some controlled environment is among the first useful areas that come to mind considering the necessity of computer generated and animated human face models. The film industry is also using related techniques with scenes that would be very dangerous or impossible to film with real actors.

As the face is so important for communication and the human brain is very talented to recognize, it's realistic and detailed animation becomes a research area in computer graphics. We can see the results such as human body animations and talking heads. The animation of the face is mainly producing realistic facial expressions on the digital face model.

There are 2D and 3D face processing systems. In order to exploit the real

[1] www.icao.int

structure of the face, 3D data is more suitable, since the 3D nature of human face. Using 3D systems is more robust to the most critical factors limiting performance: illumination and pose variation compared to 2D system. Advantages for 3D based face processing systems are the following:

- The light collected from a face is a function of the geometry of the face, the albedo of the face, the properties of the light source and the properties of the camera. Given this complexity, it is difficult to develop models that take all these variations into account. Training using different illumination scenarios as well as illumination normalization of 2D images has been used, but with limited success. In 3D images, variations in illumination only affect the texture of the face, yet the captured facial shape remains intact.

- Another differentiating factor between 2D and 3D face processing is the effect of pose variation. In 2D images effort has been put into transforming an image into a canonical position. However, this relies on accurate landmark placement and does not tackle the issue of occlusion. Moreover, in 2D this task is nearly impossible because of the projective nature of 2D images. To circumvent this problem it is possible to use more different views of the face. This, however, requires a large number of 2D images from many different views to be collected. An alternative approach to address the pose variation problem in 2D images is either based on statistical models for view interpolation or on the use of generative models. Other strategies include sampling the plenoptic function of a face using light field techniques. Using 3D images, this view interpolation can be simply solved by re-rendering the 3D face data with a new pose.

But the 3D based system have their own problems compared to the 2D based systems:

- First is the acquisition, depending on the sensor technology used, where only parts of the face with high reflectance may introduce artefacts under certain lighting on the surface. The overall quality of 3D image data collected using a range camera is perhaps not as reliable as 2D image data, because 3D sensor technology is currently not as mature as 2D sensors' technology.

- Another disadvantage of 3D face processing techniques is the cost of the hardware. 3D capturing equipment is getting cheaper and more widely available but its price is still significantly higher compared to a high resolution digital camera. Moreover, the current computational cost of processing 3D data is higher than for 2D data .

- Finally, one of the most important disadvantages of 3D face system is the fact that 3D capturing technology requires cooperation from a subject. As mentioned above, lens or laser based scanners require the subject to be at a certain distance from the sensor. Furthermore, laser scanners require few seconds of complete immobility, while a traditional camera can capture images from far away with no cooperation from the subjects.

In order to exploit this 3D structure in different applications, building a statistic 3D face model is a good choice. Once we obtain such 3D face models they can be used in different ways:

- **2D face recognition:** A key example in the 2D face reconstruction using 3D statistical face model from single 2D image is the work from Blanz and Vetter(1999) [9]. But a lot of manual operations are needed. Such 3D models can be exploited in a generative way: thay can generate new synthesized 3D faces. We can used those 3D faces to train 2D landmark detectors which are robust to pose variation in order to avoid manual labelling of landmarks for better 2D face recognition. This is important for application where automatic facial landmarks detection and face recognition systems are needed. In chapter 5, we are trying to solve 2D face recognition problem with the priori of 3D knowledge of face. One challenging problem in 2D face recognition is the large pose variation on the face images. One way to solve this problem is the technique by Blanz et al. The human faces can be treated as a manifold surface in a 3D space. The 3D Morphable Model (3DMM) for face image synthesis and face recognition is developed by Blanz et al. [9, 10]. One advantage of the 3D morphable face model is that it can easily handle variations on pose and illumination instead of 2D models. The variance of pose and illumination is always an obstacles for face recognition in 2D space. Another advantage of the 3D Morphable Model is that a 3D face surface is extracted from a single 2D face image, which avoids expensive 3D face/head

scan. Face recognition uses the shape and texture parameters of the model, which represent intrinsic information of faces. We can exploit 3D Morphable Model to reconstructed the 2D frontal image from the 2D nonfrontal image by using the priori of 3D knowledge of face.

- **Photorealistic personalized representation:** The 3D realistic avatar reconstruction (i.e. automatic 3D face reconstruction from a 2D image) is a research area overlapping with computer vision, computer graphics, machine learning and Human-Computer Interaction (HCI). 3D face processing techniques are useful for (1) extracting information about the person's identity, motions and states from images of face in arbitrary poses; and (2) visualizing information using synthetic faces for more natural human computer interaction. A general statement of the problem of 3D photorealistic personalized representation reconstruction can be formulated as follows: given still 2D image of a scene, the face is extracted from the image and reconstructed to be rotated and manipulated in 3D. The solution to the problem involves segmentation of faces (i.e. face detection) from clustered scene and localization of landmark points from face regions. This step contains the procedure of initialization of the 3D generic face model on the 2D image. In 3D face reconstruction step, the 3D statistic face model are modelled from measurements of faces, such as 3D range scanner data (i.e. 3D scans, 3D geometry and texture of neutral face) or images (i.e. 2D images or stereo images). Then, the 3D statistic face model is deformed according to the face in the 2D image to reconstruct a 3D face.

We are interested in this thesis in the reconstruction of 3D realistic face avatar from a 2D image. Given a single photograph of a face, we would like to estimate its 3D shape and texture by using 3D Morphable Model, its orientation in space and the illumination conditions of the scene. The face model created from the image can be then rotated and manipulated in 3D.

All the previous statements show the usefulness of 3D statistic model for face processing for a bunch of different applications. And it is also an actual problem and new proposal for quantitative evaluation on 3D face reconstruction are needed. In this thesis we evaluate the 3D face reconstruction in two different ways:

- Quality evaluation: Taking the image and the 3D scan from the same subject, the 3D reconstruction precision could be evaluated by computing the geometric distance between the reconstructed 3D face and the ground truth (3D scan from the same subject).

- Indirect evaluation: The 3D face reconstructed algorithm also could be evaluated by 2D face recognition. As we discussed before, the 3D Morphable Model based 3D face reconstruction could be exploited to solve the 2D face recognition across pose problem. The better 3D face reconstruction precision we can achieved the better face recognition performance we can obtain.

1.1 Thesis Outline

This thesis is organised as follows: Chapter 2 presents the proposed Combined Active Shape Model for 2D frontal facial landmark location and its application in 2D frontal face recognition in degraded condition. In Chapter 3, we study the 2D facial landmark location on nonfrontal images, and details about the new construction, training and fitting of the proposed 3D Active Shape Model are given. It can be used for the initialization step for the automatic 3D face reconstruction. Since the training data of the 3D Active Shape Model are generated from the 3D Morphable Model, using it for initialization could benefit the 3D face reconstruction. In Chapter 4, our automatic 3D face reconstruction method is quantitatively evaluated with IV^2 multimodel biometric database [50], by exploiting both 3D scans and 2D images. In Chapter 5, the methodology of using the 3D Morphable Model to solve the 2D face recognition across pose problem is studied. The final chapter concludes the work of the thesis, and highlights directions for future work.

Chapter 2

Automatic 2D Facial Landmark Location with a Combined Active Shape Model and Its Application for 2D Face Recognition

2.1 Introduction

Finding the correct position of facial **landmarks** (key points) is a crucial step for many face processing algorithms such as face recognition, modelling or tracking. It is also needed for a variety of statistical approaches in which a model is built from a set of labelled examples. Also many 2D face recognition algorithms depend on a careful geometric normalization, with the location of landmarks such as eyes and mouth centres, that is previous to the global feature extraction step. With more reliable and more precise landmarks better face recognition performance is achieved.

Depending on the application context, face recognition can be divided into two scenarios: face verification and face identification. In face verification, an individual who desires to be recognised claims an identity, usually through a personal identification number, an user name, or a smart card. The system conducts a one-to-one comparison to determine whether the claim is true or not, *i.e.*, face verification is to ask a question - "Does the face belong to a specific person?". In face identification, the system conducts a one-to-many comparison to establish an individual's identity without the subject to claim an identity, *i.e.*, face identification is to answer the question - "Whose face is this?". Throughout this thesis, the generic term *face recognition* is also used, which does not make a distinction between verification and identification.

The number and position of facial landmarks are not unique and depend on applications and algorithms. For 2D face recognition with global methods, usually eye centres, nose and mouth positions are needed. While for Active Shape Model (ASM) introduced by Cootes et al. in 1995 [19] approaches, the number of landmarks is bigger (around 50). They are located in regions of the nose tip, the nostrils, the center (iris) and corner of eyes, the mouth corners, the eyebrows and the tip of the chin. Those landmarks can be labelled by hand, but for realistic applications it is necessary to have automated methods. Due to the variety of human faces and their variability related to expressions, pose, accessories, or lighting and acquisition conditions, fully automatic landmark localization remains a challenging task.

This chapter focuses on automatic facial landmark location for face recognition, in situations where mainly illumination, scale and small pose variabilities are present. We are interested in automatic facial landmark location in 2D images for two purposes:

- To fully automate our 2D face recognition system.
- For 3D face reconstruction from 2D images.

The rest of this chapter is organized as follows: first, a brief literature review about facial landmark location is given in Section 2.2. Then a reminder of the original Active Shape Model (ASM), on which our proposed combined model is based on, is given in Section 2.3. The proposed Combined Active Shape Model, denoted as C-ASM, is explained in Section 2.4. We evaluate the precision of the landmark detection in two ways. In Section 2.5, we compare the detected landmarks with ground truth (manually annotated) landmarks. As we are interested in face recognition, in Section 2.6, we use the C-ASM to do automatic landmark location for 2D face recognition. Finally, the conclusions related to this chapter can be found in Section 2.7.

2.2 Literature Review about Automatic 2D Landmark Location

A lot of algorithms have been proposed for facial landmark location for 2D images. As suggested by Hamouz et al.(Hamouz 2005) [30], they can be classified in two categories: image-based and structure-based methods.

In **image-based methods**, faces are treated as vectors in a large space and these vectors are furthermore transformed. The most popular transformations are Principal Components Analysis (PCA), Gabor Wavelets (Fasel 2002, Vukadinovic 2005) [24, 55, 70], Independent Components Analysis (Antonini 2003) [3], Discrete Cosine Transform (Salah 2006) [55], and Gaussian Derivative Filters (Arca 2006, Gourier 2004) [4, 26]. Through these transforms, the variability of facial features is captured, and machine learning approaches like boosted cascade detectors (Viola 2001) [69], Support Vector Machines (Chunhua 2008) [21] and Multi-layer Perceptions are used to learn the appearance of each landmark. Some examples of such methods are proposed by Viola and Jones [69], Jesorsky et al. [34], and Hamouz et al. [30].

Structure-based methods use prior knowledge about facial landmark positions, and constrain the landmark searching by heuristic rules that involve angles, distances, and areas. The face is represented by a complete model of appearance consisting

of points and arcs connecting these points (Shakunaga 1998) [58]. For each point of this model, a feature vector is associated. Typical methods include Active Shape Models (ASM) (Cootes 1995, Ordas 2003) [19, 47], Active Appearance models (AAM) (Cootes 2004) [20], and Elastic Bunch Graph Matching (Wiskott 1997, Monzo 2008) [45, 73]. These methods are well suited for precise localization (Milborrow 2008) [44].

Within structure-based models, one outstanding approach is the Active Shape model (ASM) [20], because of its simplicity and robustness.

2.3 Reminder about the Original Active Shape Model (ASM)

The original Active Shape Model (ASM) was introduced by Cootes et al. in 1995 [19]. It is a model-based approach in which the priori information of the class of objects to is encoded into a template. Such template is user-defined and allows the application of ASM to work on any class of objects, as long as they can be represented with a fixed topology, such as faces.

The face template can be considered as a collection of contours, each contour being defined as the concatenation of certain key points defined in the shape analysis literature as landmarks, see Figure 2.1. The deformation of the landmarks allowed in the model template is learnt from a training database. As a result, an important property of ASM is that they are generative models. That is, once trained, ASM are able to reproduce samples observed in the training database and, additionally, they can generate new instances of an object not present in the database but consistent with the statistics learnt there from. In the original Active Shape Model (ASM) introduced by Cootes et al. in 1995 [19], there are two statistical models that exploit the global shape and the local texture prior knowledge in the segmentation process.That is Point Distribution Model (PDM) and Local Texture Model (LTM). The PDM represents the mean geometry of a shape and its statistical variations from the training set of shapes. While the LTM is used to describe the texture variations at each landmark position of the PDM.

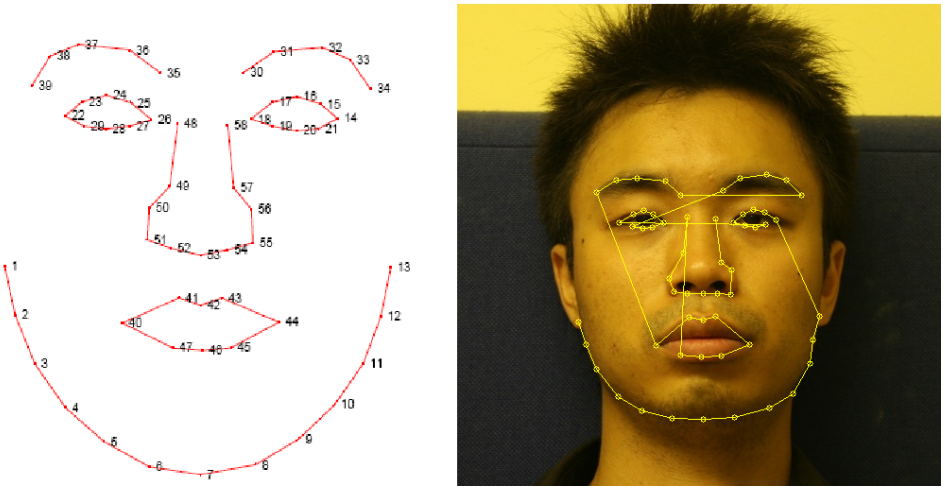

Figure 2.1: Landmark positions and the contours of a 58-points template for facial analysis.

2.3.1 Point Distribution Model

In order to construct the PDM, there is a need for a training set. The training set consists of a set of images, which represents the object class to be modelled. And those images should be annotated with the predefined template. The set of annotated landmarks on one image is referred to as the shape associated to that image.

The PDM is constructed by applying Principal Component Analysis (PCA) to the set of shapes in the training set. It is generally preceded by a 2D alignment in order to make the analysis independent from 2D rotation and scaling variations. Indeed, shape is usually defined as all the geometrical information remaining when positional, scaling and rotational effects have been filtered out from an object.

The Point Distribution Model is constructed by applying Principal Component Analysis (PCA) to the aligned set of shapes, which are presented by landmarks on the training face database. Assume there are N training images. The Point Distribution Model is a linear model, so the i^{th} shape S_i and the model parameters P_i in the shape space can be represented as follows:

$$P_i = \Phi^T(S_i - \overline{S}), \quad S_i = \overline{S} + \Phi P_i, \tag{2.1}$$

where $i = 1, ..., N$. $\overline{S}$ is the mean shape, and Φ is the eigenvector matrix of the shape space Briefly, the Point Distribution Model describes heuristic rules of the face shape. During the fitting, this model helps in the interpretation of noisy and low-contrasted

pixels.

2.3.2 Local Texture Model

As stated before, ASM have as many local texture models as the number of landmarks in the template. A typical image structure that describes the local texture around each landmark is the Grey-Level Profile (GLP) [19], calculated from the fixed-length pixels sampled around each landmark. The direction of the profile is perpendicular to the contour. The first derivative of the profile is calculated and used as the feature vector. Those vectors are extracted from all the training images, and represent the normalized derivatives profiles, denoted as $g_1, g_2, ..., g_N$. The mean profile $\overline{g}$ and the covariance matrix C_g are computed for each landmark. The Mahalanobis distance measure is used to compute the difference between a new profile and the mean profile $\overline{g}$, defined as follows:

$$Mh^2(g_{new}) = (g_{new} - \overline{g})C_g^{-1}(g_{new} - \overline{g})^T. \tag{2.2}$$

Actually different Local Texture Models are adapted to different conditions, the examples are introduced in the following Subsection 2.3.3. The dimension of the GLP is depended on the number of fixed-length pixels sampled around each landmark, the details about the parameters will be explained in Section 2.5.

2.3.3 Matching algorithm

As explained in Cootes et al. [20] and Sukno (Sukno 2007) [65], when the shape models are used for segmentation and landmark location, only two inputs are required: an image containing a face and a starting guess of the face position (i.e. provided by a face detector). The matching process alternates image driven landmark displacements and statistical shape constraints based on the PDM, usually performed in a multi-resolution fashion in order to extend the capture range of the algorithm. The matching process can be summarized in the following steps:

1. Place a first guess of the model into the image (generally, a scaled version of the mean shape, depending on the application task).

2. Search the image in the neighbourhood of each landmark. Adjust the coordinates of each landmark to the best position in this neighbourhood. In other words: move the landmarks according to their LTM. This will generate a cloud of points without shape constraints.

3. Apply shape constraints: find the best plausible shape matching the cloud of points generated in step 2. This implies finding the model parameters and some transformation (e.g. a similarity) from model coordinates to image coordinates. The $S_{constrain}$ parameter restricts the PCA coefficients to lie within $S_{constrain}$ (For example,$S_{constrain} = 3$) times the standard deviation observed in the training set.

4. Go back to step 2 until stop condition is reached.

The criterion used to displace the landmarks at step 2 is the minimization of the Mahalanobis distance based on the Gaussian model learnt during training for each LTM. Let $\{g_j(1), g_j(2), ..., g_j(k_P)\}$ be the set of local texture points for k_p candidate positions at landmark j. The position suggested by the LTM will be the one minimizing:

$$Mh^2(g_j(k)) = (g_j(k) - \overline{g_j})C_g^{-1}(g_j(k) - \overline{g_j})^T \tag{2.3}$$

for k varying between 1 and k_p, where Mh^2 denotes Mahalanobis distance Once all landmarks have been displaced to their best local position, they form a cloud of points which not necessarily describe a plausible shape for the studied object (i.e. a human face).

At step 3, shape restrictions are applied according to the PDM. As a result, landmarks are displaced again to the nearest plausible shape to the candidate points provided by the appearance models (in a least squares sense). The rationale behind shape restrictions is the assumption that facial shapes lie approximately within a hyper ellipsoid (in PCA-space) that can be learnt during training. However, for simplicity reasons it is very common to use PDM and limit the shape-space to a hyper-cuboid.

Starting from the original formulation of ASM introduced above, a considerable number of extensions have been proposed. One of the most interesting aspects of the original formulation of ASM is its simplicity. For example, the residuals of the shapes with respect to the mean are assumed Gaussian. This formulation works well for a wide variety of examples, although it is too simple to represent nonlinear shape variations.

Non linear formulations of the PDM were proposed by Sozou et al. in 1995 [62] using Multi-Layer Perceptron (MLP) to perform the PCA decomposition. The experiments reported on image search revealed comparable performance to the linear PDM, yet requiring half the number of dimensions. Chen et al. in 2004 [18] focused on a different aspect of the PDM. They decomposed the overall error of ASM fitting into two terms: representation error and search error. They analysed the behaviour of the error as a function of the variance explained by the model. Based on experiments over 400 faces they claim that the optimal percentage of variance retained by the model is lower than that generally employed.

As opposed to those modifying the PDM, a number of authors have focused on the texture model of ASM. Wang et al. in 2002 [71] combined the first order derivatives with edge information to work with facial images and Koschan et al. in 2002 [37] explored inclusion of color information. While Ordas et al. in 2003 [47] replace the 1D normalized first derivative profiles of the original ASM with local texture descriptors calculated from "locally orderless images", for reliable segmentation for cardiac Magnetic Resonance data. In Milborrow et al. in 2008 [44], the authors use the 2D profile in the square region around the landmark for a more precise fitting result.

However most of the research in this topic has concentrated in improving the precision of landmark detection in "passport" like images. As we also are interested in noncollaborative environments, we propose two extensions to increase the robustness under degraded conditions. First, we replace the gray-level profiles with SIFT features as LTM; second in order to have a better representation of faces, the landmarks on the face region and the face contour are modelled and processed separately for the PDM.

2.4 A Combined Active Shape Model for Landmark Location

2.4.1 Using SIFT Feature Descriptor as Local Texture Model

Some previous work exist that try to choose better Local Texture Models. Different Local Texture Models are adapted to different conditions. For example, using the gray-level profiles is simple and fast and suitable for real-time processing, but it is less precise. While the 2D profile in the square region around the landmark used in

[44] is precise in controlled conditions. As we are more interested in non collaborative environments, such as video-based face recognition, we propose to use the SIFT [?, 42, ?, ?] feature descriptor, which is robust to degraded conditions, such as illumination or small pose variations.

Scale Invariant Feature Transform (SIFT) Features

In 2004, David Lowe presented a method to extract distinctive invariant features from images [42]. He named them Scale Invariant Feature Transform (SIFT). The process consists of four major stages: (1) scale-space peak selection, (2) key point localization, (3) orientation assignment, and (4) key point descriptor. In the first stage, potential interest points are identified by scanning the image over location and scale. This is implemented efficiently by constructing a Gaussian pyramid and searching for local peaks (termed key points) in a series of Difference-of-Gaussian (DoG) images. In the second stage, candidate key points are localized to sub-pixel accuracy and eliminated if found to be unstable. The third step identifies the dominant orientations for each key point based on its local image patch. Finally, local image descriptors are built for each key point. Local gradient data is used to create key point descriptors. The gradient information is rotated to line up with the orientation of the key point and then weighted by a Gaussian with variable scale. This data is then used to create a set of histograms over a window centred on the key point. Key point descriptors typically use 16 histograms, aligned in a $4x4$ grid, each with 8 orientation bins, one for each of the main compass directions and one for each of the mid-points of these directions, resulting in a feature vector containing 128 elements. In our application we use this SIFT local image descriptor as our Local Texture model. In this thesis we implement the SIFT descriptor ourself.

Advantages of the SIFT Feature Descriptor

Using SIFT features for object matching is very popular, because of SIFT's ability to find distinctive key points that are invariant to location, scale and rotation, and robust to affine transformations (changes in scale, rotation, shear, and position) and changes in illumination [42]. It seems to be a reliable choice for solving the problem of illumination and pose variability during the facial landmarks location. Since it is based

on the local gradient histograms around the landmark, the SIFT descriptor is highly distinctive and partially invariant to variations, like illumination or 3D view point, as introduced in [42]. In our application, we use the SIFT descriptor to replace the Grey-Level profiles. In order to make the ASM shape model rotation invariant, the gradient orientations of the descriptor are always computed relative to the edge normal vector at the landmark point which could be obtained by interpolation of neighbouring landmarks, as depicted in Figure2.2.

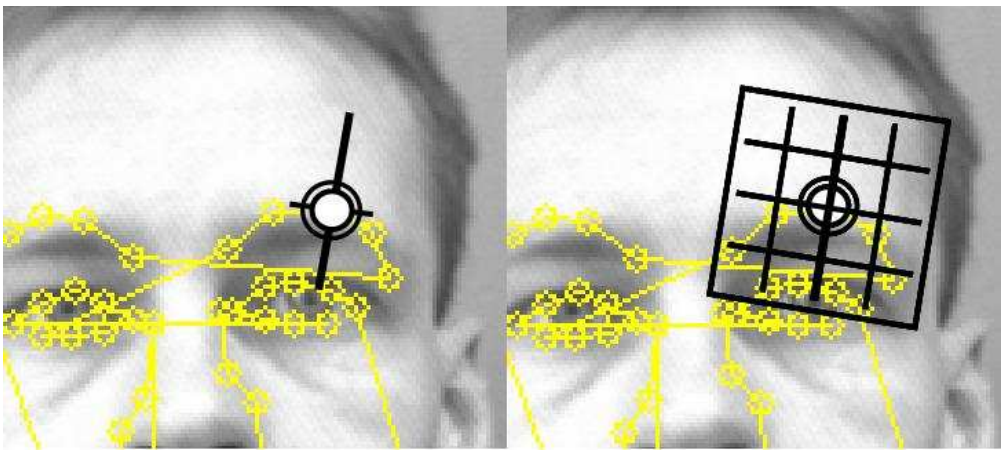

Figure 2.2: Difference of the Grey-Level Profile and the SIFT descriptor. Left: The Grey-Level Profile (GLP) is extracted from the neighbourhood pixels perpendicular to the contour. Right: The SIFT descriptor is computed over a patch along the normal vector at the landmark (the original image is from the BioID database [34]).

There are the two main advantages of the SIFT feature descriptor. The first advantage is that SIFT descriptors encode the internal gradient information of a patch around the landmark, thus capturing essential spatial position and edge orientation information of the landmark while Grey-level Profile only captures the one-dimensional pixel information that is perpendicular to the contour. Though the Mahalanobis distance measure assumes a normal multivariate unimodel distribution of Grey-level Profile, in practice, they can be any statistical distribution. The SIFT descriptors have a more discriminative likelihood model which is distinctive enough to differentiate between landmarks.

In Figure. 2.3, we calculated the Mahalanobis distance of the neighbourhood points over a $21x21$ pixels window around the landmark highlighted in Figure. 2.2. The SIFT descriptor has a unambiguous minimal point in the center of the neighbourhood

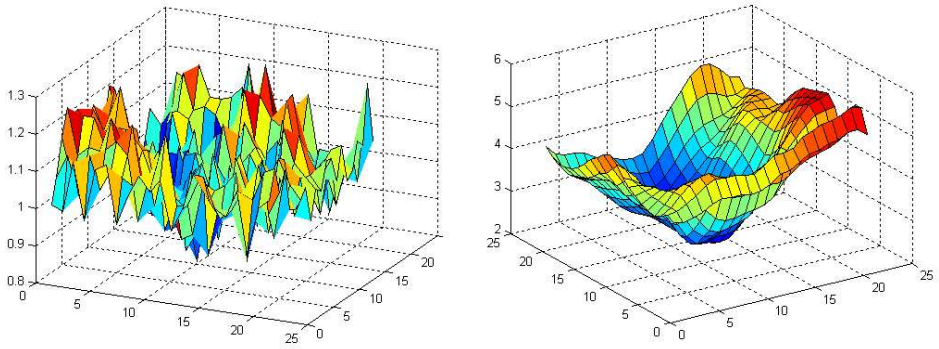

Figure 2.3: Comparison of the Grey-Level Profile and the SIFT descriptor cost function from Eq 2.3. Left: Gradient profile matching cost of the landmark highlighted in Figure 2.2 over a window of size 21x21. Notice the multiple minima resulting in poor alignment of shapes. Right: SIFT descriptor matching cost for the same landmark point.

region, while the results of GLP contains more noises. Also the SIFT descriptors are invariant to affine changes in illumination and contrast by quantizing the gradient orientations into discrete values in small spatial cells and normalizing these distributions over local blocks. Such features are important in challenging real-life situations presenting illumination variabilities.

The second advantage of the SIFT descriptors is that they are more stable to changes that occur due to changes of pose, that can occur when dealing with faces .

2.4.2 Combined Active Shape Model (C-ASM) Based on Facial Internal Region Model and Facial Contour Model

One of the novity of our work is applying different feature descriptor for different landmarks on the faces.As shown above, using SIFT feature descriptor, we can find correspondences between landmarks in two images that have small pose variability, even when the landmarks used to train the ASM are in 2D. The points in the face region that we denote as “internal” (such as eyes’ corners), could be considered as the perspective projection of 3D face on the image plan. While the contour points are different, and are more dependent on the 3D view point. In that case the SIFT descriptor dose not work when the acquisition angle of testing images is different from the training images. Especially when those points are occulted because of minor head pose rotation, see Figure 2.6.

So in our proposed approach, two models are used to represent the human face. One model represents the landmarks of what we call "internal region", including the landmarks on the eyes, nose, eyebrows and mouth. Those points could be considered as 3D position invariant during perspective projections. So we use the SIFT descriptor for this model, and we name it **facial internal region model**. The other one models the contour point on the face only. For those points using SIFT representations will result in wrong matches. The gradient of the profile is more suited for the contour points, so we use Grey-Level Profile to describe them, and we name it **facial contour model**.

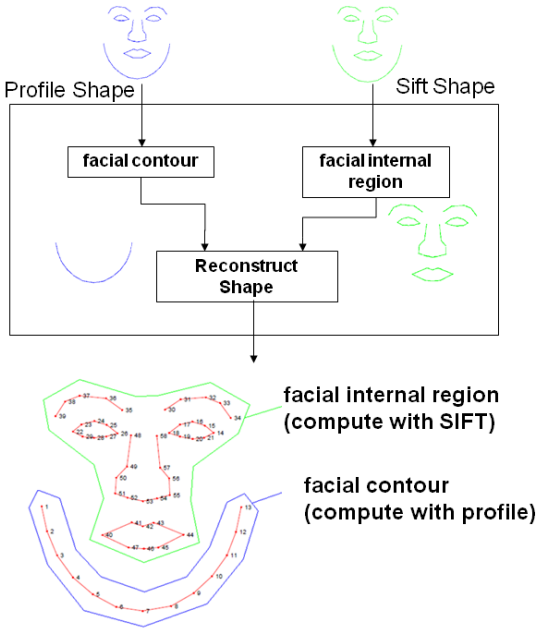

Figure 2.4: Combined landmark detection model: 45 landmarks define the facial internal region model (represented with SIFT features) and 13 landmarks define the facial contour model (represented with GLP features).

The facial internal region model is represented with 45 points, while 13 points are used for the facial contour model, as depicted in Figure. 2.5. Each of them has its own shape model and shape variability. The combination of the two models makes the final results.

In the fitting step, for each iteration the two models are matched to the face image separately. After matching, we combine them into a new face model and use the Point Distribution Model to constrain it to a plausible shape in the shape space,

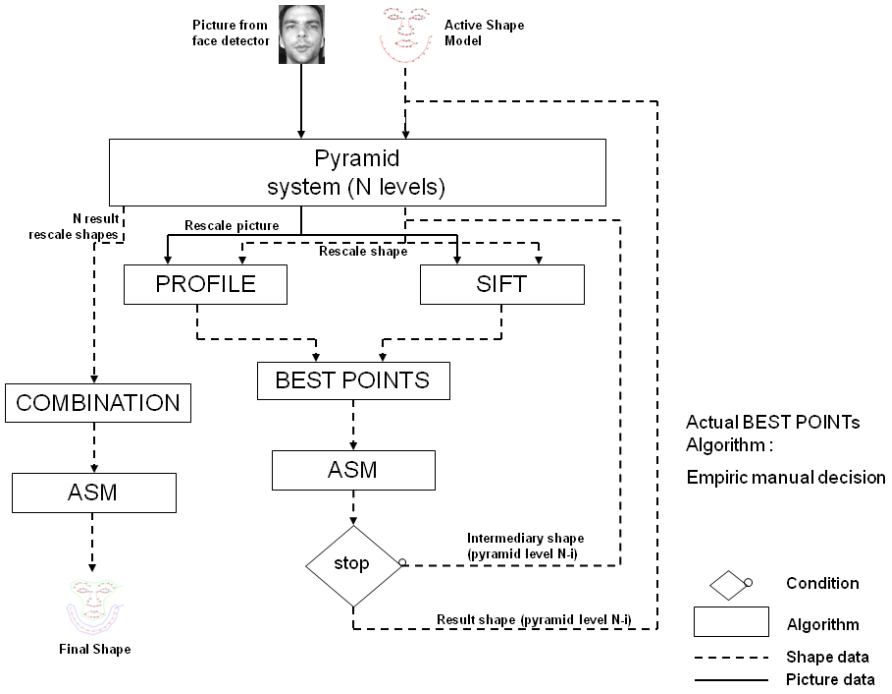

Figure 2.5: Combined Active Shape Model marching algorithm flow chart.

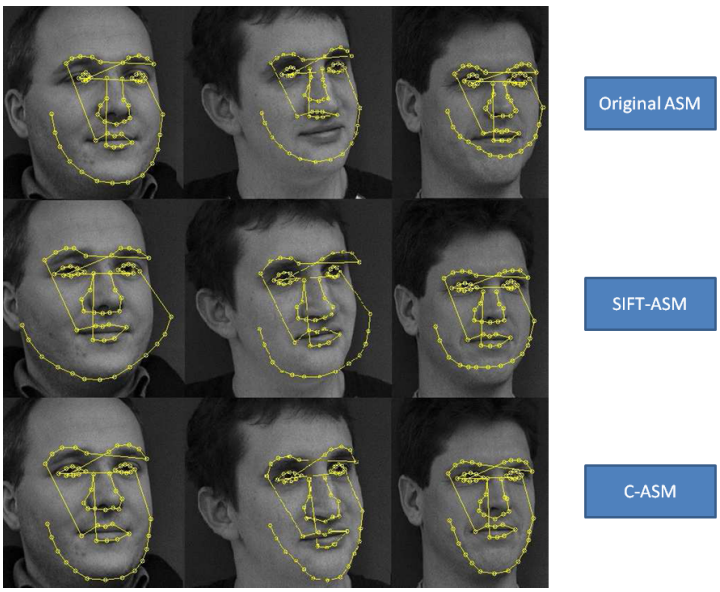

Figure 2.6: Typical fitting result of non frontal faces achieved by original ASM (top row), SIFT-ASM (middle row) and C-ASM (bottom row). (The original images are from the IMM database [63]).

as shown in Figure. 2.4. This is repeated iteration by iteration at each resolution until convergence is reached.

In Figure. 2.6, for comparison purposes, we show on the first line the results of the automatic landmark detection with the original ASM (using only profile features), the ASM that is based on SIFT features (middle row), and the result of our Combined Active Shape Model (C-ASM). We can observe that the Grey-Level Profile has better performance on the contour points of side-view images, while SIFT features seem to be more adapted for the internal face region points. But we should note that, the idea to use the Gray-level Profile as the LTM for the facial contour is not only to improve the precision of the detected landmarks on the contour, but also to constrain the contour points around the mouth. This will increase the robustness of the landmarks on the facial internal region such as mouth and eye, as shown in Figure 2.6, which are the points needed for face normalisation in the applications of face recognition. The experimental results for face recognition are given in Section 2.6.

2.5 Experiments for C-ASM Landmark Location Precision Evaluation

In this section, we compared the detected landmarks with ground-truth (manually annotated) landmarks. We are mostly interested in automatically detecting the two eyes and mouth centres which are used for our face normalization step for our face recognition system explained in Mayoue et al. in 2009 [48].

2.5.1 Experimental Protocol for Landmark Location Precision

In this section, we will introduce the databases we used for training and evaluation of our C-ASM for landmark location, the evaluation criteria and the parameters of the C-ASM. Also we will briefly explain the existing methods with which we compared our results.

Training Database

The IMM Face Database [63] comprises 240 still images of 40 different human faces, all without glasses. The gender distribution is 7 females and 33 males. Images

were acquired in January 2001 using a 640x480 JPEG format with a Sony DV video camera, DCR-TRV900E PAL. The following facial structures were manually annotated using 58 landmarks: eyebrows, eyes, nose, mouth and jaw. A total of seven point paths were used; three closed and four open. The landmark's positions and contours are shown in Figure 2.7.

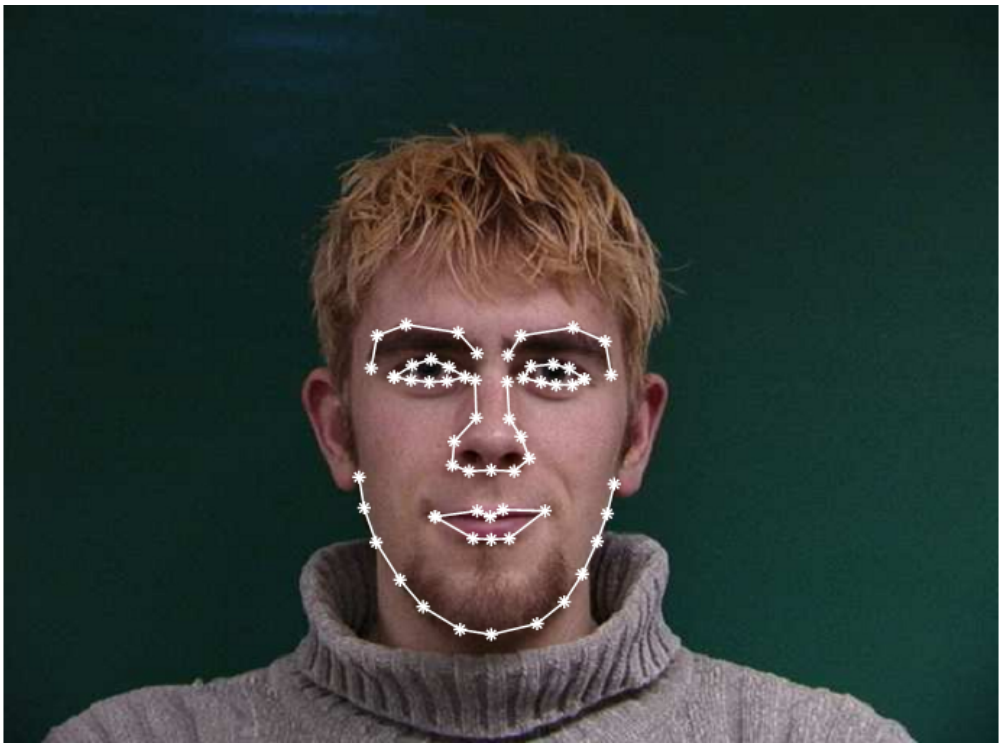

Figure 2.7: Annotated face image from the IMM face database [63].

Evaluation Database

The BioID dataset consists of 1521 gray level images with a resolution of 384 × 286 pixels. Each one shows the frontal view of a face of one out of 23 persons. The number of images per subject is variable, as is the background (usually cluttered like in an office environment). The positions of the eyes are provided.

For our experiments on face landmark location, a subset of the FRGCv2.0 (Face Recognition Grand Challenge version 2) face database [51] is selected. The full FRGCv2.0 database contains images from 466 subjects and is composed of 16,028 controlled still images captured under controlled conditions and 8,024 non-controlled still

images captured under uncontrolled lighting conditions. There are two types of expressions: smiling and neutral and a large time variability exists. The positions of eyes, mouth and nose centres are manually labelled.

Evaluation criteria: Because we are interested in face recognition, we evaluate in this chapter only the points which we use for our face normalization step [48]. These points are the centers of the eyes and the mouth center. In order to be able to evaluate the landmarking methods a well-defined error measure is required. Since the images in the databases are of various scales, the measure that was proposed by Jesorsky et al. [34] is used, where the localization criterion is defined in terms of the eye center positions:

$$d_{eye} = \frac{max(d_{lefteye}, d_{righteye})}{||C_l - C_r||} \tag{2.4}$$

where C_l, C_r are the ground truth eye center coordinates and $d_{lefteye}, d_{righteye}$ are the distances between the detected eye centres and the ground truth ones. In the evaluation, we treat localizations with d_{eye} above 0.05 as unsuccessful. Mouth center is evaluated in the some way but normalized with the distance ($d_{eyeC,mouth}$) between the average point of two eyes $C_{twoeyes}$ and mouth center C_{mouth} from ground truth:

$$C_{eyeC} = \frac{C_l + C_r}{2}, \; d_{mouth} = \frac{d_{eyeC,mouth}}{||C_{eyeC} - C_{mouth}||} \tag{2.5}$$

It has to be noted that prior to the landmark location step, we apply a face detection algorithm in order to have a rough location of where the face is located. We use the AdaBoost approach proposed by Viola and Jones in 2001 [69], freely available from the OpenCV library introduced by Bradski in 2005 [11]. After the face region is located, we scale it to region of 260x260 pixels, so its size is similar to the training data of the Combined Active Shape Model.

Experimental Parameters: As explained in Section 2.4, for the Combined ASM model, we use the 58 landmarks (present in the IMM database), divided into 45 landmarks for the internal facial region and 13 points that belong to the facial contour regions. As eyes' and mouth centers are not present among these 58 landmarks (see Figure. 2.4), we calculate them by averaging the landmarks detected around the eyes and mouth. We use coarse to fine search over 2 levels of Gaussian scale pyramid. The SIFT block contain $4x4$ cells with $4x4$ pixels and 8 gradient orientation bins, thus having descriptor size of 128. The length of the grey level profiles is set to be 17 pixels.

Comparison with STASM: For comparison purposes, we use the publicly available STASM software [44], developed by Stephen Milborrow. The STASM method extended the original ASM by using 2D profile among the points and more landmarks, during the training step (using annotated images from the XM2VTS database). We denote it as STASM-original. The effect of the number of landmarks on the detection performance is out of scope of this work. To be able to make a fair comparison, we also trained STASM with the same training data from the IMM database that we are using. We denote it as STASM-modify.

2.5.2 Experimental Results for Landmark Location Precision

For the evaluation of our automatic landmark detection algorithms we use the BioID and FRGCv2.0 databases. The BioID database is chosen because there are already published results on that database for facial landmark detection, while the FRGCv2.0 was chosen because it includes a huge number of subjects (around 500) with variabilities including illumination, pose and expression, and because the ground truth position of eyes, and mouth is available.

Evaluation on the BioID Database

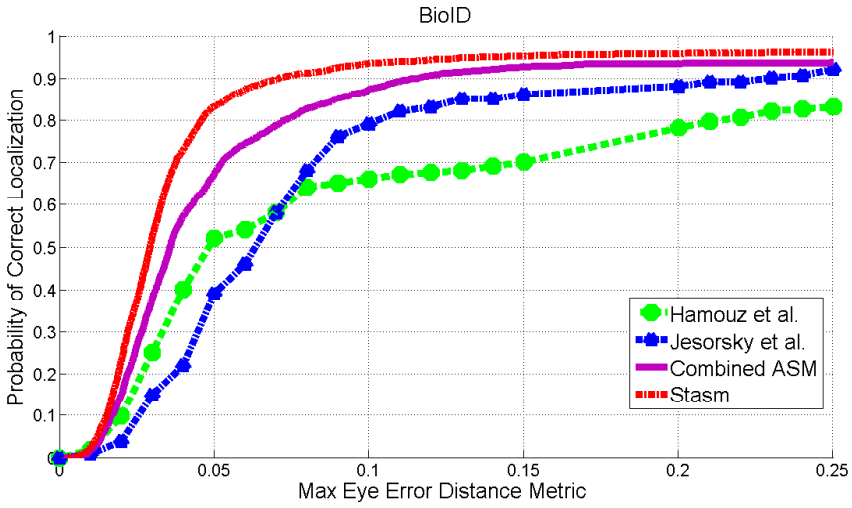

Figure 2.8: Comparison of the proposed Combined-ASM with already published results for eyes detections on the BioID database.

For comparison purposes, in Figure. 2.8, we reproduce three published results

related to the d_{eye} measurements including results of Jesorsky et al. [34], Hamouz et al. [30], and the results of the STASM software by Milborrow [44]. These results are compared with our Combined ASM model implementation. The first two methods are image-based methods, while the last two ones are structure-based methods. It is obvious that structure-based methods have better performance than image-based methods even at the error level of $Error < 0.1$. Our Combined ASM method performs better then the two image-based methods, but worse then the available STASM software.

The results of the STASM software that we trained with a different training data (the IMM database) are presented in Table 2.1. Using different training database results in different results the for Stasm Software.The XM2VTS database contains more training images and more landmarks, this results in better detection performance. However if we use same training database, the proposed C-ASM gives better results, see Table 2.1.

Table 2.1: Evaluation results on the BioID Database. Spatial error rate (at 10 %) of eyes and mouth centers detection, of various landmark detection algorithms on the BioID database [34](in %).

Method	Result	Training Database
Stasm-original	95	XM2VTS
Stasm-modify	78.5	IMM
SIFT-ASM	75	IMM
Combined-ASM	86	IMM

Evaluation on the FRGCv2.0 Database

From the FRGCv2.0 database [51], we used the subpart called **spring2003** which contains 11, 204 images, to evaluate our landmark location precision. There are not published results available on the FRGCv2.0 related to landmark location. Therefore we can only compare our results with the results of the Stasm software.

Because with the Stasm software (that also uses a face detection part as a fist step) there are about 39 % of the above mentioned **spring2003** set of the FRGCV2.0 database images where the STASM face detection algorithm fails, we applied our landmark location software on the same set, for sake of comparison.

In Figure 2.9 we compare the results of the Stasm software, with the proposed

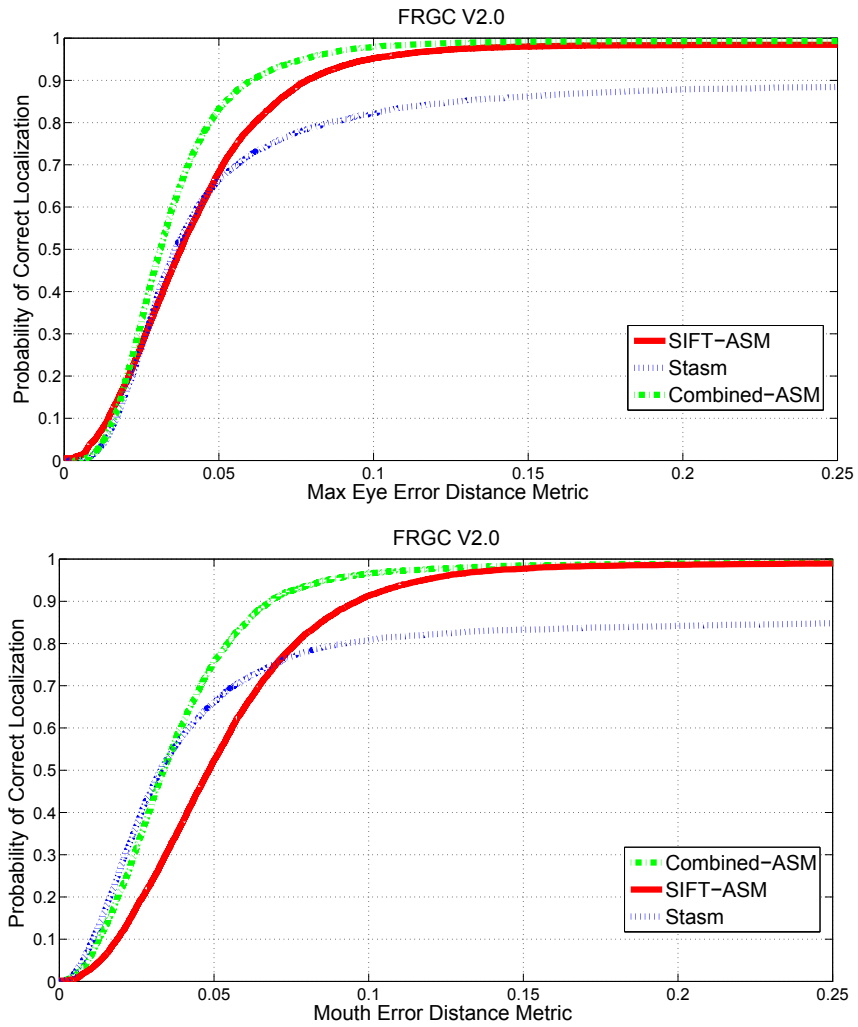

Figure 2.9: Cumulative histograms on FRGCv2.0 database with maximum eyes and mouth error, the Stasm in this experiential we used the default training data (STASM-original).

Combined ASM. In order to evaluate the contribution of using different features for different parts of the face, we also report results of using the ASM model with the SIFT features instead of the originally proposed grey level profile features (denoted as SIFT-ASM). The result show that the C-ASM method gives better performance for the eyes and mouth locations then the Stasm software and the ASM method with new SIFT features.

The C-ASM highly improves the precision of the mouth center, because the SIFT feature descriptor works more inaccurately in the contour points, and those points will affect the mouth region landmarks during the location phase. The results of the STASM software that we trained with a different training data (the IMM database) are presented in Table 2.2. Using different training database also results in different results the for Stasm Software in this experiment. But the C-ASM got 95% correctly detected rate, using the IMM database for training.

Table 2.2: Evaluation results on the FRGCv2.0 Database. Spatial error rate (at 10 %) of eyes and mouth centers detection, of various landmark detection algorithms on the FRGCv2.0 Database [51].

Method	Result	Database
Stasm-original	81.6	XM2VTS
Stasm-modify	77.4	IMM
SIFT-ASM	86.8	IMM
Combined-ASM	95	IMM

In this experiment, we only measure the landmark location precision on the FRGCv2.0 database by comparing the detected landmarks and the ground truth. We will study the influence of the landmarks location for face verification on the MBGCv1.0 and v2.0 databases, where no manually annotated landmarks are provided.

2.5.3 Experimental Discussion

The above presented experiments show that structure-based methods have better performance than image-based methods for facial landmark location. The Stasm software which uses the 2D profile and an extended set of landmarks for the training phase, presents better results on the BioID database compared to the C-ASM. But it is not as good as the proposed C-ASM for the FRGCv2.0 Database.

One possible reason is due to the different characteristics of the databases. In the BioID database, all the images are captured when the person is near the camera, so the face is the largest part of the image. In the FRGCv2.0 database, there are uncontrolled images where the human face occupies a smaller area in the image. When the face area is small in the image, the initialization from the face detector will not be as precise as for "passport style" photographs. Actually the ASM is an iteration strategy whose performance highly depends on the initialization. The Stasm software uses 2D profile as Local Texture Models which will increase the precision, while C-ASM use the SIFT descriptor which increase the robustness when bad initialization happens. Because the SIFT descriptor is scale and rotation invariant, even if the face area detected by the face detector is enlarged and decreased or distorted, it will not affect the Local Texture Models matching phase.

In the BioID database the average distance between two eyes is about fifty pixels, one pixel costs two percent error rate, and that error can be ignored when normalizing the face for face recognition. In that case the Combined-ASM algorithm seems to be robust without losing much of its accuracy for facial landmark detection for face recognition in cases when illumination, scale and small pose variation is present in the recording conditions (such as present in the FRGCv2). That is to say the proposed C-ASM is more suitable for uncontrolled images under degraded conditions, because the robustness of the SIFT descriptor and the separation in facial internal and facial contour region.

2.6 Application for 2D Face Recognition

As summarized in [1] by Zhao et al. in 2006, automated recognition of faces has reached a certain level of maturity, but technological challenges still exist in many aspects. For example, robust face recognition in outdoor-like environments is still difficult. It is also important to precise that in a broader sense face recognition implies usually face detection, face normalization, feature extraction and face recognition tasks (modules). A lot of methods need facial landmark location for the normalization step. Therefore all those problems have to be solved separately in order to have good face recognition systems. In that case a robust and automatic facial landmark location al-

gorithm is needed.

In Section 2.5 we have evaluated the precision of the C-ASM landmark detector on databases with annotated ground truth. It is also important to evaluate the performance of the final face recognition system related to its components, in order to be able to pinpoint the existing challenges separately. These points are a necessary step if face recognition should be applied for challenging applications, such as video surveillance. It is difficult to find relevant databases that have the good annotations with ground truth data, such as position of faces and landmarks, so that we will also evaluate the landmark detection by face recognition performance.

2.6.1 Fully Automatic Face Recognition with Global Features

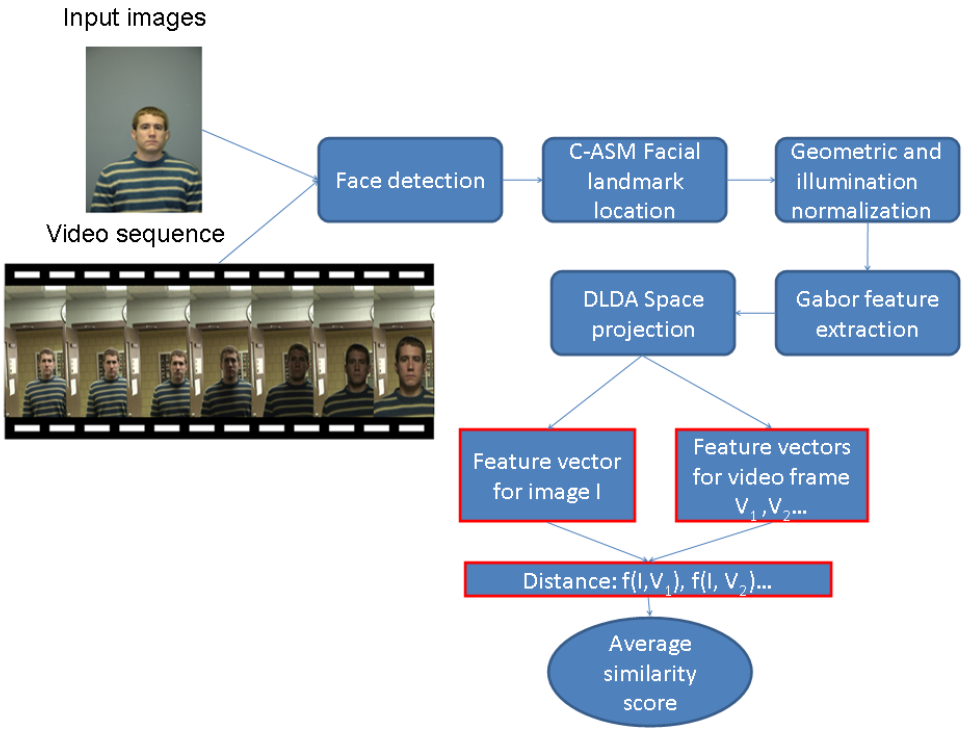

Figure 2.10: Flow chart of the system for our fully automatic face recognition based on C-ASM and global features. Images and video from MBGCv1 portal challenge [46].

For global feature based recognition algorithm we use the open source sudfrog system. More details about this software can be found in `svnext.it-sudparis.eu/svnview2-eph/ref_syst/sudfrog/`. The flow chart of the system could be seen in Figure 2.10. Given an input reference image or a video frame the process includes the

following steps:

1. Face detection: For this step we used the well known face detection framework by Paul Viola and Michael Jones in 2001 [69]. This algorithm is implemented in OpenCV [11].

2. Facial landmark location: in this step, our proposed C-ASM is used to locate 58 landmarks on the face. Some examples could be seen in Figure 2.11. From the 58 landmarks, we compute the positions of the eyes and mouth.

3. Geometric and illumination normalization: Thanks to the centres of the eyes and mouth we got from the previous step, we applied a geometric normalization to the face area of the overall image. It provides a reduced image (128x128 pixels) of the face, where eyes ($(32, 42)$ and $(96, 42)$) and mouth $(64, 102)$ positions are predefined. Finally, an anisotropic smoothing is used to reduce the influence of the variation of illumination in uncontrolled conditions (see Figure 2.12).

Figure 2.11: Typical landmark location results from the MBGCv1 portal challenge [46]. Top: landmark location results on still images. Bottom: landmark location results on video frames (Not all the video frame we have the same detection of landmarks, in here we just show some typical examples).

4. Gabor feature extraction: we use Gabor filters with several resolutions and orientations (5x8) convoluted with the normalized images and only magnitude value are used in our experiment. After convolution 5x8=40 Gabor magnitude images with size (128x128) are obtained, which is too big for processing. Therefore for each Ga-

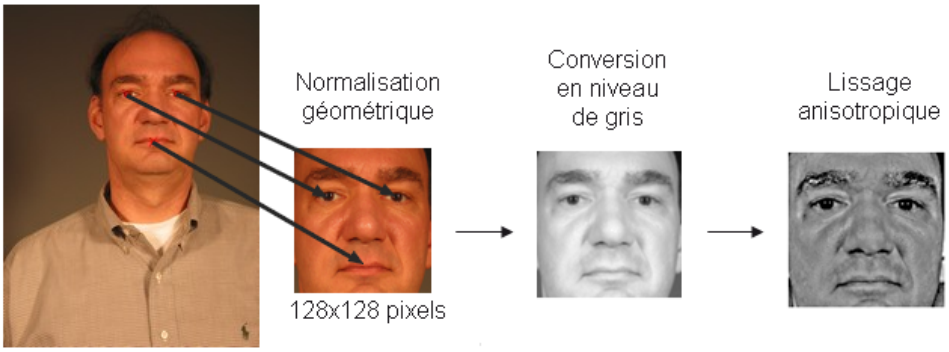

Figure 2.12: The geometric and illumination normalization, image from [48].

bor magnitude images we down-sample it to reduce the size to be (32x32). Finally we present the face as a vector with size (32x32x40) as the extracted feature.

5. DLDA (Direct Linear Discriminant Analysis) space projection: We project the vector in the DLDA space to reduce the dimension of the space. So the output of this step will be a vector with length related to the size of the DLDA training database to represent a face image. The learning data is from FRGCv2.0 training set, with 120 persons with approximately 10 images for each person, so the output vector length will be 119.

Those parameters setting are used in the following experiments. After the above precessing, the reference image is presented as a vector FV_{Image}, and the video can be presented as a set of vectors $FV_{Video_1}, FV_{Video_2}, ..., FV_{Video_K}$, where K is the number of video frames where we have successfully detected the landmarks (ignoring the images which we can not detected face from).In calculating scores, the cosine distance is used to estimate the similarity between two feature vectors. Given that the comparison is between an enrolment image and test video, we calculate the final score as the mean of these K distances. This is a simple way to calculate the score. More attention is needed in order to choose and select better reference image from the video sequence. This is not the purpose of our work.

2.6.2 Face Recognition Databases

Related to video surveillance, the most relevant existing publicly available databases seem to be the MBGC portal and video challenge. Examples of data from

those challenges are given in Figure 2.13. In the first evaluations, MBGC in Decem-

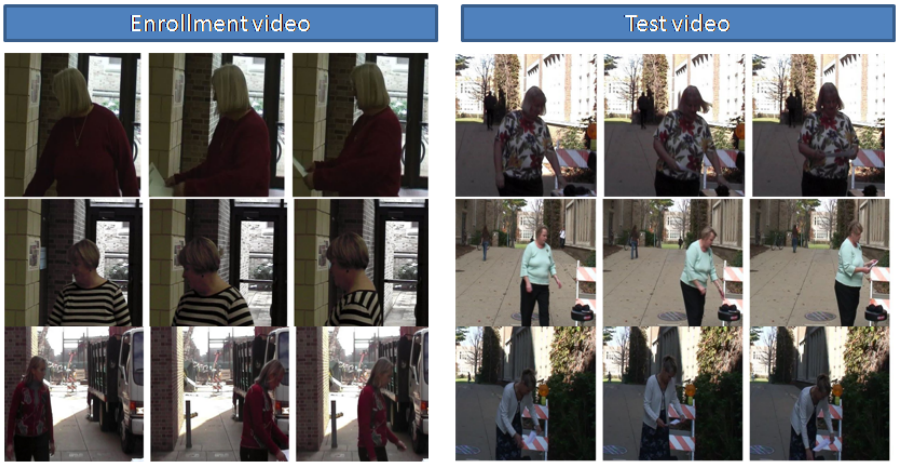

Figure 2.13: Images from the "Video MBGC challenge" videos of enrolment (first 3 columns) and test (last 3columns).

ber 2008 [www-mbgc-2008], three databases were provided: **Still**, **Portal** and **Video**. The results of the **Still** and **Portal** challenge were acceptable. But the algorithms submitted for the **Video** Challenge, presented the most difficult case, with two submissions that are close to a random guessing result (see MBGC webside `http://www.nist.gov/`).Figure 2.13 shows examples of the **Video** Challenge. These examples illustrate the fact that under such conditions, other information instead of face are needed for recognition of people. To address the variety present in these sequences is through detection of the silhouette, and also face detectors for poses other than frontal are needed. Moreover, when comparisons are made under conditions of face-profile comparison, the "traditional" 2D face recognition methods are insufficient. Because the **Video** Challenge is too difficult, in our application we are using the **Portal** Challenge to evaluate our 2D face recognition system.

The scenario studied in the **Portal** Challenge is to compare a reference image to a video sequence where the person walks through a metallic gate toward the camera. A practical application of this scenario is identity verification at border posts. This database contains data from 111 people including 29 people who provided two acquisition sessions. During a session, a reference image and a video are recorded (see Figure 2.14). Thus, the overall enrolment consists of 140 reference images while the test set contains

140 video sequences. Enrolment images (1504x1000 or 998x1500 pixels) were acquired under controlled illumination and contain only frontal faces with an average of 110 pixels between the eyes. The test videos (1080x1440 pixels / frame) had an average of three seconds. The resolution between the eyes varies from 50 to more than 200 pixels as the person approaches the camera. The evaluation protocol has 200 client-client and 5592 client-impostor tests.

Figure 2.14: Example of biometric data extracted from the MBGCv1 - Portal Challenge (http://www.nist.gov/itl/iad/ig/mbgc.cfm).

In the December 2009, we also participated the the portal challenge experiment of the MBGCv2 evaluations, which is the extension evaluation of MBGCv1 with more images (about 2000 enrolment images) and 512 test videos. The still images are no more under controlled conditions as in the MBGCv1.

2.6.3 Face Verification Experimental Results

MBGCv1 results

Table 2.3: Face verification result on MBGCv1 portal challenge as a function of EER. EER denotes Equal Error Rate and VR denotes face verification rate, SIFT-ASM denote the preliminary version of the proposed method (using only ASM with SIFT features.) The confidence interval at 99.9% [] is calculated as explained in [49].

The face images are geometrically normalized using eyes and mouth centers.

Landmark detector	EER Session I	VR Session I	EER Session II	VR Session II
Combined-ASM	4.1%[±0.16]	80%[±0.08]	1.8%[±0.05]	86%[±0.06]
SIFT-ASM	5.1%[±0.22]	74%[±0.08]	3.9%[±0.15]	78%[±0.07]
OpenCV eyes detector	27.1%[±0.82]	34%[±0.68]	24.1%[±0.75]	28%[±0.57]

Table 2.3 lists the results of our recognition system on MBGCv1 portal chal-

lenge. There are two sessions in total, In Session I, the reference images are acquired with flash compared to the reference images from Session II. We also compared its performance with publicly available OpenCV eyes detector [11]. It shows the influence of the landmarks on the face recognition result. For the global feature based face recognition, the face normalization step is very important. More precise landmarks we have, the better face recognition result we can get. The OpenCV eyes detector doesn't work well for our face recognition system. By using the Combined-ASM, we have improved the Equal Error Rate (EER) from 5.1 % and 3.9 % to 4.1 % and 1.8 % compared to SIFT-ASM.

For the error analysis, we considered a new experiment, that is a face identification experiment in order to find the false accept images. For each query video we find the enrolment images of no matching subjects that produced a smaller similarity score than the corresponding enrolment image of the same subject. In Table 2.6.3, we have listed the errors for the match and no match comparisons. By visualizing the errors, we can obtain the information that the problems result in false alarms and false acceptance. From Figure 2.15, in the top row, the problem may be the pose and image distortion and pose variation. The video captured by the Digital Video have different quality than still images. In the middle row, the problem could be the expression variation, we can see in the enrolment still image the lady closes her eye, this may bring the difficulty for the recognition proceeding. In the third row, video and images has high quality, but the two enrolment images are too similar, so that our face recognition algorithm can not distinguish them.

Also we have found that the enrolment images are very important, for each video we have several images, it does't affect a lot if some image have bad quantity, but for the enrolment images we only have one per subject. If they are bad, then we have no chance to have good results in client-client and client-impost tests.

To summarize, the are two aspects to improve our system. First, preprocessing images and videos, for example, to select high quantity frame from the video sequences, or to correct pose of face before passing them to the face recognition system. Second, to enhance ability of 2D face recognition system to distinguish between similar faces , this could be done by using more suitable training data and better classification algorithms.

In our experiment, only the global feature based face recognition algorithm is

Figure 2.15: Some examples of wrong identified examples on MBGCv1 database. The left column images are from the query video frame. The middle column are the enrolment images of non-matching subjects to video, that produced a smaller similarity score than the corresponding enrolment images of the subject. The right column are the corresponding enrolment images of the same subject.

Table 2.4: Error analysis on the MBGCv1 portal challenge using our automatic face recognition system.

	Session I	Session II
Total number of match comparisons	98	102
False alarms	8	7
Total number of no-match comparisons	2856	2736
False acceptances	32	18

tested. However the local feature based face recognition method is also very interesting for us, as we have detected the landmarks on the fiducial points such as eyes conner, mouth conner and so on, to use their landmark for local feature based face recognition can be tested in the future.

MBGCv2 results

For the portal challenge of MBGCv2 evaluations, the results we have obtained is $EER = 10\%$. The video sequence is almost similar to MBGCv1, while the most problems are from the diversity of the enrolment still images. Some difficult examples are show in Figure 2.16. In that case, the face detector can not give the correct initialization to our C-ASM detector. There are around 200 images in the enrolment set that we can not detect or have wrong detection. This is the main reason which drop down the performance of our face recognition system. The public results of MBGCv2 could be found on the MBGC webside (http://www.nist.gov/itl/iad/ig/mbgc.cfm).

Figure 2.16: Some challenging examples of enrolment still images from MBGCv2 database [46]. The images are too big (left) , to small (middle), or incomplete (right).

2.7 Conclusions

In this chapter, we present a new algorithm that successfully localizes facial landmarks for 2D face recognition experiments in degraded condition, such as video surveillance. We assess the localization performance of the proposed method on two datasets (BioID and FRGCv2.0). Also the proposed method is experimented on the portal challenge experiment of the MBGCv1 and v2 evaluations, where an automatic landmark detector was needed, in order to find the position of the normalization points in the video frames and enrolment images.

In the proposed Combined Active Shape Model (Combined-ASM) we extend the original ASM by using the SIFT descriptor as a new local texture model and split the facial landmarks in facial internal region and facial contour landmarks. The proposed Combined Active Shape Model algorithm is more robust for eyes and mouth center localization in more challenging lighting conditions, and also where some pose and expressions variabilities are present.

Chapter 3

Automatic 2D Facial Landmark Location using 3D Active Shape Model

3.1 Introduction

In the previous chapter, we presented the Combined Active Shape Model, to enhance the performance the original Active Shape Model for the landmark location under degraded conditions, but with the hypothesis that we deal with 2D near frontal images. However, one of the important obstacles in image-based analysis of the face is the 3D nature of human faces. When fully automatic face analysis systems are designed, capturing frontal-view images cannot be guaranteed. Examples of such situations can be found in video surveillance systems, car driver images or whenever there are operational constraints that prevent from placing a camera frontal to the subject. In such situations, 2D facial landmark location systems working across large pose variations are needed.

In this chapter, a 3D Active Shape Model (3D-ASM) algorithm is presented to automatically locate facial landmarks from different views. The proposed 3D-ASM system is based on the well known Active Shape Model [19] with the following improvements:

1. Taking advantage of 3D scans of faces as training data, we propose to exploit 3D statical shape models and projective geometry across different views. The 2D face shape can be considered as the projection of a 3D model. Compared to the original 2D ASM proposed by Tim Cootes in 1995 [19], we separate shape variations into intrinsic changes (caused by the character of different person) and extrinsic changes (caused by model projection).

2. Our system in based on searching with a set of view-dependent local patches to locate facial landmarks, and using these to update the face shape model parameters of 3D-ASM. In that case the self-occlusion problem can be solved efficiently.

3. We propose to train the 3D-ASM with data generated from the 3D Morphable Model (3DMM). Using 3DMM to synthesize training data offers us two advantages: first, few manual operations are need, except labelling landmarks on the mean face of the 3DMM. Second, since the learning data are obtained directly from the 3DMM, landmarks have one to one correspondence between the 2D points detected from the image and 3D points on the 3DMM. This kind of correspondence will also benefit 3D face reconstruction processing.

The rest of chapter is organized as follows: first a brief literature review about facial landmark location across pose is given in Section 3.2, that completes the review of facial landmark location in 2D frontal face images given in the previous chapter. The proposed 3D-ASM construction and fitting are explained in Sections 3.3 and 3.4 respectively. In Section 3.5, we explain how to use a 3D Morphable Model to train our 3D-ASM. The databases and experimental protocols necessary for the training and evaluation phases are presented in Section 3.6. The results are reported in Section 3.7. Finally, the conclusions can be found in Section 3.8.

3.2 Literature Review about Facial Landmark Location across Pose

In the previous chapter, we gave a brief review about facial landmark location on frontal view images. In this chapter we will focus on the problem of landmark location across pose. As summarized in (Sukno 2007) [65] by Sukno in 2007, facial images present large changes in shape and appearance when the relative angle between the camera and the face is modified. The three-dimensional nature of the head is further complicated by the non-rigid motion it can involve. A lot of algorithms have been proposed for facial landmark location across pose. They can be classified in three categories: appearance-based, statistical model-based and 3D model-based methods.

- Appearance-based methods: Several works tackle pose variation by learning the relationships between different views. Fan et al. [23] learn a pose change model from example images. A Gaussian skin-color model is used to coarsely detect faces under varying viewpoints and a feature-based strategy provides further refinement and rejection of false alarms. Sanderson et al. [56] learn prior information of the face from multiple 2D views of a prototype training set. The authors used maximum likelihood linear regression (MLLR) and standard multivariate linear regression. In the MLLR approach, a generic face model is constructed for each viewpoint.

- Statistical model-based method: Another group of approaches can be identified as based on statistical models. As a general rule the models for facial images are bi-

dimensional and cannot handle large pose variations. Combining a number of them to extend their viewpoint range has been a popular solution: view-based Active Appearance Models (AAMs) [20] and view-based Direct Appearance Models [77] are some examples. This idea is followed also by Li et al. [59] and Xin et al. [75]: the whole range of views from frontal to side views is partitioned to construct separated statistical models. The model to be used for an unknown image is determined with the help of a multi-view face detector [80]. The use of a single statistical model to deal with the whole range of views was proposed by Romdhani et al. [54]. They used KPCA (Kernel Principal Component Analysis) to make the point distribution model non-linear and added the viewing angle as an additional parameter to the landmark vector.

- 3D model-based methods: More sophisticated solutions tackle the problem by dealing with a 3D model of the face. Projective geometry theory was used by Buxton [13] to deal with the alignment of shapes under different viewpoints in ASMs. By restricting themselves to affine imaging conditions, the authors propose a method to remove pose variation based on two reference views, appropriately selected from a multi-view dataset. Their Integrated Shape and Pose Model (ISPM) is presented as an extension to the Linear Combination of Views (LCV) under affine conditions. An important point in the work of Buxton is the selection of a subset of facial landmarks (although manually) for the alignment, based on the observation that the face is not a rigid object and substantial shape differences may be present in the different views to be aligned.

  There are also some approaches half way between 2D and 3D, which derive 3D shape models from multiple 2D views but perform the image search in 2D. This is the strategy followed by Xiao et al. [74] and Mathews et al. [43], based on AAM. Li et al. [41] jointly optimize overall appearance, local appearance (around landmarks), and the difference to the previous frame. An interesting point in the combined loss function of Li et al. is the introduction of a visibility weight for the appearance of each landmark, which depends on pose (based on the normal to the landmark in the 3D shape). The method was reported to behave reliably in the range of $[-70, 70]$ degrees in yaw though no quantitative results were provided.

Tong et al. [67] combine 2D and 3D: the authors assume that a 3D model is available and use it to estimate head pose. Then, the shape model is corrected to match the estimated pose using an affine approximation.

Gu et al. [29], use 3D deformable model to segment a single face image. A single frontal-view per person is used to synthesize a multi-view database from which to learn prior information about pose changes. The model consists of a set of sparse 3D points and the view-based patches associated with every point. Assuming a weak perspective projection model, the algorithm iteratively deforms the model and adjusts the 3D pose to the image in a EM (expectation maximization) framework, but no quantitative results are provided. In [76], the author defined a 3D general shape to align face shapes in 3D instead of 2D alignment. Those two methods are both related to an ASM framework. Efraty et al. [22] proposed to create training landmarked samples from 3D scan faces database. These landmarks are further employed to train a profile view 2D ASM, but even the learning data are 3D, the model only works on near profile faces in their experiments. In [16], Caunce et al. built a sparse 3D shape model from 923 head meshes, they used the normalised view-based local texture patches which is similar to Gu and Kanade [29], but continuously updated to reflect the current model pose.

A common drawback of all the above techniques is that they need somehow large databases to construct the facial models. In the work of Gu et al. [29], a frontal-view image per person have to be manually segmented and annotated, which is time consuming, tedious and subjective. So we propose to use a 3D Morphable Model (3DMM) to automatically generate landmarks needed to train a 3D Active Shape Model (3D-ASM). The advantage of the proposed 3D-ASM method is that there is no need to manually annotate the landmarks in the training 2D images. It is only necessary to define the set of landmarks that are needed on the 3DMM, that are going to be propagated automatically on any newly generated 3D faces related to the original 3DMM. Such automatically generated landmarks can serve as training examples for the 3D-ASM.

3.3 3D Active Shape Model Construction

Like the original 2D ASM introduced by Tim Cootes [19] (and explained in the previous chapter), our 3D-ASM is composed of a 3D Point Distribution Model (3DPDM) and a 3D view based Local Texture Model (3DLTM), in order to handle statistical information of the 3D shape geometry and texture variations for each landmark. For training a 3D-ASM two things are needed, a set of 3D scans and the corresponding landmarks on the 3D scans. The landmarks on the 3D scans are needed to build our 3DPDM, which is a sparse 3D point set, with a 3D shape prior of human faces. The 3D scans are used to synthesize 2D images in different views to train the view based Local Texture Model (LTM) associated with every point on the 3DPDM.

3.3.1 3D Point Distribution Model

The 3D Point Distribution Model (3DPDM) is very similar to the 2DPDM, except the addition of the z coordinate. A 3D shape can be described by a vector of 3D coordinates $S_{3D} = [x_1, y_1, z_2, ..., x_N, y_N, z_N]$, where N is the number of landmarks. The 3DPDM is obtained from the PCA spaces of the 3D faces with the corresponding landmarks on different 3D scans:

$$S_{3D} = (\overline{S_{3D}} + \Phi_{3D} \cdot p_{3D}), \tag{3.1}$$

where $\overline{S_{3D}}$ is the mean shape in the 3D space, and Φ_{3D} is the eigenvector matrix.

The 2DPDM explained in previous chapter deals both with intrinsic changes (caused by the change of expression and different persons) and extrinsic changes (caused by camera projection) with a single model. While 3DPDM reflects only the intrinsic changes. The extrinsic changes are handled by the camera model and 3D geometric transformation parameters.

The detected 2D shape located in images $S^*_{2D} = [x^*_1, y^*_1 ... x^*_N, y^*_N]$ could be considered as the observation of the 3DPDM projection on the 2D image plan:

$$S^{pro}_{2D} = P(sR(\overline{S_{3D}} + \Phi_{3D} \cdot p_{3D}) + t), \tag{3.2}$$

where P is a projection matrix, R is a $3x3$ rotation matrix, t is a translation vector, and s is the scale parameter. In this chapter we assume an orthogonal projection where

$$P = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{vmatrix}.$$

3.3.2 3D view-based Local Texture Model

Using SIFT descriptor as Local Texture Model in ASM can make the algorithm more robust to out-of-plan rotations as explained in previous chapter. In this work we use the same SIFT descriptor but extend it by using view-based statistical models for each landmark. This enables us to increase the ability to deal with large rotations and avoid the self-occlusion problem during fitting. The concept of 3DLTM is described in Figure 3.1. In order to describe the landmarks from different view , we render all 3D faces in different views by setting different roll angles . For example in Figure 3.1, the roll angles of the 3D model are set to be: $(-90, -60, -30, 0, 30, 60, 90)$. For new each landmark, a view-based statistical texture model is built separately:

$$f(g^v_{new}) = (g^v_{new} - \overline{g^v})C_g^{v-1}(g^v_{new} - \overline{g}^v)^T. \tag{3.3}$$

The $\overline{g}$ is the mean local texture model, and C_g is the covariance matrix of texture model. 2DLTM ($\overline{g^v}$,C_g^v) is trained for each view v separately. During the fitting (of the 3D-ASM to 2D images), the LTM are chosen dynamically according to the current pose (see Section 3.4).

As shown in Figure 3.1, for occlusion points, we extract the SIFT descriptor from the synthesized training images at the positions where those 3D points are projected. This kind of " virtual " descriptor can make all the landmarks have a uniform presentation, and we don't need to consider which points are occluded during the fitting procedure. All the points are treated in a same way. This will benefit top simplify the fitting procedure and solves the self-occlusion problem, this is one of novelty of the thesis, see next Section 3.4.1.

3.4 2D Landmark Location: Fitting the 3D Active Shape Model to 2D Images

Once the 3D-ASM is trained, in this section, we explain how to exploit it for landmark location in 2D images (by a fitting procedure).

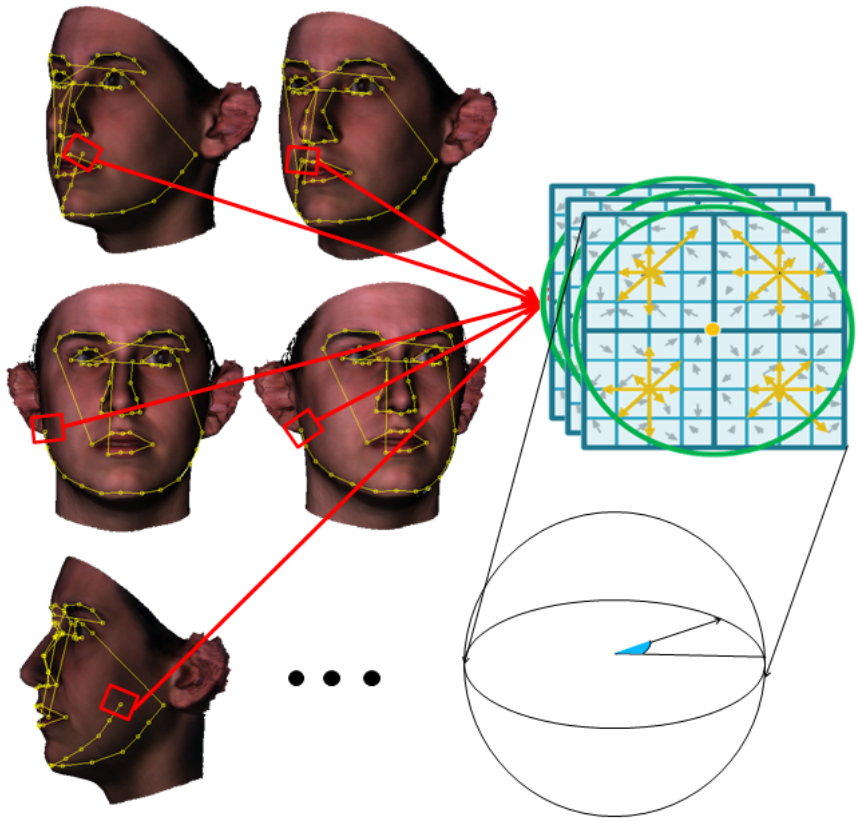

Figure 3.1: Illustration of 3D Local Texture Model. For each landmark, one 2DLTM is built separately for each viewpoint. 7 view-based 2DLTM compose our 3DLTM.

3.4.1 Framework of Matching Algorithm

We construct a two-layered Gaussian pyramid, and apply the alignment algorithm sequentially from the coarsest to the finest layer. Then the algorithm goes iteratively as follows:

1. **Local Search:** For the i^{th} landmark point, compute Mahalanobis distance using the local texture model of the current view around the current location, then select the best candidate (x_i^*, y_i^*) which has the smallest distance as new location.
2. **Estimation of Parameters:** The estimation of 3D shape (p_{3D}) and pose parameters $(\mathbf{R}, s, \mathbf{t})$ from 2D shape $S_{2D}^* = [x_1^*, y_1^*, ..., x_i^*, y_i^*]$ is an ill-posed problem. We consider it as an over-constrained non-linear optimization problem that can be solved by generalized Gauss-Newton iterations as described in 3.4.2.
3. **Texture Model Update:** The view-based local texture models $(\overline{g^v}, C_g^v)$ is chosen according to pose parameters obtained from step 2.

Because of using a 3D face model, there could be some landmarks that have the self-occlusion problem under the corresponding viewpoint. Normally for this problem, the occluded landmarks can be estimated by the Z-buffer algorithm [64, 17], and then the observed non-occluded shape points are used to recover the shape parameters by non linear parameter estimation. For example, Gu et al. [29] use the expectation-maximization (EM) algorithm to deal with this problem.

We propose a new solution for to occluded points: as shown in Figure 3.1, during the training phrase for occlusion points in 3D we only consider the 2D projection position on 2D images without taking into account the occlusion. Since the 3D view-based LTM is a discontinuous model depending on the view points, it can approximate the LTM of the occluded points in the specific view points, because the texture model of occluded points are learned in the training phase. This kind of approximation makes the fitting procedure much simpler and efficient.

3.4.2 Shape and Pose Parameters Optimization

Given the observation shape S^*_{2D} and the 3D Point Distribution Model S_{3D}, the objective function for the optimization is:

$$E = E_d + \lambda E_p; \tag{3.4}$$

$$E_d = \sum_{i}^{N} w_i \left\| S^*_{2Di} - S^{pro}_{2D\,i} \right\|; \tag{3.5}$$

$$E_p = \sum_{j=1}^{m} \frac{p^2_{3D}}{\delta^2_j}; \tag{3.6}$$

where E_d is the error between the observation shape and the projected shape. m is the number of the shape eigen vector.The contribution of the i^{th} landmark is weighted with a landmark specific weight w_i, which is inversely proportional to the Mahalanobis distance of each landmark on the observation shape to the mean of the corresponding Local Texture Model. The purpose of this weight is to define the quality of the location of each landmark. The weights w_i are normalized between $(0.1, 1)$ and updated dynamically during the fitting. The E_p specifies the a priori term, which constraints the shape deformation to reasonable values and δ_j is the eigenvalue associated with the j^{th} eigenshape of the 3DPDM.

At the beginning of optimization, we set λ equal to be zero,so only the pose parameters are optimized. After that λ is set such as E_d is proportional to E_p.

3.5 How to Synthesize Training Data from 3D Morphable Model

One of the problems of the Active Shape Model is the availability of training data. As described in Section 3.3, for training a 3D-ASM two things are needed: a set of 3D scans to synthesize 2D images in different views, and the corresponding landmarks (the same feature points)on the 3D scans. They are used to build the 3DPDM and 3DLTM separately. One solution is to manually label landmarks on the 3D scans. In our work, we exploit the characteristic of the 3D Morphable Model, with which training data can be obtained in a more simple way. We first remind the 3D Morphable Model

(3DMM) on which our approach is based. Then we introduce the process of training the 3D-ASM, including 3D Point Distribution Model and 3D view-based Local Texture Model, using synthetic data from 3DMM .

3.5.1 Reminder about 3D Morphable Model

3D Morphable Model (3DMM) was introduced by Blanz and Vetter [9]. It is a parametrized model that can generate synthetic 3D faces constructed from a set of 3D facial scans. A vertex-to-vertex correspondence of all 3D training faces is a condition to build a properly working morphable model. Such models are based on the key observation that given two 3D faces, if they are previously registered, their linear interpolation (also known as 'morph") will still describe a human face, which make human faces lying in the 3D space intrinsically.

In [9, 10], the morphable model is acquired from 3D scans of 100 males and 100 females, aged between 18 and 45 years. These scans are recorded with a *Cyberwave* 3030PS laser scanner. The scans represent face shape in cylindrical coordinates relative to a vertical axis centred for the head. There are 512 angular steps covering 360 and 512 vertical steps at a spacing of 0.615mm. After the raw scans are obtained, some preprocessing is needed.

1. Holes are filled and spikes are removed on the face surface.
2. 3D data are aligned with a 3D-3D Absolute Orientation.
3. Heads are trimmed along the edge of a bathing cap.
4. Heads are cut vertically behind the ears to remove the back of the head.
5. Heads are cut horizontally at the neck to remove the shoulders.

After the above preprocessing, a modified optic flow method is applied to establish dense point-to-point correspondence between a new face and a reference face. The shape and texture vectors of the reference face are:

$$\mathbf{S}_0 = (x_1, y_1, z_1, x_2, \ldots, x_n, y_n, z_n)^T \tag{3.7}$$

$$\mathbf{T}_0 = (R_1, G_1, B_1, R_2, \ldots, R_n, G_n, B_n)^T \tag{3.8}$$

where the reference face is a triangular mesh with n vertices ($n = 75972$),(x_k, y_k, z_k) are the Cartesian coordinate for each vertex, and (R_k, G_k, B_k) are the colour values from for vertex ($k = 1, \ldots, n$).

The PCA is performed on the set of shape and texture vectors $\mathbf{S}_i$ and $\mathbf{T}_i$ of m example faces. The $m-1$ eigenvectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{m-1}$ are computed the shape vectors from PCA by a Singular Value Decomposition (SVD) . The eigenvectors construct an orthogonal basis on the shape vectors:

$$\mathbf{S} = \bar{\mathbf{s}} + \sum_{i=1}^{m-1} \alpha_i \cdot \mathbf{s}_i \tag{3.9}$$

where $\bar{\mathbf{s}}$ is the average from each shape vector, $\bar{\mathbf{s}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{S}_i$. So as to the texture vectors, an orthogonal basis is constructed:

$$\mathbf{T} = \bar{\mathbf{t}} + \sum_{i=1}^{m-1} \beta_i \cdot \mathbf{t}_i \tag{3.10}$$

The model parameters (coefficients) α_i and β_i are used to represent a face in an image. Assuming a uniform Gaussian distribution, the probability for coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is given by

$$p(\boldsymbol{\alpha}) \sim exp\left[-0.5\sum_{i=1}^{m-1}(\alpha_i/\delta_{S_i})^2\right], \quad p(\boldsymbol{\beta}) \sim exp\left[-0.5\sum_{i=1}^{m-1}(\beta_i/\delta_{T_i})^2\right]. \tag{3.11}$$

with δ_{S_i} and δ_{T_i} being the eigenvalues of the shape and texture covariance matrices respectively. Figure 3.2 shows the morphing effect achieved as the first shape component α_1 is varied within the ranges $[-2\delta_{S_1}, 2\delta_{S_1}]$.

3.5.2 The 3D Active Shape Model Construction Using 3D Morphable Model to Generate Data

We propose to construct a 3D Active Shape Model by using 3D Morphable Model as follows:

1. On the average shape of 3DMM, select the vertex points corresponding to the desired landmarks;

2. Choose different sets of shape and texture parameters to generate new 3D face data, typically examples could be found in Figure 3.10.

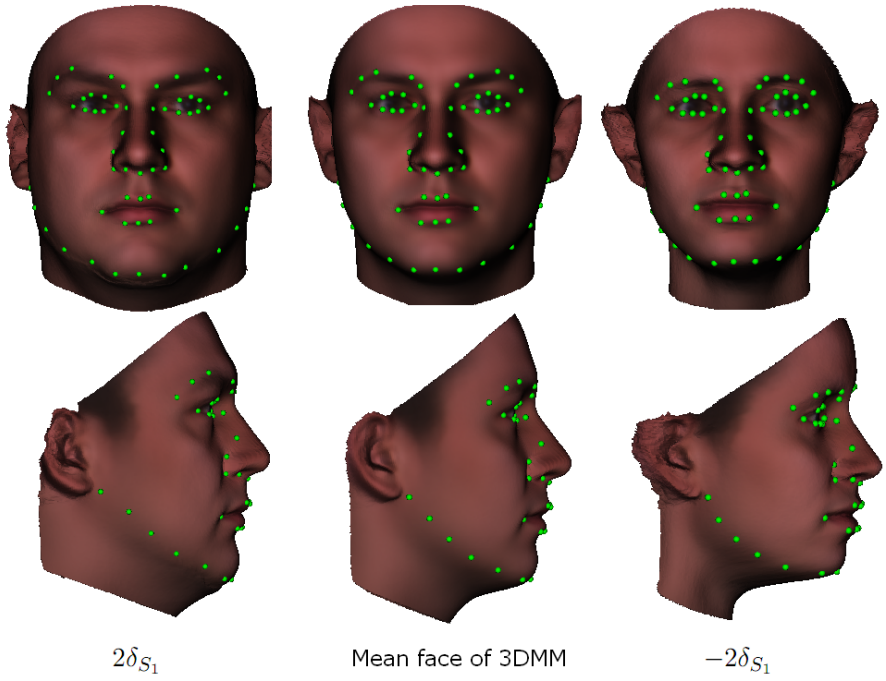

Figure 3.2: 3D Morphable Model from [9] and the 58 manually selected 3D landmarks. Middle: the average face model with 58 landmarks. Left and right: Change the first component ($\pm 2\delta_{S_1}$) of shape parameter and the corresponding 3D landmarks.

3. From those synthesized 3D faces, use the positions of selected vertices to construct the 3DPDM (see Section 3.3.1);

4. Render images from all 3D faces in several views by setting different roll angle.

5. For each view of step 4, extract the SIFT descriptors from previously synthesized images with two-layered Gaussian pyramid processing in order to learn the local texture from different image size;

6. Build a 2DLTM for each landmark, which composes our 3D view-based Local Texture Model, as explanation in the Section 3.3.2.

We can select as much landmarks as we can. For comparison purposes, we manually selected on the 3D average face of the 3DMM, the same 58 vertices as the ones defined in the IMM database [63] (see Figure 3.2). Thanks to the morphing characteristics of 3DMM, by setting different shape and texture parameters we can obtain different 3D faces with the 3D position of the 58 vertices previously defined. Those vertices could be considered as landmarks in a 3D space, as shown in Figure 3.2.

The advantages of using 3DMM to automatically generate training landmarks are the following:

1. We can generate as much landmarks as the resolution of the 3DMM is capable of with few manual operations (only need to label the landmarks on the average shape of 3DMM).
2. We can also generate faces with landmarks for different subjects with pose, illumination and expression variabilities (if they are present in the 3DMM).
3. Since the 3DPDM and the 3DLTM are directly generated from the 3DMM, the 2D landmarks can be considered as a projection of 3D vertices. This is a strong advantage for 3D face reconstruction from 2D images.

3.6 Databases

In this section we first give some details about the databases needed to train the 3D-ASM and the databases to evaluate our 3D-ASM landmark detector. We compared it with our previously reported similar experiments for 2D landmark detection, with the Combined ASM (C-ASM) method which is introduced in the previous chapter. The C-ASM was trained with manually annotated images. The test databases include, BioID [34], IMM [63] and PIE [60] database. A part of the IMM database (80 images with pose variation) was taken in order to make the comparison with our previous C-ASM results. As the proposed 3D-ASM method should be more robust to pose variations, we evaluated it also on the PIE database, which contains more pose variability and for which ground truth information about some landmarks are also available.

3.6.1 Training Database for the 3D-ASM

In this chapter, we use the 3D Morphable Model (3DMM) provided with the USF Human-ID 3D Database [9]. The USF Human ID 3D face database consists of 136 scans of 136 subjects acquired with the Cyberware 3030PS laser scanner. There are more than 90K vertices and 180K triangles for each face model.

The publicly database 3D Morphable Model is built from 3D scans of 100 individuals, aged between 18 and 45 years. These scans are recorded with a *Cyberwave* 3030PS laser scanner. The scans represent face shape in cylindrical coordinates relative

to a vertical axis centred as the head. In the database there are original scans together with the trained 3D Morphable Model for the synthesis of 3D faces. In a word, a 3D face PCA space of shape and texture able to synthesise faces of any individual in 3D. The PCA coefficients of the 100 original scans which are used for training are also provided in a file *TRAIN100.FSC*, which can be used to generate the correspondent subjects. It should be noted that this is the only publicly available 3DMM. And the 3DMM database used in [10] is built from 200 scans.

3.6.2 Evaluation Databases

In order to validate the performance of using automatically generated landmarks for the training phase of the proposed 3D-ASM, we conducted different experiments on different databases:

- Experiment with 3D-ASM training with real scans denoted as **train-real-scan** and evaluated on the following databases:
  - BioID Database (to compare the performance with the C-ASM in frontal images);
  - PIE Database (to compare the performance with the C-ASM in nonfrontal images);
  - IMM Satabase (to compare the performance with the C-ASM in nonfrontal images with more landmarks).
- Experiment with 3D-ASM training randomly generated 3D face denoted as **train-random** and evaluated on the following database:
  - IMM database (to compare the performance with the 3D-ASM with real scans).
- Experiment with 3D-ASM training real scan with more views denoted as **train-views** and evaluated on the following database:
  - IMM database (to compare the performance of 3D-ASM training with 7 and 9 view points).

Figure 3.3: Typical images from the IMM database [63].

IMM database: The IMM Face Database [63] is composed of 240 still images of 40 different human faces, all without glasses. The gender distribution is 7 females and 33 males. Images were acquired in January 2001 using a 640x480 JPEG format with a Sony DV video camera, DCR-TRV900E PAL. The following facial parts were manually annotated using 58 landmarks: eyebrows, eyes, nose, mouth and jaw. A total of seven point paths were used; three closed and four open.

PIE database: The CMU (Carnegie Mellon University) Pose, Illumination, and Expression (PIE) database [60] is collected between October 2000 and December 2000 consisting of over 40,000 images of 68 subjects. To obtain significant illumination variation the data is acquired in a 3D Room with a "flash system with 21 flashes to obtained different illumination conditions. To obtain a wide variation across pose, 13 cameras are composed to a camera array in the CMU 3D Room. Furthermore, several different expressions are presented for each subject. The camera positions are shown in Figure 5.2 .

BioID database: The BioID dataset consists of 1521 gray level images with a resolution of 384 $\times$ 286 pixels. Each one shows the frontal view of a face of one out of 23 different persons. The number of images per subject is variable, as is the background (usually cluttered like in an office environment). The positions of eyes are provided.

Figure 3.4: Images taken from all cameras of the CMU PIE database for subject 04006. The nine cameras in the horizontal sweep are each separated by about 22.5° [60].

3.7 Experimental Setup and Results

In this section we will evaluate the performance of our proposed 3D-ASM landmark detector. The experiments include using real scans and the randomly generated data. The experimental protocol is similar to the previous chapter, where we evaluate detected position of the eyes and the mouth or the another defined points with manually ground truth. It should be noted that, our 3D-ASM needs a initialization to give the coarse region of the face. This can be done by the well known Adaboost face detector [69] which is implemented and freely available in OpenCV library [11], and we used it for the initialization step. Also there are multi-view face detectors as described in [80], which work on different poses and can give the coarse pose categories. Although this can be helpful to give more precise initialization for our 3D-ASM facial landmarks detector, the study of multi-view face detectors is out of scope of our work. We used the OpenCV frontal face detector and the 3D-ASM is also initialized for frontal view faces. So we do not benefit of a prior pose information and make the hypothesis that the face is in frontal position.

3.7.1 Evaluation Using the Real Scans

In this experiment, we used the PCA coefficients of the 100 original scans to generate learning data and compare the performance of our 3D-ASM with C-ASM which we have introduced in previous chapter.

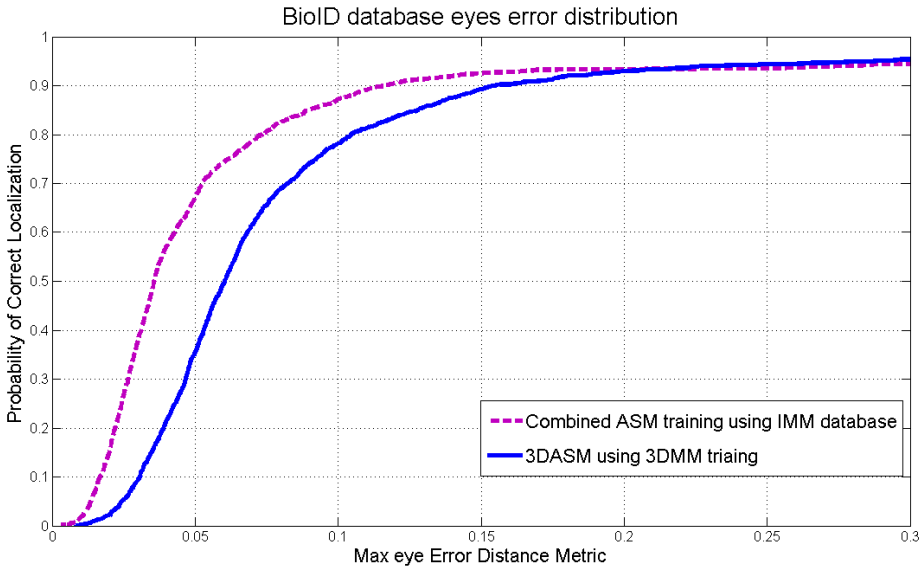

Figure 3.5: Comparison of our 3D-ASM and Combined-ASM from Chapter 2 on the BioID database for the two eyes.

Evaluation on the BioID Database

For comparison purposes, in Figure 3.5, we reproduce the published result related to the maximum error of the two eyes measurements. These results are compared with our 3D-ASM. From Figure 3.5, the performance of the 3D-ASM is much worse than Combined ASM facial landmark detector. One of the reasons is that no expression variation is present in the USF Human-ID 3D Database, while the images from BioID database are in near frontal view with expression variability. So the C-ASM is more suited to frontal face image with expression variability.

Experimental Results on the PIE database

We have done two experiments on the PIE database. In the first experiment we use the OpenCV face detector for the initialization, in order to evaluate the performance of the whole system including face detection and landmark location. In the second experiment, we suppose that we have the face detection, so the performance of 3D-ASM landmark detector is evaluated only. For first experiment, we choose 280 images from cameras c37, c05, c29, c11 and without expressions as shown in Figure 3.7. This limited choice is because the face detector [11] works only for those sets. The ground truth

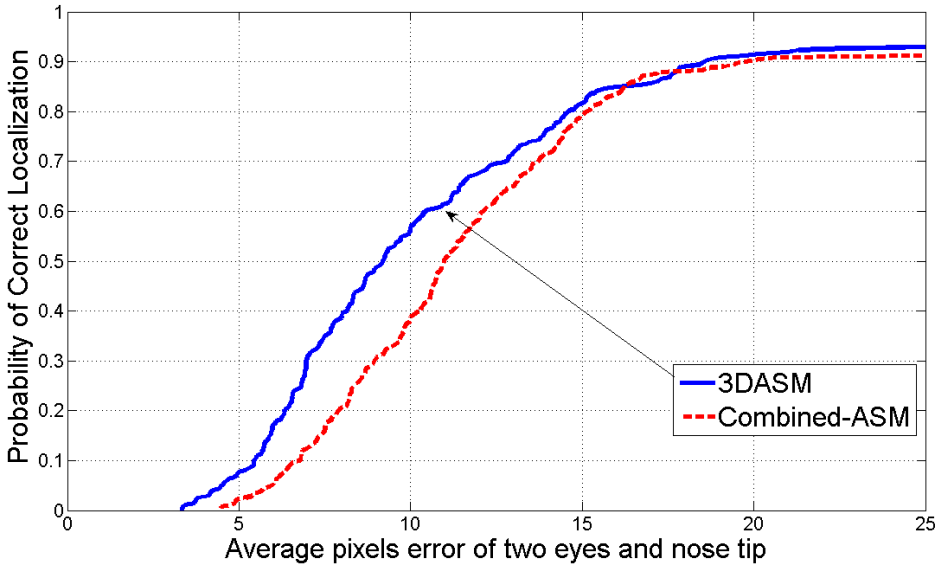

Figure 3.6: Comparison of our 3D-ASM and Combined-ASM from Chapter 2 on the subset of PIE Database.

Figure 3.7: Typical rotation images in PIE database [63]. From left to right image are captured by camera: c37, c05, c29, c11.

for the landmarks comes from [27], where we have position of eyes and noise tip. As shown in Figure 3.6, the 3D-ASM gives better results. By visual inspection we have also noted that the performance is not only better on eyes and noise tip, but also on mouth corners and contour landmarks, which are needed for the 3D face reconstruction. In this experiment, we were limited by the performance of the face detector.

In the second experiment, we assume that we have a multi-view face detector as described in [80], that works on different poses and can give the coarse pose categories. We use the manual labelled face region and pose categories to initialize our 3D-ASM facial landmarks detector. In that case, we can use 884 images from 13 camera to evaluate our 3D-ASM. The results are shown in Figure 3.8. The initiation of pose

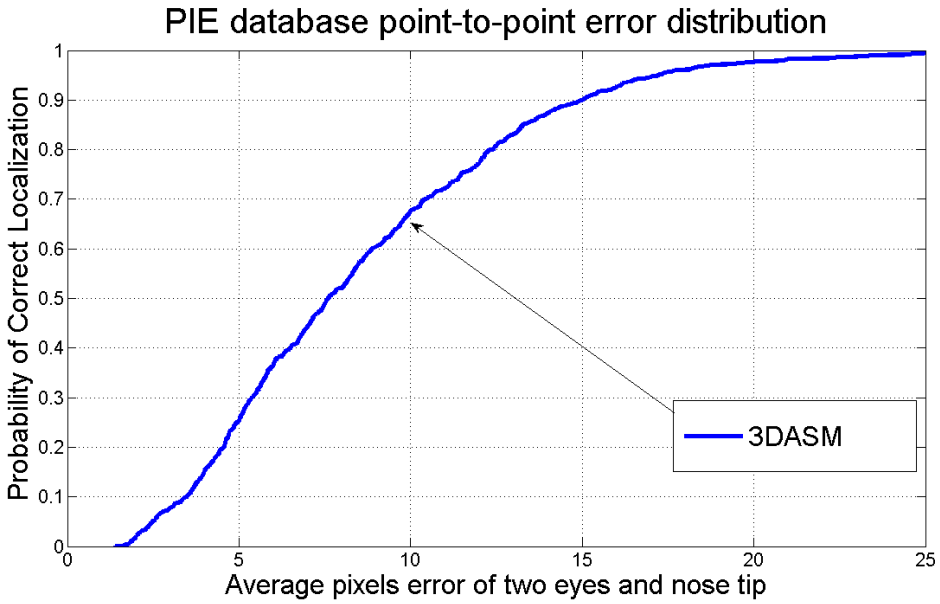

Figure 3.8: The 3D-ASM facial landmarks detector point-to-point error distribution on all 13 cameras on a subset of PIE database, for eyes and nose center points.

and facial region is important to our 3D-ASM algorithm. Actually this is a common characteristic of all ASM-based methods [20].

We also study the mean error and standard deviation corresponding to the camera position in Table 3.1. The nine cameras in the horizontal sweep are listed in the table, the order is according to the Figure 5.2, from left to right: c22, c02, c37, c05, c27, c29, c11, c14, c34. The out of plan rotation is a big challenge to the landmark location even if the face is well detected, the frontal face is more easy to locate than the profile face.

Table 3.1: Evaluation results on the PIE Database. Mean error (in pixels) of eyes and nose centers detection, of various different camera position.

Camera Position									
	c22	c02	c37	c05	c27	c29	c11	c14	c34
Mean error	13.1	8.3	6.5	6.1	4.4	5.7	6.7	7.6	14.0
standard deviation	2.2	2.1	2.5	2.4	2.0	2.2	2.2	2.0	3.1

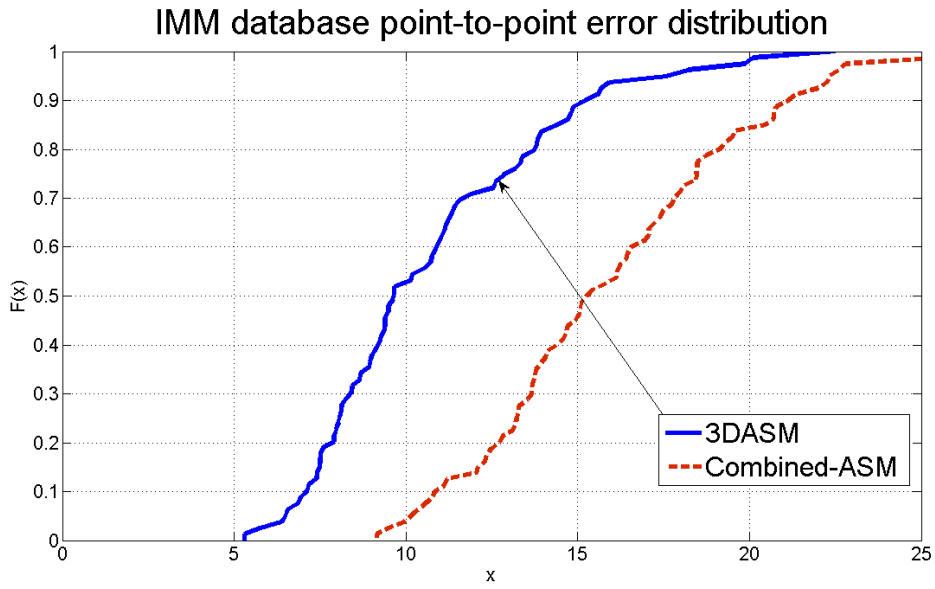

Figure 3.9: Comparison of 3D-ASM and Combined ASM on the subset of IMM Database (80 images) on 58 landmarks.

Experimental Results on the IMM database

From IMM Database we have chosen 80 images which have pose variation, and the landmark location result is evaluated on all the 58 points for each image. We take the mean error of the 58 points. From Figure 3.9, we can see that the proposed 3D-ASM gives better performance than the Combined-ASM, therefore validating our proposal of using automatically generated landmarks to train the 3D-ASM instead of manually annotated landmarks.

3.7.2 Evaluation Using Randomly Generated 3D Faces

To evaluate the influence of the training data to the landmark location precision, we have used different synthesized data from the 3D Morphable model to train our landmark detector. As in the "face space" of the 3D Morphable Model, a 3D human faces can be defined as a shape vector and a texture vector. So we can generate "new" 3D faces by randomly setting the shape and texture parameters, some typical examples could be found in Figure 3.10. In our experiment, two kinds of probability distribution functions are used to generate the random shape and texture parameters: uniform

distribution and normal distribution. The number of synthesized 3D face which are used for training are 100 and 300. In order to avoid no human like faces, the variation of shape and texture parameters are constrained within a reasonable interval, that is $(-0.8, 0.8)$ in our experiments. We choose this value empirically, if the value is too big we will have unrealistic examples.

From Figure 3.11, we can see that by increasing the number of randomly generated 3D faces for training, we can obtain better landmark detector, however the performance is still worse than using the 100 real scans. The statistical information of the 3D Morphable Model and the 3D-ASM are both learned from the real scanner data, so no mater how many synthesis data are generated for training the statistical information of the model is not increased. But the advantage is no manual annotations are needed.

Another way to increase the performance of the 3D-ASM landmark detector could be to increase the view categories for training. As we explained in Section 3.3.2, our 3DLTM is a discontinuous function to the texture patch from different view point, the more view categories we are using in the training phase the better performance we can obtain. For comparison, we render training data in 9 different views by setting roll angle to be: $(-90, -60, -40, -20, 0, 20, 40, 60, 90)$. In Figure 3.12, we can see the improvement.

3.8 Discussion

In this chapter, we proposed a 3D Active Shape Model for 2D landmark location on non-frontal face images. The novelty of our proposal is that we do not need to manually annotate the 2D landmarks on all training 2D images. Instead, we need to annotate manually only once the defined landmarks on the mean face of a 3DMM. We use this 3DMM for learning a 3DPDM which describes the prior of intrinsic changes caused by the characters of different persons in 3D space, and a 3DLTM which describes the prior of each landmark's local texture characteristic in different poses separately. Our fitting framework for landmark location is simple and efficient. We mainly compare the performance with our previous C-ASM (trained with manually obtained landmarks). The results show that our proposed algorithm based on automatically generated training landmarks gives better performances than C-ASM. Therefore we have validated the

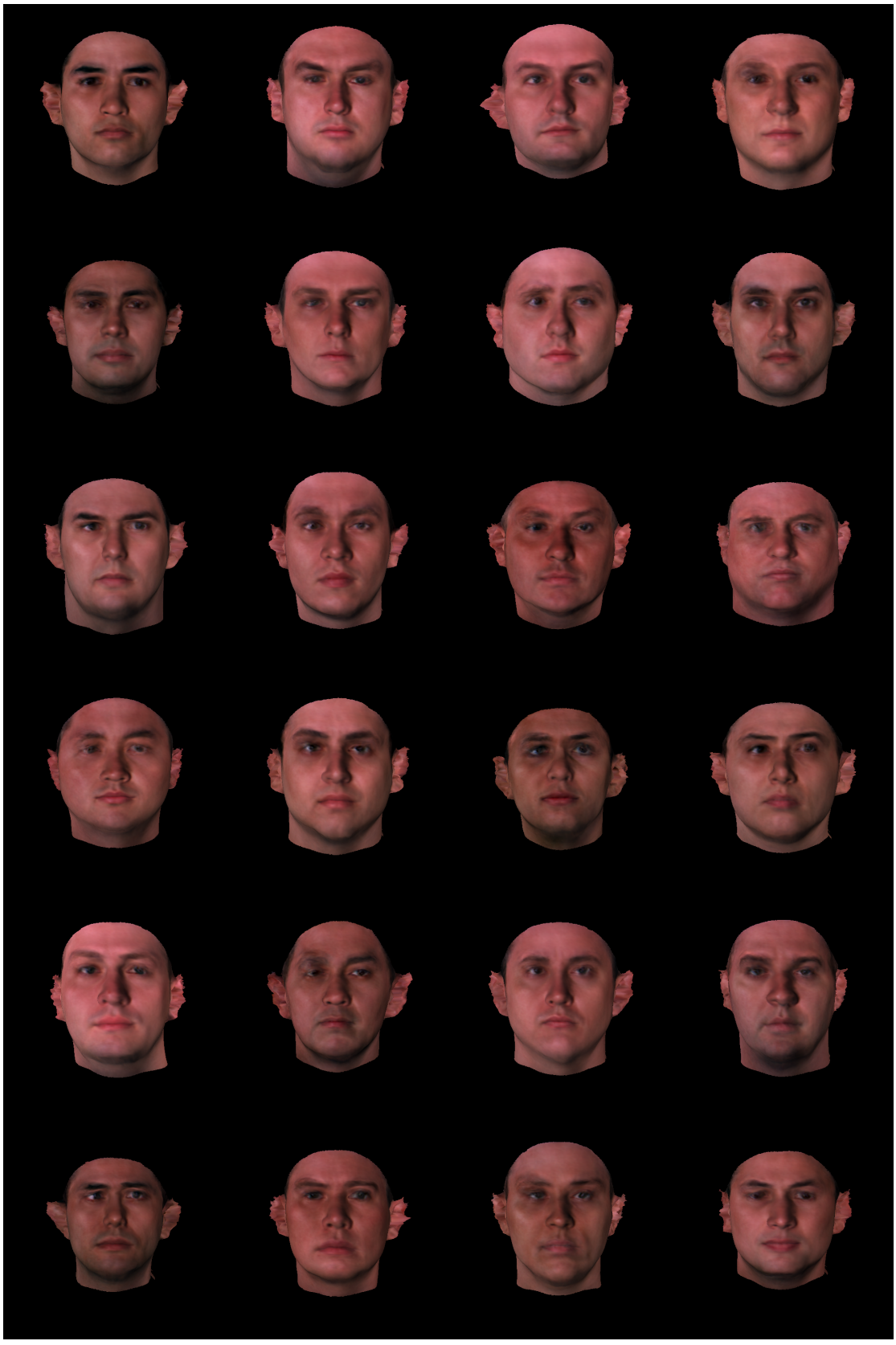

Figure 3.10: Typically randomly generated 3D faces from 3D Morphable Model.

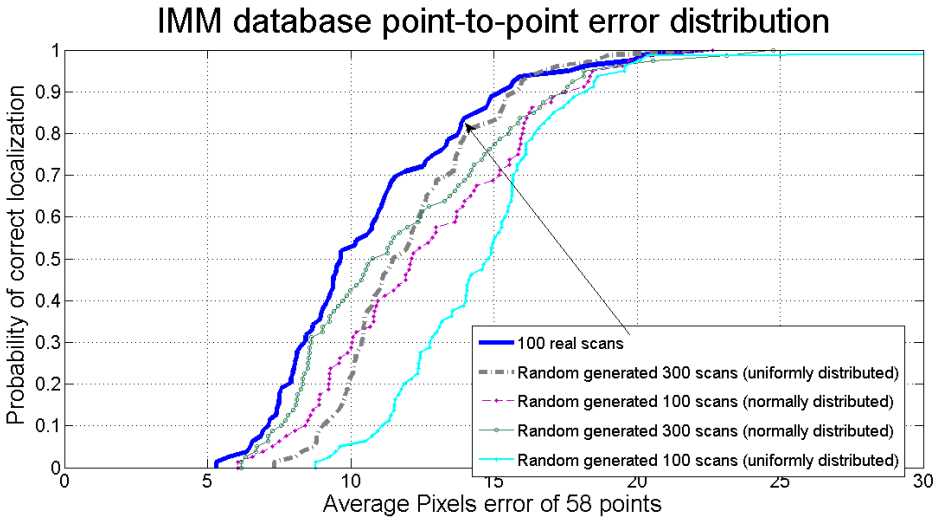

Figure 3.11: Influence of the training data to 3D-ASM. Comparison of landmarks location precision using different training data on the subset of IMM database.

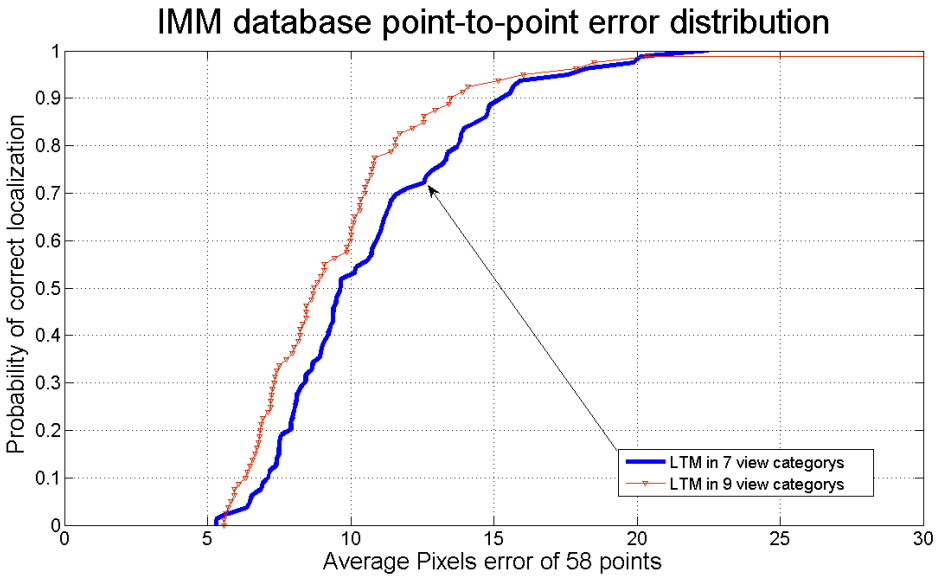

Figure 3.12: Comparison of landmarks location precision using different view categories for training.Evaluated on IMM database.

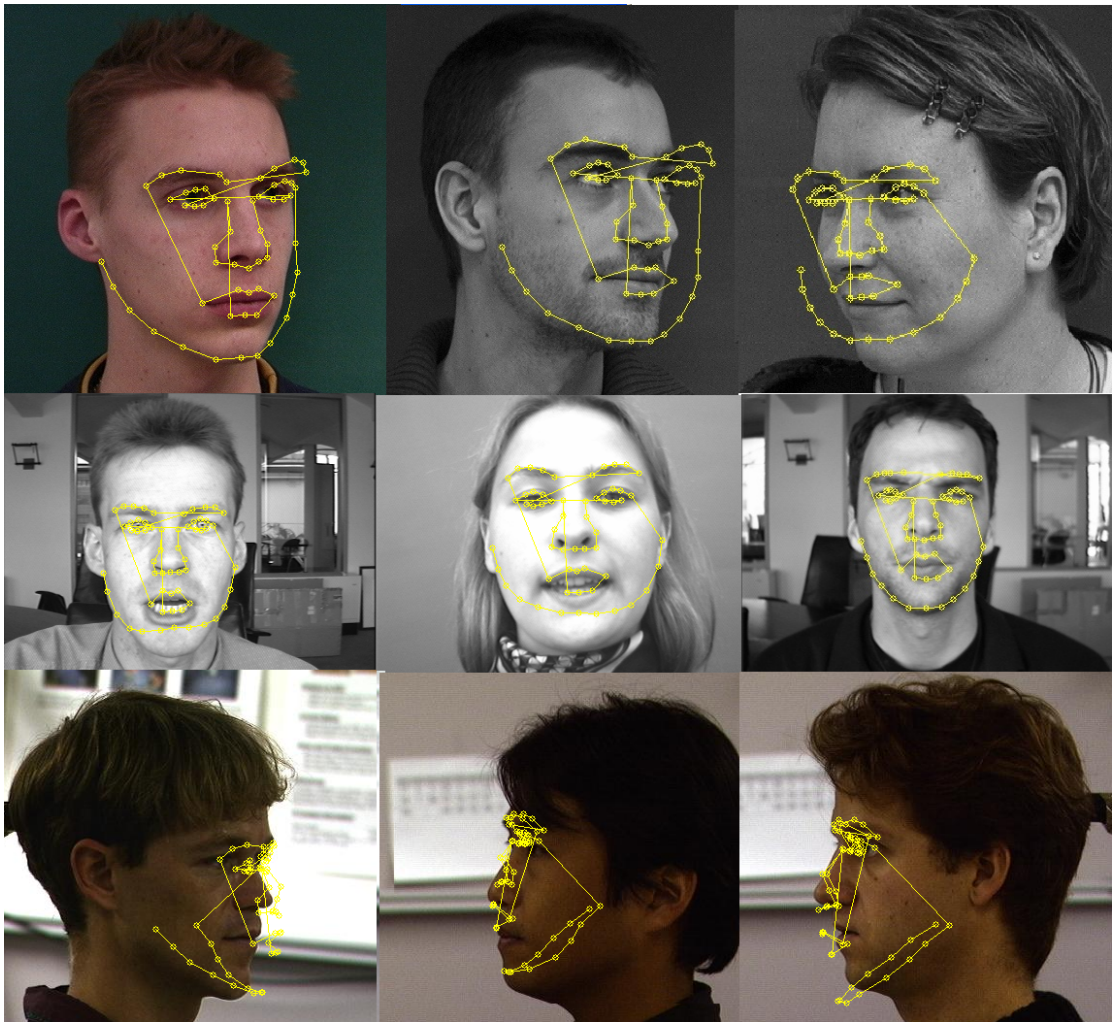

Figure 3.13: Error analysis: Bad landmark location examples from IMM, BioID and PIE database.

proposal of automatic landmark generation for training.

We only compare the 3D-ASM with our own system, but not with another published results. This is because, there are few public database suitable for evaluating the landmark location across pose. We don't have a common protocol, neither a standard definition for the landmarks, and also the ground truth landmarks are not easy to obtain. All of those make the comparison itself a big challenge.

In this work, we focus on solving the pose variation problem during the landmark location, but expression can be handled by the same framework by increasing expression variability in 3DMM. The 3D-ASM can not deal with expressions because the 3DMM that we used in USF human ID database is built with faces without expressions. As all facial landmarks location algorithms, the 3D-ASM is sensitive to the initialization step. In Figure 3.13, we show some difficult examples. Note that, the profile view of the faces is very difficult, and the landmarks we selected for the frontal view may be not suitable for the profile. The background has more effect for the side view images.

The proposed automated 2D landmark algorithm is exploited for of 3D facial reconstruction from 2D images in the next chapter.

Chapter 4

Automatic 3D Face Reconstruction from 2D Images

4.1 Introduction

During the last ten years, the 3D face reconstruction problem received a continuously increasing scientific attention [72]. There are many interesting applications that rely on 3D information, such as face recognition, human computer interaction and animation. Working with 3D data raises also many research challenges, such as model initialization, subspace learning, illumination effects, etc. There are different 3D face reconstruction methods which can be separated into three principal categories: reconstruction from a single image, stereo-based methods and video-based methods. Our main goal is to have a fully automatic system to reconstructed 3D model from single 2D image instead of using manual landmarks. In order to fully automate the 3D reconstruction process from a single 2D image, in provoius chapters we presented two methods for automatic facial landmark detection. Those methods are used for the initialization steps of 3D Morphable Model based face reconstruction.

In this chapter we are using the concept of analysis-by-synthesis loop introduced Blanz et al.(Blanz 1999) [9] as shown in Figure 4.1. We are interested in fully automatic 3D face reconstruction. We use the 3D-ASM landmark detector (introduced in Chapter 3) for automatic 3D face reconstruction from 2D nonfrontal face images. A 3D Active Shape Model (3D-ASM) is used to automatically detect 58 landmarks. Those landmarks are exploited to recover the initial pose and the main facial shape parameters of the 3D face model, followed by 3D Morphable Model (3DMM) fitting for face surface reconstruction. Our 3D-ASM landmark detector is a pose robust landmark location algorithm, whose training data originate from the 3D Morphable Model. The landmarks of the 3D-ASM have one-to-one correspondence between the 2D points detected from a 2D image and 3D points defined on the 3DMM. This kind of correspondence leads to a robust and precise initialization of the pose and coarse facial shape parameters. Then we fit the 3D Morphable Model to the input image by minimizing pixel-by-pixel color difference in an analysis-by-synthesis loop.

In this chapter we will introduce a novel technique to exploit the 3D Active Shape Model (3D-ASM) based 2D landmark detection in order to facilitate and improve the initialization step of the 3DMM.

The rest of chapter is organized as follows: first a brief literature review about

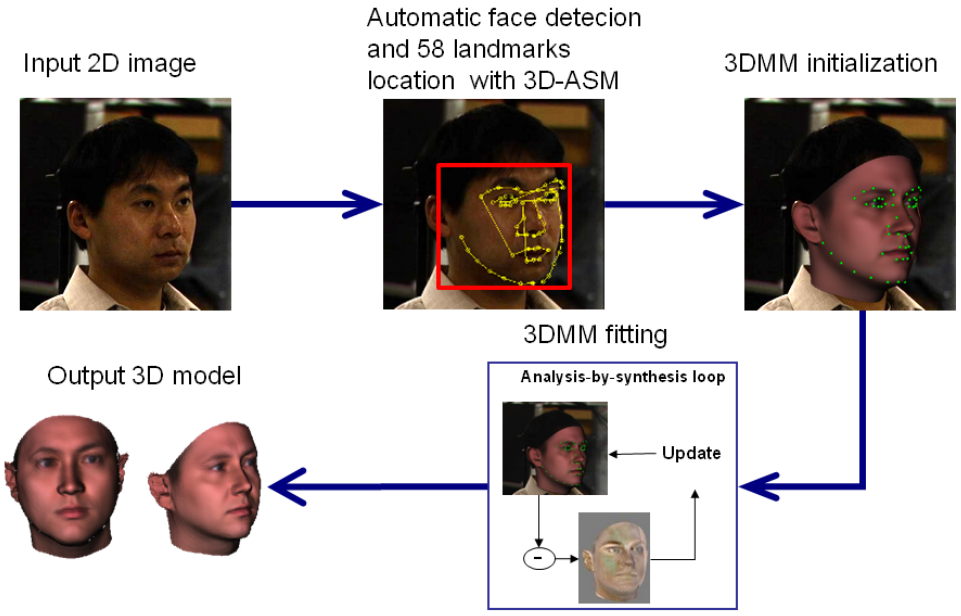

Figure 4.1: The framework of our automatic 3D face reconstruction algorithm from a single image with nonfrontal face. By using the landmarks detected by 3D-ASM, the pose of the 3DMM and the main facial feature are recovered. Example input 2D image is from the PIE database [60].

automatic face reconstruction is given in Section 2, then the main idea of our method of automatic 3D face reconstruction from 2D images, including using the 3D-ASM for automatic landmark location, model initialization and 3DMM-based face reconstruction is explained in Section 3. The evaluation methodology for the 3D face reconstruction is presented in Section 4. The databases that underlie our experiments and the results are shown in Section 5. In Section 6, we also study the problem about how the different view points and the image quantity effect the 3D face reconstruction results. The conclusions can be found in Section 6.

4.2 Literature Review Related to 3D Face Reconstruction and Its Evaluation

Among related works to 3D face reconstruction we will first summarize works in 3D face reconstruction, followed by recent research on automatic 2D landmark detection for 3D face reconstruction. Because the evaluation of 3D reconstructed faces is an important point, we will also report and comment some relevant publications in this

topic.

4.2.1 3D Face Reconstruction

As explained in (Widanagamaachchi 2008) [72], the first notable attempt to build a model of facial appearance using a 3D model was undertaken by Atick et al. (Atick 1997) [6]. In analogy with Sirovich and Kirby(Sirovich 1987), [61] and Turk et al. (Turk 1991) [68], they showed that human faces whether imaged or as surfaces have few degrees of freedom and thus can be represented with a relatively small number of parameters. Turk et al. applied PCA to a set of laser range scanned face surfaces represented in cylindrical coordinates to derive a set of eigenvectors describing perturbations from the mean head shape. They coined the modes of variation eigenheads. These eigenheads seem to capture modes of variation which are easily identifiable as facial characteristics. They found that an out-of-sample head could be represented with approximately 1% error using 100 modes of variation. Yan and Zhang (Yan 1998) [78] extended Atick et al's technique to allow the model to be fitted to nonfrontal images. However, it is clear that rendering a 3D head using Lambertian reflectance without variation in albedo yields very unrealistic images. Evidently, their model lacks sufficient complexity to realistically capture facial appearance. Nevertheless, they used a minimization technique to fit their model to frontal face images assuming known illumination and found that the recovered shape appeared qualitatively accurate. For synthetic Lambertian images the error was on the order of 2%, though this would obviously be higher for real world images with variation in texture and which exhibit non-Lambertian reflectance.

Blanz and Vetter (Blanz 1999) [9] enhanced this model by using a device which simultaneously captures shape and texture (in the three color channels). This allowed them to construct a statistical model (from a set of 3D scans) whose appearance parameters controlled both 3D shape and surface texture. Further, they used a method based on optical flow to find the dense correspondence between each head. This ensured that every vertex in the model corresponds to the same point on each face in the training sample. When combined with a complex rendering process which simulated camera settings and illumination conditions, near photo-realistic face images can be generated. One of the weaknesses of their approach is the lack of a realistic model of skin reflectance. They used the generic Phong model (Phone 1975) [53] which combines

ambient, diffuse and specular reflectance to capture the reflectance properties of skin. They use a similar technique to Atick et al. (Atick 1997) [6] to recover model parameters from a given input image of a face, though their optimisation procedure is far more complex. Besides shape parameters, they also adjust albedo, camera parameters, pose and illumination until an optimal match is achieved. This is a very computationally intensive process. But near photo-realism is achieved for input images. The technique also relies heavily on its optimisation procedure, which may return a local rather than a global minimum and is dependent upon a good initialisation. The reconstruction starts from a number of feature points (landmarks) on a face image. Those landmarks are used to align the pose of the 3D face model to the input image. In the majority of published works [9, 10, 39, 40], these landmarks are annotated manually. One promising research direction is to fully automate all the steps. Therefore, a 2D facial landmark location algorithm which is suitable for this purpose is required.

4.2.2 Automatic 2D Facial Landmark Location for 3D Face Reconstruction

In Chapter 2 we have reviewed the work for the frontal view facial landmark location. For 3D face reconstruction from single image we need a landmark detector which works on nonfrontal images for the initialization. In chapter 3 we introduce the previous works on the problem of landmark location across pose. Because we are interested in the fully automatic 3D face reconstruction from single 2D images, in this section we will complete the reviews presented in the Chapter 2 and 3 for automatic 2D Facial landmark location for 3D face reconstruction.

Hu et al. (Hu 2004) [33] proposed an automatic linear algorithm to recover the shape information from sparsely corresponding 2D facial landmarks. They first automatically detected 83 landmarks, and then 3D shape parameters were computed only from these landmarks. The method is reported to be efficient, but it works only with faces in frontal views with normal illumination. In (Breuer 2008) [12], the authors proposed an automatic facial landmark location algorithm, which is based on a classification algorithm (i.e. support vector machine), to initialize the 3DMM on nonfrontal faces. The authors created five view-specific component detectors (for nose tip, corners of the mouth, and external corners of the eyes) in order to detect facial landmarks. In order

to increase the robustness of the detectors, the processing is iterated using a criterion that is related on the 3D model based confidence measure. The evaluation of their 3D reconstruction algorithm is done with human visual inspection. Their results indicate that the automated algorithm is competitive in many cases, even though it does not fully match the quality of manual initialization.

4.2.3 Evaluation of the Quality of the 3D Face Reconstruction

Different criteria can be considered for the automatic 3D face reconstruction from 2D images. One straightforward method is with perceptual experiments where participants compare by visual inspection (Brerer 2008, Widanagamaachchi 2008) [12, 72] two reconstructed models. This kind of subjective evaluation is heavily affected by the relation among the tested faces, the subjects and poses. Other authors [10, 33] proposed to evaluate 3D face reconstruction by face recognition. With better 3D face reconstruction algorithms higher recognition accuracy is expected. But the performance of face recognition systems depends on both shape and texture, and the absolute geometric accuracy of the reconstructed face shape is still unclear. Le et al. [38] proposed a quantitative method to evaluate the accuracy of 3D face reconstruction algorithms. They suggest describing the shape difference between the reconstructed 3D faces and the 3D ground truth using Signal to Noise Ratio (SNR). They used synthetic 2D data as input images. Amberg et al. [2] have done quantitative evaluation using 3D scans and real 2D image from the same persons by using the geometric distance between the reconstructed 3D face and the 3D scans. Their evaluation is done on a database with only 20 subjects.

We would like to specify some differences and connections between our work and Breuer's work [12]. Both of them consider to automatically construct 3D face from a single image with pose variation. Our work focuses on the geometric precision of the reconstruction, and we propose an unified framework to quantitatively evaluates the accuracy of 3D face reconstruction algorithms, with possible applications for face recognition. While Breuer's work focused on visual inspection, their application is more about how to construct a photo-realistic representation for video games.

4.3 Automatic 3D Face Reconstruction from Nonfrontal Face Images

This chapter aims to automatically reconstruct a 3D face model from a single photograph under pose variations. Our task is similar to the one presented by Breuer et al. [12], but we focus on single 2D images, not video sequences. The framework of our automatic 3D face reconstruction algorithm is described in Figure 4.1. It is decomposed in three steps: 1. Face and landmark location, 2. Pose and shape initialization, and 3. Model fitting.

During the first step, we detect the region of the face and our 3D-ASM landmark location algorithm robust to pose variations is used to automatically detect 58 landmarks. In the second step those landmarks are exploited to align the 3D face model to the input 2D image (In Blanz et al. [10], 6 $\backsim$ 9 manual landmarks are used). During this initialization step, not only the pose is estimated, but also the major facial shape parameters are coarsely recovered using the large number of detected landmarks. After this initialization stage, we fit the 3D Morphable Model to the input image by minimizing pixel-by-pixel color difference in an analysis-by-synthesis loop. In such a way the computing time for the fitting part should be diminished, because the main facial shape parameters are already recovered in the initialization phase. Also thanks to this initialization with shape parameters, we can expect that the optimization of the analysis-by-synthesis loop should be more robust to local minima in the fitting phase.

4.3.1 Automatic 2D Face Landmark Location with 3D-ASM

In order to facilitate the fitting part, we propose to exploit the characteristics of an 3D-ASM 2D landmark detector. Active Shape Models need training data with manually annotated landmarks. In order to be independent of the manual annotation part, and to be able to choose specific landmarks, the 3D-ASM detector uses as training data synthetically generated from a 3DMM. The landmark points detected on 2D images with this 3D-ASM method have one-to-one correspondence to the corresponding 3D landmarks defined on the 3D model. This fact benefits the 3DMM fitting step. To build 3D-ASM, we use synthetic 3D faces data from the 3DMM [57]. The 3D landmarks can be easily manually located on 3D models, but it is more confusing to select them

on 2D images. Probably it is much easier for humans to find landmarks on corners and edges instead of visualizing 3D projections with self-occlusion. When input 2D images contain frontal faces, the 2D landmarks are much closer to the projection of the 3D landmarks [35]. But when 2D images represent faces in nonfrontal views, the manual annotation of 2D landmarks is more confusing. Figure 4.2 illustrates the difference between 3D landmarks and 2D landmarks for the same person in same head pose.

Furthermore during the landmark detection, 3D rotation parameters are taken into consideration. The landmarks detected by 3D-ASM on 2D images are considered as projections of 3D shape on the 2D image plan. Therefore we can recover the projection parameters and 3D shape information in the detected 2D landmarks, which also increases the precision of the following initialization step and induces robustness to pose variations.

As explained in Chapter 3, to obtain a 3D personalized face model, the fitting process starts from the 3D landmarks which are recovered from a set of 2D landmarks detected by 3D-ASM. The method recovers the shape and texture parameters of the face on the 2D image and is explained with further details in subsections 4.3.2 and 4.3.3.

4.3.2 3DMM Initialization Using Detected 2D Landmarks

Before using analysis-by-synthesis loop to optimize the global transformation $\boldsymbol{\theta}$, shape $\boldsymbol{\alpha}$, and texture $\boldsymbol{\beta}$ parameters, the 3D model needs to be aligned to the 2D image with reasonable initial pose and shape parameters. The more accurate the initialization is, the better 3D face reconstruction is expected. This processing is done in two steps: pose initialization and initialization of shape parameters.

Let $(\hat{x}_i, \hat{y}_i)$ be the estimated position of the i^{th} landmark on the 2D image. For those landmarks of interest, we also know the 3D coordinate of the corresponding vertex on the 3DMM. Let (X_i, Y_i, Z_i) be the 3D coordinates of the i^{th} landmark, and $(x_i, y_i) = P \circ T \circ S(X_i, Y_i, Z_i)$ be the projection of those points on the 2D image. The shape transformation with respect to the shape variation on the 3D Morphable Model of the vertex is denoted as $S(X_i, Y_i, Z_i)$. The 3D-coordinates of the vertices of the face model are defined according to an object centred coordinate system. A rigid body transformation applied to each (shape-transformed) vertex of the model is denoted as T, and P is the camera projection transformation. In our experiment we use week perspective projection, so only the focal length needs to be estimated during

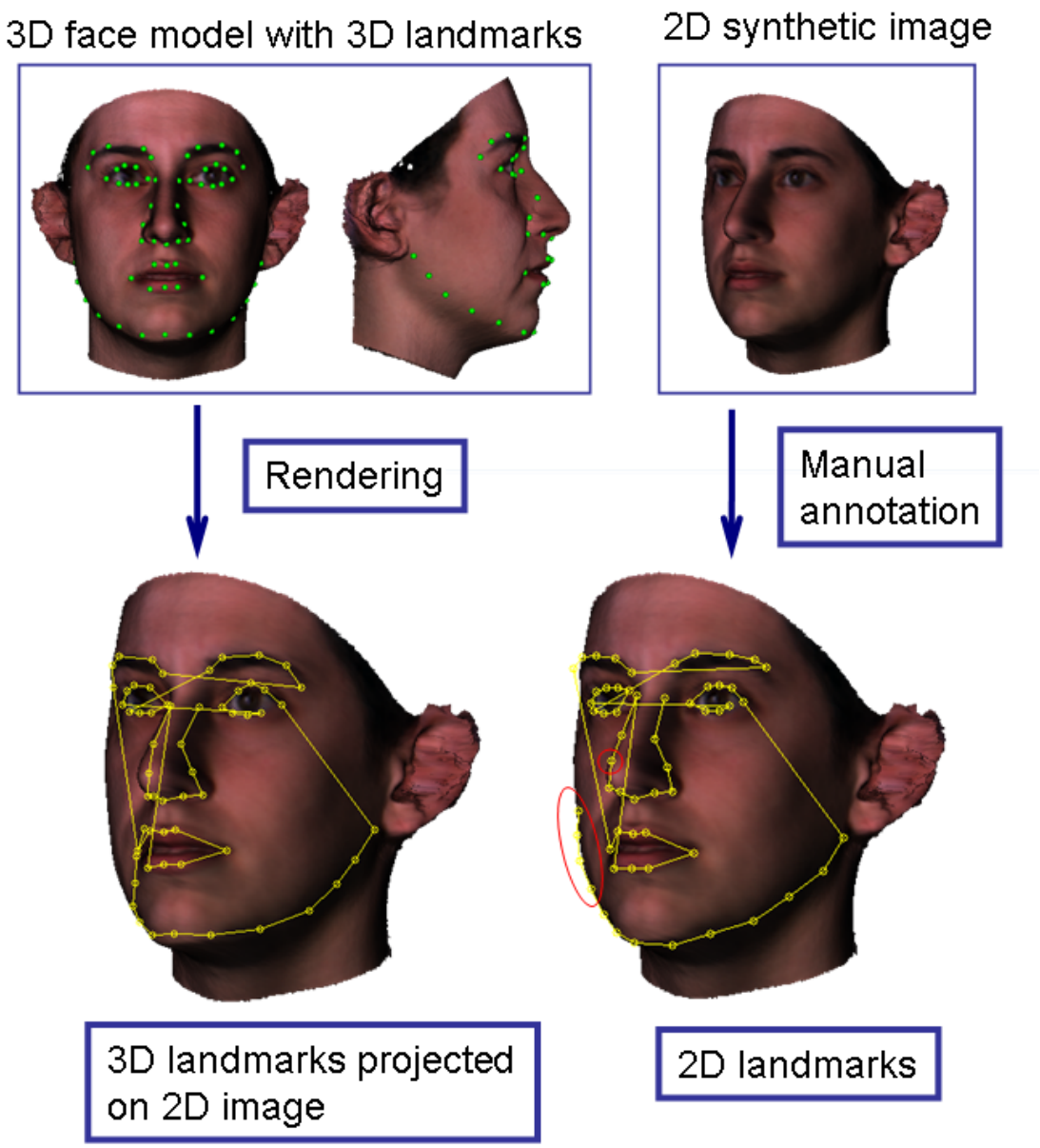

Figure 4.2: The difference of the 3D landmarks and 2D landmarks. The left image is the rendering image with 58 landmarks on the 3D model. The right one is 2D image rendering in same angle, while with 58 2D landmarks manually located on 2D image. The red points show the significant different points. The 3D scan data are generated from USF database [57]

.

the optimization. Another camera projection models can also be applied. The energy function we minimize is the following:

$$E_{Initial} = \sum_{i=1}^{N_p} (\hat{x}_i - y_i)^T (\hat{x}_i - y_i) \tag{4.1}$$

where N_p is the total number of landmarks (58 in our experiment). During the pose initialization step, the initial model position is found by minimizing $E_{Initial}$ with respect to the 6 parameters related to rotation, translation and focal length.

After a coarse pose is estimated, it is possible to further optimize $E_{Initial}$ with respect to the model shape parameters, since the number of detected landmarks is large enough to render a unique solution. Considering the shape prior, the objective function can be modified to:

$$E_{Initial} = \sum_{i=1}^{N_p} (\hat{x}_i - y_i)^T (\hat{x}_i - y_i) + \lambda_S \sum_{j=1}^{n} \frac{\alpha_j^2}{\delta_j^2} \tag{4.2}$$

where n is the number of the shape model, α_j is the j^{th} element of the shape parameter $\boldsymbol{\alpha}$, and λ_j is the corresponding shape eigenvalue of the model. This energy consists of two parts: the first part measures the difference between the detected 2D landmarks and corresponding 3D landmarks projection positions. The second part is the shape prior which constraints the shape deformation to reasonable values, so that we can avoid non-face-like surfaces. The parameter λ_S, which we take proportional to the sum of all the weights in the first part, allows us to balance the influence of matching quality and shape prior probabilities. We minimize this energy using Levenberg-Marquardt algorithm.

After the initialization step, the pose and the main shape parameters are recovered. For the final reconstruction, we fit all model vertices to the image in an analysis-by-synthesis loop that optimizes all facial details and compensates for lighting and other imaging parameters.

4.3.3 3D Face Reconstruction by Model Fitting

In the third stage of our method, the full facial surface is reconstructed by fitting the 3DMM to the input image. The 3DMM uses a linear subspace (i.e. a PCA) to model the facial shape and texture from 3D scans. The coefficients of shape and texture model define person intrinsic variations (such as identity). The objective of the

fitting is to minimize pixel-by-pixel color difference using the analysis-by-synthesis loop method. Our fitting processing is similar to Blanz and Vetter [9], except we use Gauss-Newton instead of Newton algorithm for the nonlinear optimization. The analysis-by-synthesis aims not only to optimize these coefficients but also the pose (rotation, scale and transformation), color and intensity of directed light and ambient light, color contrast as well as gains and offsets in each color channel. Once the shape and pose parameters are recovered, the face texture can be extracted from the 2D input image to make the 3D face model more realistic. With our proposed initialization phase the computation time for the fitting process is diminished, because the main facial shape parameters are already recovered in the initial phase. Also thanks to the initialization of shape parameters, the optimization of the analysis-by-synthesis loop should be more robust to the local minima problem that can occur during the fitting phase.

Face image synthesis defines the positions of vertices of the 3-D model with illumination and colour. During the processing of fitting a model with a novel image, not only the shape and texture parameters α_i and β_i are optimised, but also the following rendering parameters are optimised. There are 22 rendering parameters concatenated into a vector ρ:

- pose angles ϕ, θ, and γ
- 3-D translation $\mathbf{t}_w$
- focal length f
- ambient lighting intensities $L_{r,amb}, L_{g,amb}, L_{b,amb}$
- directed light intensities $L_{r,dir}, L_{g,dir}, L_{b,dir}$
- the angles θ_l and ϕ_l of the directed light
- colour contrast c
- and gains and offsets of colour channels $g_r, g_g, g_b, o_r, o_g, o_b$

In analysis-by-synthesis iterations, the fitting algorithm finds model parameters and rendering parameters, and produces an image as similar as possible to the input image $\mathbf{I}_{input}$ as shown in Figure 4.3. The goal of the fitting is to find shape and texture

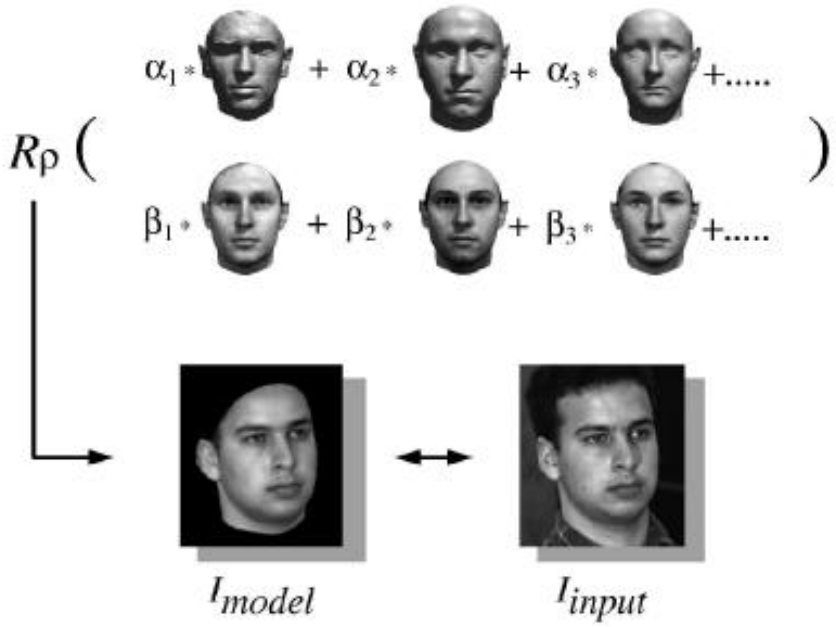

Figure 4.3: Fitting a morphable model: analysis by synthesis iterations [10].

coefficients α and β such that rendering R_ρ produces an image $\mathbf{I}_{model}$ that is as similar as possible to $\mathbf{I}_{input}$.

It should be noted that for initialisation, in the work of Blanz et al. [10], seven facial feature points, such as the corner of the eyes or the tip of the nose, are marked in image coordinates. On the morphable model, these 7 points are also defined as vertices of the mesh corresponding to the points in the image. While in our work, the 58 automatic detected landmarks are used in the same way. The primary objective in analysing a face is to minimise the sum of square differences over all colour channels and all pixels in the input image and the symmetric reconstruction

$$E_I = \sum_{x,y} \|\mathbf{I}_{input}(x,y) - \mathbf{I}_{model}(x,y)\|^2 \tag{4.3}$$

A stochastic version of Gauss-Newton's method is used to minimise the cost function in the fitting procedure. Because the face model is separated into four regions - eyes, nose, mouth and the surrounding face area, the optimisation is also separated by each region to obtain local parameters, i.e., $\alpha_{r_1}, \beta_{r_1}, \ldots, \alpha_{r_4}$ and β_{r_4}.

4.4 Evaluation Method for 3D Face Reconstruction

In order to evaluate the quality of the 3D reconstructed models, we have chosen to compare the shape of the reconstructed 3D data with the shape of the ground truth of the same subjects obtained from a 3D laser scanner. For this purpose the multimodal biometric database IV^2 [50] is well suited. For more then 100 subjects this database contains various 2D images (that we are going to use as input 2D images for the 3D reconstruction), and 3D laser scans (from where we can obtain the ground 3D truth for the same subject). The Mean Squared Error (MSE) of the geometric distance over the surface is used to measure the shape difference. In this chapter we evaluate the 3D face reconstructed results using geometric distance, in next chapter the performance of the 3D face reconstruction will be evaluated by face recognition.

The framework of our 3D face reconstruction evaluation methodology is illustrated in Figure 4.4. For each person, one 2D nonfrontal image is selected as the input image for the 3D face reconstruction. The evaluation aims to compare the 3D reconstructed face and the ground truth available in the IV^2 database.

In order to measure the shape difference between the reconstructed shape and the ground truth, an alignment step is necessary. We use a 3D alignment phase which can be separated in two parts using a coarse-to-fine strategy. The coarse step is based on a manual annotation in which the user must select three points on the ground truth (the outer corner of left and right eyes and the nose tip). This manual step is only needed in the evaluation phase, not for the 3D face reconstruction. The corresponding points of the 3D reconstructed face could be easily obtained since the index of those points on the 3DMM is known. Thanks to these corresponding points, a coarse alignment of the reconstructed face to the ground truth can be done with an affine transformation. Then we apply a fine alignment which finds the minimal distance between two surfaces starting from the last initial solution. This step is based on the well-known Iterative Closet Point (ICP) algorithm [7]. It is an iterative procedure minimizing the MSE between two surfaces. At each iteration of the algorithm, the geometric transformation that best aligns the 3D scan and the 3D reconstructed face model is computed. Let $\boldsymbol{P} = p_0, ..., p_i, ..., p_N$ be a set of points on the 3D reconstructed face, and $\boldsymbol{Q} = q_0, ..., q_i, ..., q_M$ the corresponding points on the 3D scan. The goal is to find the rigid transformation (R, t)

which minimizes the distance between these two sets of points. The rigid transformation (R, t), minimizing the least square criterion bellow, is calculated for each iteration.

$$e(R,t) = \frac{1}{N}\sum_{i=1}^{N} \|(Rp_i + t) - q_s\|, \tag{4.4}$$

where q_s is the nearest point on the 3D scan to p_i. Since we are using the week perspective camera model during the 3D face reconstruction step, the absolute scaling of the reconstruction face is unknown. So during the ICP alignment, the scale parameter s is added to the rigid-body transformation. The criterion is:

$$e(R,t) = \frac{1}{N}\sum_{i=1}^{N} \|s(Rp_i + t) - q_s\|. \tag{4.5}$$

The ICP algorithm presented above has the problem that it can sometimes converge monotonically to a local minimum. So during our experiment we discarded the 3D reconstructions for which the ground truth scans conducted to local minima during the ICP fine alignment step. The 3D scans that caused this errors presented for example a large hole on the mesh.

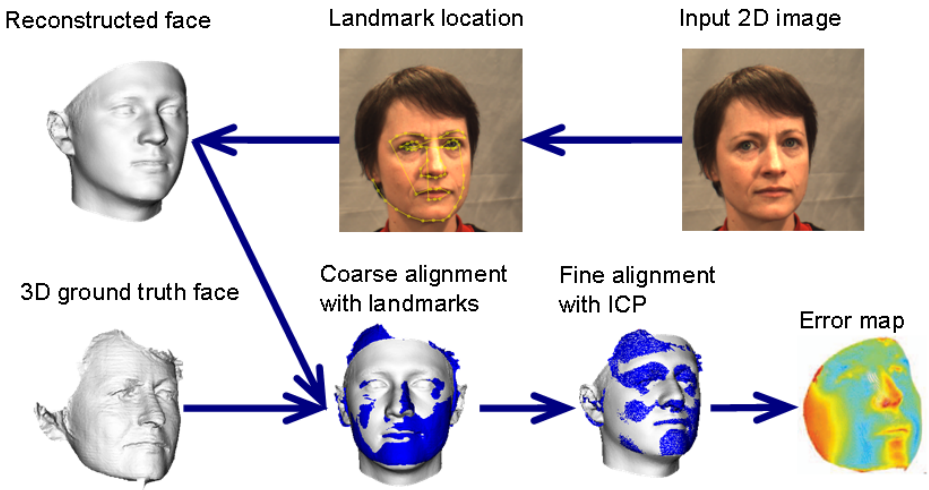

Figure 4.4: The framework of our 3D face reconstruction evaluation protocol. The input 2D image and the 3D ground truth scan are from the IV2 database [50].

4.5 Database and Experimental Results

4.5.1 Databases

In this section more details about the databases that underlie our experiments and the results are given. The USF Human ID database [57] which contains the 3D Morphable Model is used for training the 3D Active Shape Model. The IV^2 Multimodal Biometric Database [50] is used for the evaluation of the reconstructed 3D personalized models.

USF Human ID 3D face database: The USF Human ID 3D face database [9] consists of 136 scans of 136 subjects acquired with the Cyberware 3030PS laser scanner. There are more than 90K vertices and 180K triangles for each face model.

IV^2 **Multimodal Biometric Database:** The IV^2 database [50] contains face and iris data. Among the the face data we extracted from the videos 2D face images with some pose and illumination variability, and the 3D facial data acquired with a laser scanner (Minolta Vivid 700), as the ground truth. The resolution of the 2D images are 640×480, and the distance between two eyes is about 40 pixels. The resolution of the 3D scans (are manually merged from 3 partial original scans taken from 3 views (left, right, frontal) respectively) have about 7000 vertices and 13000 triangles. This database is composed in total of 430 records from 315 different subjects, 219 subjects have only one session, 77 subjects with two sessions and 19 subjects with three sessions. It should be noted that there are 104 subjects that have glasses in the acquisition and 45 subjects have beard. And there are some subjects that have incomplete 3D scans. We have to ignore those subjects during our experiment. The total number we can use for our experiment is 68 subjects.

In our experiment, in order to have head pose variation, we extracted from the IV^2 frontal stereoscopic video data one 2D input image for the 3D face reconstruction data. For practical reasons the 2D images were extracted at a given timestamp, where some illumination variability is present, and the head is not in a frontal position. From those extracted images, we ignored some bad quality images due to the out of focus, no face present, image blurred or subject with beard and glasses. In total, we obtained 87 such images from 68 subjects, for which the 3D models were constructed. Notice that one of the reasons that we choose the frontal stereoscopic video data is that they contain

the head pose variation with roll angle approximatively equal to (-40,40).

4.5.2 Experimental Results

In our experiment, the USF Human ID Database [9] is used to train the 3D-ASM 2D landmark detector. The 3DMM available also with this database is used for the fitting part. The evaluation of the 3D reconstructed models is conducted on the IV^2 database. The 2D images from IV^2 stereoscopic video are extracted and are used as the input images for the 3D face reconstruction.

For comparison purposes, for each input image we built three 3D models, with different landmarks for the 2D input image. Three different landmark location algorithms are used for 3D face reconstruction:

- C-ASM
- 3D-ASM
- manually labelled landmarks

For comparison, related to the precision of the 2D automatic facial landmarks algorithms, another set on automatically obtained 2D landmark points with the Combined Active Shape Model (C-ASM) and 2D manual landmarks, are used in the 3D face reconstruction step. So for each input 2D image, we compared three 3D reconstructed models. Typical reconstructed face models are shown in Figure 4.5. The experimental and qualitative results of the proposed method are illustrated as in Table 4.1 and Figure 4.6. In Figure 4.6, the histogram of the face reconstruction over 87 input images are plotted. Figure 4.6 shows that the 3D-ASM gives good performance and even better than the 2D manual landmarks on nonfrontal face images. The explanation could be the following: the 3D-ASM is training with the 3DMM which is also used for face reconstruction. The 2D landmarks detected by 3D-ASM have a natural correspondence with 3D points defined on the 3DMM. Also during the landmark detection, the 3D-ASM considers 3D out-of-plane rotation parameters. This leads to a better initialization for the analysis-by-synthesis loops. While the 2D manually annotated landmarks suffer probably from the problem depicted in Figure 4.2, that it is not so easy for humans to precisely locate 2D landmarks on nonfrontal face images.

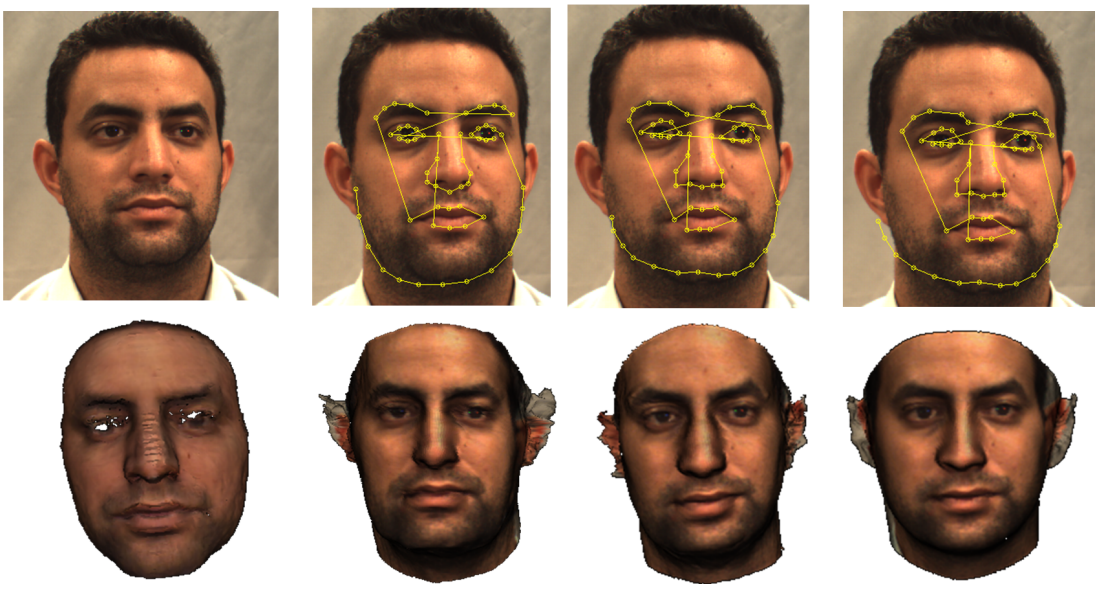

Figure 4.5: 3D face reconstruction using three different landmarks for initialization. First column: the input 2D image (above) and 3D ground truth scan (bellow). Second to fourth column: three different landmarks detected on the 2D image (above) and the corresponding 3D face reconstruction results (bellow), from left to right: CASM, 2D manual, 3D-ASM landarks. The input 2D image and the 3D scan are from the IV^2 database [50].

Table 4.1: Performance of the 3D face reconstruction initialized by 3D-ASM, CASM and 2D manual landmarks in a side by side comparison. STD = standard deviation.

	CASM	2D manual	3D-ASM
MSE (mm)	2.85	2.63	2.38
STD (mm)	0.98	0.96	0.75

In Table 4.1, we list the average Mean Squared Error and the standard deviation of the 87 face reconstruction results using the three different landmarks detection methods. From the Table 4.1 and the reconstructed 3D face examples in Figure 4.5, the results shows the difference between different landmarks location strategy are not so significant.

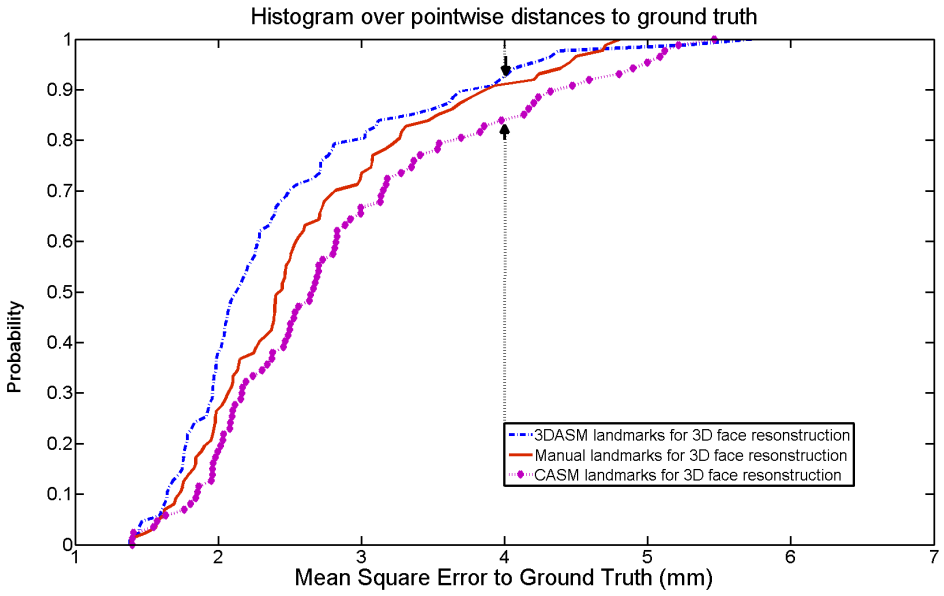

Figure 4.6: Histogram of geometric Mean Squared Error distance from the 3D reconstructed models to the ground truth surfaces.

4.6 Influence of View Point Change to the 3D Face Reconstruction Results

In this section, we will study the problem how the view angle of the 2D images affects the 3D face reconstruction. Our reconstruction algorithm only needs single input image for 3D face reconstruction. To construct a 3D shape by using analysis-by-synthetics loop, the input image could be take from different view points.

Figure 4.7 explains this problem, the left column shows the input images for face reconstruction, from top to bottom which are taken from frontal, side and profile view separately. The reconstruction results are rendered in the right three columns. They look different. Take the first row for example, the 3D reconstructed face looks more similar in frontal view rather than from another point of views.

To study which is the optimal view point for the face reconstruction, we use synthetic head pose database which we generated from the 3DMM and the evaluation method described in Section 4.4. Firstly we randomly generate 50 3D faces from the 3D Morphable Model, and then we render the 3D face by setting roll angle from -90 to 90 degree, 10 degree for each image, see Figure 4.8. So we have 50x19 images from 19 different view points. For each image, we reconstruct a 3D face. The ICP distance between the reconstructed 3D face and 3D faces which are used to render the synthetic head pose database is exploited to measure the influence of the view point related to the reconstruction.

The experimental results are show in Figure 4.9. We find out that the frontal view seems to be the optimal view for 3D face reconstruction. While images captured from profile views are less suitable for 3D face reconstruction.

4.7 Influence of Image Quality to the 3D Face Reconstruction Results

It should be noted that, the acquisition view point is not the only factor that affect the 3D face reconstruction performance, though it is a important one. For example the resolution and quality of the input image are also important. During our experiment, the best resolution for the face reconstruction is 512x512, with 50 to 100 pixels between

Figure 4.7: Typical examples of 3D face reconstruction from different view points. The left column show the input images for face reconstruction. The second to the fourth columns present the corresponding reconstructed 3D face rendered in frontal, side and profile view separately.

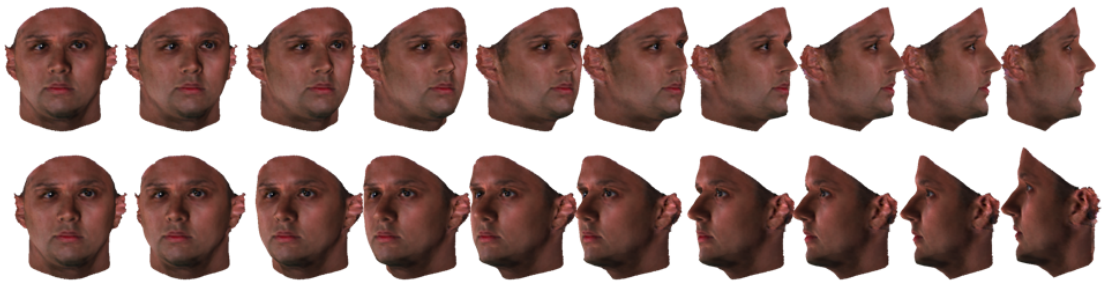

Figure 4.8: Examples of the synthetic head pose database for the evaluation of the influence of the view point to the 3D face reconstruction precision. The top row are the synthetic image rendered by setting roll angle from 0 to 90 degree, 10 degree fro each image. The botton row are the synthetic image rendered by setting roll angle from 0 to −90 degree, 10 degree fro each image.

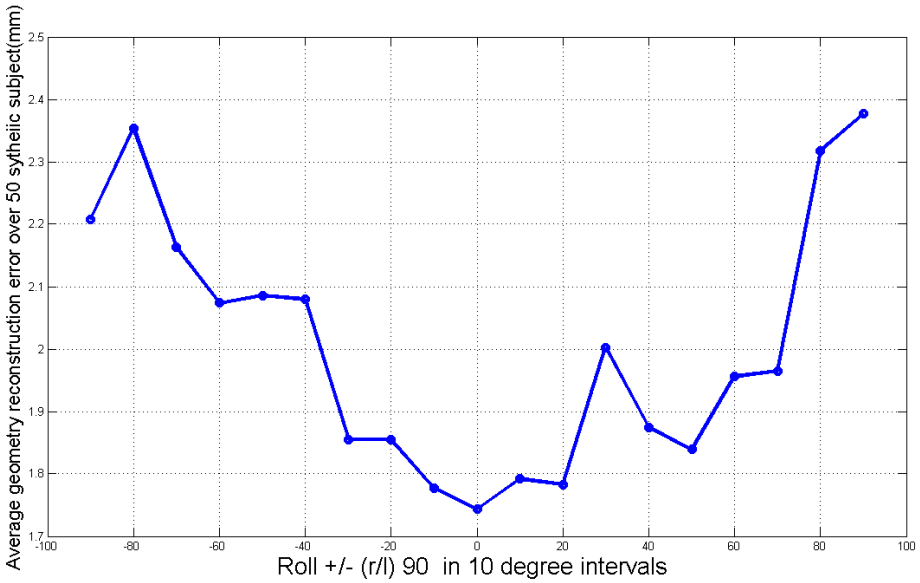

Figure 4.9: Evaluation result of the influence of the view point various to 3D face reconstruction algorithm .

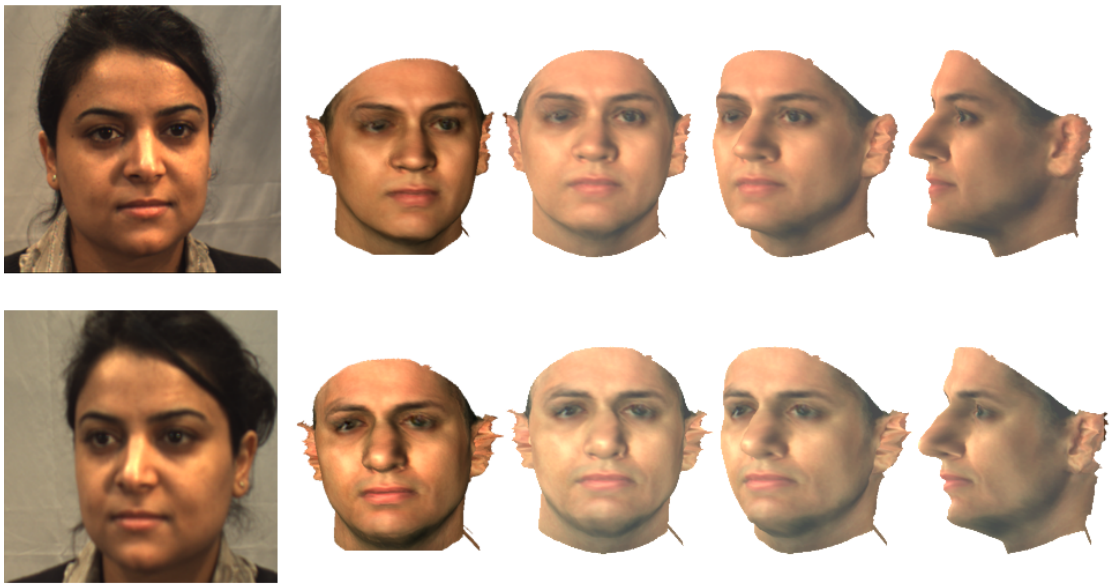

Figure 4.10: The influence of image quality to the 3D face reconstruction results. The left column show the input images for face reconstruction with different quality. The second column is the reconstructed face rendering with the illuminate and pose parameters extracted from the input image. The third to the fifth columns list the correspondence reconstructed 3D face rendered in frontal, side and profile view separately.

the two eyes. Actually this is the same resolution of the reference plan image of 3D Morphable Model in the USF Human-ID database. Given an input image , we first have to re-sample it to this resolution. Compared to the image resolution, the image quality is more important. In here the high quality image means no-blurred face image. Figure 4.10 shows face reconstruction results with blurred and no-blurred face images taken from the some subject. We can see the algorithm is sensitive to the image quality. The reason could be the following: the 3D face reconstruction using 3DMM is done by analyse-by-synthesis loop. During the fitting, the pixel value on the synthesis image where the 3D face model is projected depends on two parts: the texture values from the 3DMM and the environment illumination condition. In our implementation we use the Phone model, which assumes one pixel value is the reflection of the 3D model corresponding vertex. This is not the case when images are blurred, and we can not find one by one correspondence between them, in other word blurred image does't fit our Phone light model hypothesis. In Figure 4.10, we only show the example mapping with the texture from 3DMM, the influence of the different texture mapping strategies (texture from 3DMM or texture from input image) are discussed in next Section.

4.8 Influence of Texture Mapping Strategies to the 3D Face Reconstruction Results

In Figure 4.11, the reconstructed 3D faces with different texture mapping strategy are shown. Taking texture directly from 2D input images gives more details of the human face and makes the result more realistic for visual inspection. While 3D reconstructed faces mapping with the texture from the 3DMM looks smoother and reduces the influence of illumination. It would be interesting to combine the two kinds of texture to improve the 3D reconstruction result, and to study their influence on the face recognition. More details about the evaluation will be given in the next chapter.

4.9 Conclusions

This chapter presents a fully automated algorithm for reconstructing 3D models of face from single photograph with nonfrontal faces. The algorithm is based on a

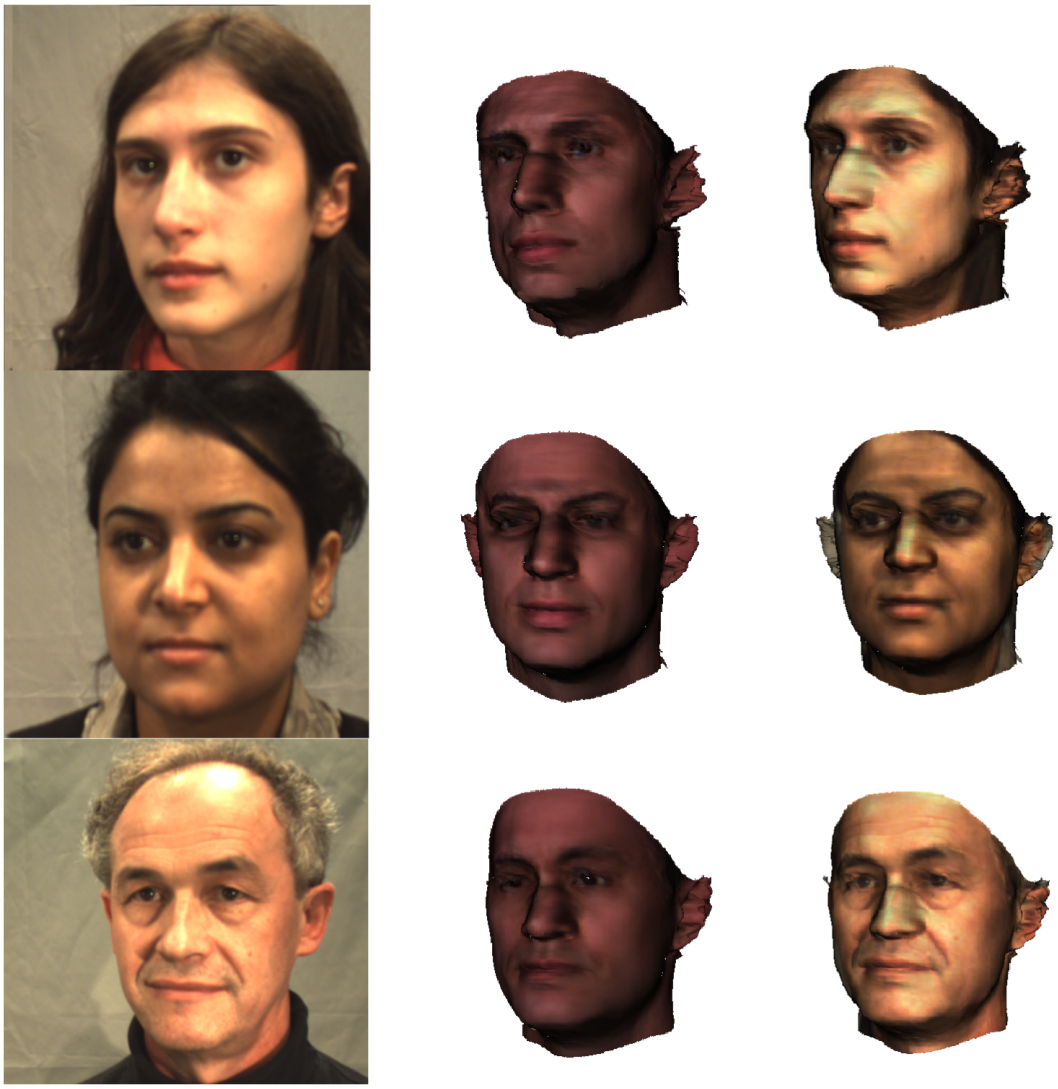

Figure 4.11: Typical 3D face reconstruction results using 3D-ASM landmark lactation for initialization. First column: the input 2D images. Second column: 3D reconstructed faces mapping with the texture from the 3DMM. Third column: 3D reconstructed faces mapping with texture extracted from the input 2D images . The input 2D images are from the IV^2 database [50].

combination of 3D-ASM and 3DMM. We evaluate the face reconstruction result quantitatively on the IV^2 Biometric Database. Different automatic landmark location algorithms are compared. The influence of the precision of landmark location due to the 3D face reconstruction is evaluated. The results show that the 3D-ASM provides excellent initialization for the 3D face reconstruction with nonfrontal faces. It seems to be even better compared to the manual annotation for nonfrontal images.

Our contribution is also related to the usage of the 3D-ASM for the step of automatic landmark location on the 2D image. Thanks to the characteristics of this step, the processing should be more robust to the local minima problem that can occur using the fitting phase.

Chapter 5

2D Face Recognition across Pose using 3D Morphable Model

5.1 Introduction

Our main purpose is in this chapter is to continue the validation our implementation of the 3D reconstruction procedure which presented in previous chapter. In Chapter 4, we use the 3D geometric distance between the 3D scanner and the reconstructed 3D face from 2D image to evaluate our 3D face reconstruction accuracy. As explained in previous chapter, the face reconstruction could also be evaluated by 2D face recognition indirectly, the better face reconstruction result we have the better face recognition rate we should obtain. In this chapter we will introduce how to use our fully automatic 3D face reconstruction algorithm to solve the problem of pose variations problem in the field of 2D face recognition. In previous chapter, our focus was about the 3D shape information of the face reconstruction, while in this chapter the texture information is also taken in account.

The rest of this chapter is organized as follows: first a brief literature review about face recognition across pose is given in Section 5.2. Reminders about using the 3D Morphable Model for 3D face reconstruction and the experimental protocol are given in Section 5.3. The proposed algorithm of coefficients-based comparison, the ICP distance of the 3D surfaces based measure and viewpoint normalization approach are explained in Sections 5.5, 5.4 and 5.6 respectively. Finally, the conclusions can be found in Section 5.7.

5.2 Brief Literature Review about Face Recognition across Pose Problem

In recent surveys of face recognition techniques by Zhao et al. (Zhao 2003) [81] and Tan et al. (Tan 2006) [66], pose variation was identified as one of the prominent unsolved problems in the research of face recognition. Therefore it gains great interest in the computer vision and pattern recognition research community. This is important in the degraded condition such as video surveillance. Consequently, a few promising methods have been proposed to tackle the problem of recognising faces in arbitrary poses, such as tied factor analysis (TFA) introduced by Phllips et al. (Phillips 1998) [52], 3D Morphable Model (3DMM) introduced by Blanz et al. (Blanz 2003) [10], eigen light-field (ELF) (Gross 2002) [28] introduced by Gross et al., illumination cone

model (ICM) (Georghiades 2001) [25] introduced by Georghiades et al., etc. However, none of them is free from limitations and is able to fully solve the pose problem in face recognition. Continuing attentions and efforts are still necessary towards ultimately reaching the goal of pose-invariant face recognition.

According to Zhang et al. (zhang 2009) [79], techniques of face recognition across pose are broadly classified into three categories, i.e., general algorithms, 2D techniques, and 3D approaches. By general algorithms, we mean those algorithms that did not contain specific tactics on handling pose variations. They were designed for general purpose of face recognition equally handling all image variations (e.g., illumination variations, expression variations, age variations, and pose variations, etc.). Generally, there are two trends in developing face recognition techniques, i.e. (1) improving the capability and universality of general face recognition algorithms so that image variation can be tolerated and (2) particularly designing mechanisms that can eliminate or at least compensate the difficulties brought by image variations (e.g., pose variations) according to its own characteristics, such as through 2D transformations or 3D reconstructions.

Recently, face recognition with assistance of 3D models is becoming one of the successful approaches when dealing with pose variations. The success of 3D model-based approaches in handling pose variations is due to the fact that human heads are 3D objects were fine structures and changes in viewpoints all take places in the 3D spaces. 3D reconstruction is an active research area in computer vision, which inversely estimates 3D shape information from 2D images. Generalised 3D reconstruction considers all of the shape modelling, the surface reflectivity descriptions and the estimation of environmental parameters (e.g., lighting conditions). The clues for reconstructing 3D objects in 2D images are usually image features (e.g., edges and corners) and image intensities.

Blanz and Vetter (Blanz 2003) [10] proposed a successful face recognition system using 3D morphable model based on image-based reconstruction and prior 3D knowledge of human faces. The morphable model was fitted with a single face image in an arbitrary condition by iteratively minimising pixel differences of image intensities and reconstructed virtual intensities using the set of parameters controlling the variations of shape, texture, illumination, pose, specularity, camera parameters, etc. Using stochastic Newton optimisation method, the process first makes use of several facial landmarks defined on both image and 3D model to find a rough alignment and then

relies more and more on the comparison of pixel intensities. The principal components of shape model and texture model were obtained in this process which was then used to reconstruct personalised 3D models and used for recognition using a modified angular (dot product) similarity measure based on linear discriminative analysis. In their experiments they show outstanding published results for images with pose variability from the PIE database.

In (Blanz 2005) [8],Blanz et al. propose to use the 3D Morphable Model in another way for non frontal face recognition in 2D still images: it serves as a preprocessing step by estimating the 3D shape of novel faces from the non frontal input images, and generates frontal views of the reconstructed faces at a standard illumination using 3D computer graphics. The transformed images are then fed into state of the art 2D face recognition systems that are training and optimized for frontal views. The 3D Morphable Model is used as a preprocessing tool for generating frontal views from non-frontal images. This method was shown to be extremely effective in the Face Recognition Vendor Test FRVT 2002, but still needs manually landmarks for the 2D images.

Jiang et al. [35, 32] used facial landmarks to efficiently reconstruct personalised 3D face models from a single frontal face image for recognition. Their method is based on the automatic detection of facial landmarks on the frontal views using Bayesian shape localisation. A set of 100 3D face scans was used as prior knowledge of human faces. Facial landmarks on both input images and 3D scans were used to find principal components of face shapes on the shape spaces spanned by the training 3D shapes. Personalised 3D face shapes were reconstructed and the facial textures were directly mapped onto the face shape to synthesize virtual views in novel conditions. Because the facial landmarks all have semantic meanings, this method is also capable to synthesise virtual views with different expressions through changing locations of the facial landmarks on the reconstructed 3D models. On CMU-PIE database, the method was shown to improve both PCA and LDA recognition algorithms, especially for LDA in half-profile views.

Castillo and Jacobs [14, 15] proposed to use the cost of stereo matching of gallery face image and probe face image to recognise faces. The stereo matching algorithm used in this method defined four planes which were left and right occluded planes and left and right matched planes. It involved fourteen transitions such as state preserv-

ing transitions and between state transitions. The cost of the stereo matching is defined as the sum of all the matching rows of the first image (say left) to the second (right) image. Exhaustively performing stereo matching using every view in the gallery to the probe image, the match was selected when the cost of stereo matching was the smallest. Tested on PIE database with 13 poses per face of 68 faces, this method achieved 73.5% recognition accuracy using any one pose as gallery and the remaining 12 poses as probe.

In [5], Ashraf et al. presented a novel strategy refereed to as stack-flow for aligning a stack of images at the patch-level. This approach is able to learn correspondences between gallery and probe viewpoints in a superior manner as compared to conventional image-to-image alignment techniques. Based on this learnt correspondence they have proposed an extension to Kanade and Yamadas viewpoint invariant face recognition work [36] to model the discriminative power of corresponding gallery and probe patches. The experiment on FERET database (http://itl.nist.gov/iad/humanid/feret/) results have also demonstrated the benefit of composing incremental warps (composite warp) to handle large view-point variations.

In this chapter, we will validate how to use 3D Morphable Models to solve the problem for face recognition under uncontrolled imaging conditions with pose variation. First, we reconstructed the 3D shape by automatic fitting the 3D Morphable Model to an image which gives a full solution of the 3D vision problem. For face recognition, different algorithm and information from reconstructed 3D face can be exploited, including:

1. The ICP distance of the 3D surfaces based measure;

2. Shape and texture coefficients-based comparison;

3. Viewpoint normalization approach.

The first approach is based on the geometry distance between the reconstructed 3D model from different pose, in that case only shape information are used. The second and the third methods are proposed by Blanz et al. in [10] and [8] separately.

5.3 Background of Automatic 2D Face Reconstruction Across Pose

In this section, we discuss face recognition system across large changes in viewpoint and the experimental protocol. For enrolment, the face recognition system is provided with one gallery image of each individual person, and in testing, each trial is performed with a single probe image. In an identification task, the system reports the identity of the probe person.

The Morphable Model of 3D faces is a vector space of 3D shapes and textures spanned by a set of examples. Derived from 100 textured Cyberware (TM) laser scans, the Morphable Model captures the variations and the common properties found within this set. The model parameters (coefficients) α_i and β_i are used to represent a face in 3D. By fitting the model to the facial image we can recover the parameters of the specific person, this is the produces of the face reconstruction. The detail and the fitting are described in Chapter 4 and summarized in Figure 5.1.

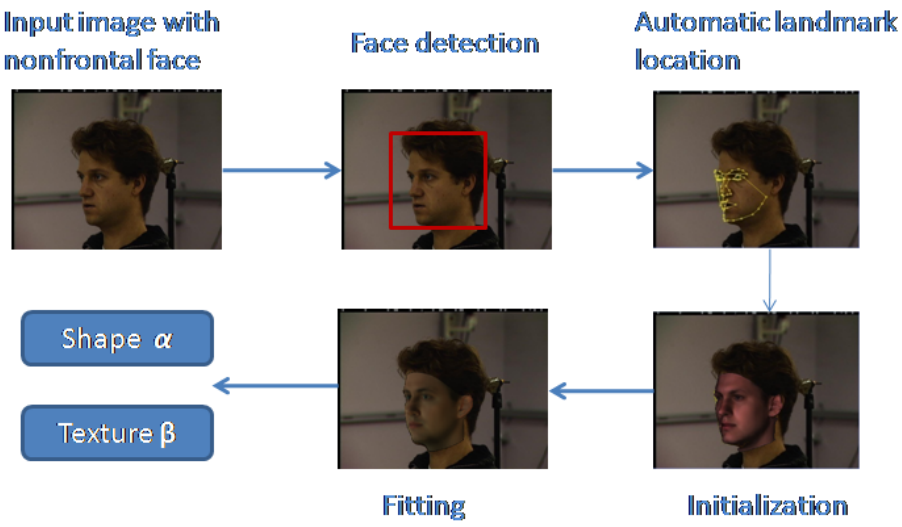

Figure 5.1: Face reconstruction procedure. For each input image a shape $\boldsymbol{\alpha}$ and a texture $\boldsymbol{\beta}$ parameter vector can be extracted separately.

As shown in Figure 5.1, for face reconstruction the following steps are needed:

1. Face detection: in here, we assume we have the face detection as described in [80], which can work on multi-view face image and give the coarse face pose.

2. Automated landmarks detector across view: the 3D-ASM landmark detector introduced in chapter 4 are used in this step. 58 landmarks are detected to initialize the morphable model to the input image automatically.

3. Model fitting: This is actually the 3D face reconstruction by analysis and synthesis loop as described in chapter 5. The shape $\boldsymbol{\alpha}$ and the texture $\boldsymbol{\beta}$ parameters are recovered during the processing.

All the following algorithms are based on the 3D Morphable Model reconstruction, the reconstructed result can be exploited in different ways to solve the face recognition across pose problem as described in Sections 5.4, 5.5 and 5.6.

5.3.1 Experimental Protocol

In our experiment, all the algorithms are evaluated on a subset (images from the cameras (02, 37, 05, 29, 11 and 14)) of the hPIE database and our Morphable model comes from USF human-ID database [9]. The PIE database contains 41, 368 images of 68 people, each person is under 13 different poses, 43 different illumination conditions, and with 4 different expressions. In our experiment, we use a subset of the PIE database, containing images of 68 individuals persons taken under 7 views with the same illumination and expression. Only one frontal image (from camera 27) for each person is stored in the gallery and some nonfrontal images (images from the cameras (02, 37, 05, 29, 11 and 14) see Figure 5.2 are used for testing. The pose variation is from 75 degrees to 75 degrees. The image size is 640x480 and the average distance better the two eyes is 50 pixels. It should be noted that, since the OpenCV face detector only works

Figure 5.2: Images taken from all cameras of the CMU PIE database for subject 04006. The nine cameras in the horizontal sweep are each separated by about 22.5.

for near frontal images (images from cameras (05 27 29)). During the experiment, we

have to adapt the 3D-ASM landmark detector to those images where the OpenCV face detector doesn't work. So we suppose that we have the approximate position and pose of every input image. We are using the 3 points (eyes centres and nose tip) to initialize our landmark detector (3D-ASM) and we suppose that we know the angle of head pose. The landmarks of the 3 points are given in (http://www.ralphgross.com/) .

5.4 ICP Distance of 3D Surfaces Based Measure

3D face recognition has the potential to achieve better accuracy than its 2D counterpart by measuring geometry of rigid features on the face. This avoids such pitfalls of 2D face recognition algorithms as change in lighting, different facial expressions, make-up and head orientation. One idea is to use the 3D reconstructed surface to do face recognition, this changes the 2D face recognition problem to a 3D face recognition problem.

5.4.1 The ICP Distance Measure

The Iterative Closest Point (ICP) algorithm, first proposed by Besl and McKay [7], is most widely used for 3D registration. This minimizes the cost as a function of the Euclidean distance between all the registered points in the two scans. The closed form solution for the rotation and translation pertaining to the local minima can be obtained using unit quaternions [31]. A brief description of the ICP is presented below.

Let $P = \{p_j\}$ be the set of scan points taken of the object to be registered to the reference scan points, $R = \{r_i\}$. The aim is then to find the rotation R and translation t which minimizes the following cost function:

$$E(R,t) = \sum_{i=1}^{|R|} \sum_{j=1}^{|P|} w_{i,j} ||r_i - (Rp_j + t)||^2 \tag{5.1}$$

where $w_{i,j}$ is 1 if r_i is matched to p_j, is 0 if r_i is not matched to p_j. Initially, the scan P is transformed using estimates of R and t. Then, for each point in P, the closest point using Euclidean distance, in the reference scan R is determined. The point correspondences are then used to compute the least squares solution for R and t that minimizes Eq. (5.1). These refined estimates of R and t are then used to transform P

and the process is repeated until the solutions do not change enough in iterations. The residual $E(R, t)$ could be used as our distance measure.

5.4.2 Experimental Results of the ICP Distance Measure on a Subset of PIE Database

We use the frontal images from the c27 for enrolment (gallery) (68 images), and the other (408) images for test. The experimental result on PIE database could be found in Table 5.1.

Table 5.1: Face identification rate using ICP distance measure approach on PIE database.

	c02	c37	c05	c29	c11	c14
Face identification rate	0	1.4	3	3	1.4	0

From the Table 5.1, we can see it is the ICP distance almost doesn't work for the face recognition. The reasons for the error are manifold. To estimate the 3D geometry shape information from the single 2D image is an ill-posed problem, the estimated 3D shape is affected by the pose and image quantity when large pose variation is presented.

5.5 Face Identification with 3D Shape and Texture Parameters

As described in (Blanz 2003) [10], after estimating the shape $\boldsymbol{\alpha}$ and texture$\boldsymbol{\beta}$ parameters from images by the fitting algorithm. Face recognition can be based on model coefficients, which represent intrinsic shape and texture of faces, and are independent of the imaging conditions. For identification, all enrolment images (c27) are analysed by the fitting algorithm, and the shape and texture coefficients are stored. Given a probe image (from 02, 37, 05, 29, 11 or 14), the fitting algorithm computes coefficients which are then compared with all gallery data in order to find the nearest neighbour. There are a number of options for distance measures between 3D faces to rely on for face recognition [10]. In our implementation we choose the cosine of the angle between

two vectors as our similarity measure:

$$d_A = \frac{\langle \boldsymbol{c}_1, \boldsymbol{c}_2 \rangle}{\|\boldsymbol{c}_1\| \cdot \|\boldsymbol{c}_2\|} \tag{5.2}$$

where the $\boldsymbol{c}_1$and $\boldsymbol{c}_2$, are the combination of shape and texture parameters.

5.5.1 Experimental Results of Face Identification with 3D Shape and Texture Parameters on subset of PIE database

In our experiment, we have tested the different strategy for face recognition, only using the shape parameters ($\boldsymbol{c} = \boldsymbol{\alpha}$),only using the texture parameters ($\boldsymbol{c} = \boldsymbol{\beta}$), and combining shape and texture parameters ($\boldsymbol{c} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$). The first 50 shape (Energy: 84%) and texture (Energy: 78%) PCA parameters are used. The results are list in Table 5.2 .

Table 5.2: Face identification rate using parameters-based approach on PIE database.

	c02	c37	c05	c29	c11	c14
shape parameters-based	0	1.4	6	6	1.4	0
texture parameters-based	15	25	34	34	25	15
shape texture parameters-based	1.4	6	13	13	6	1.4

From the table we can see that, during the 2D face recognition, texture parameters are much more important than shape parameters. Extracting 3D shape parameters from nonfrontal face image from different pose is a ill-posed problem. We can extracting it by using the statistical priori information from 3D scan but it is not enough for face recognition.

It should be noted that, our results are far away from the state of the art result by Blanz et al. [10]. The reason could be from two aspect: first in our face reconstruction algorithm we have used the automatic landmark while Blanz et al. use manually landmarks, second, we are using the USF human ID model which is built from 100 scans. While in [10], the 3DMM is constructed from 200 scans. Its ability to represent a face is dependent on the training set having contained similar faces.

5.6 Viewpoint Normalization Approach

Most face recognition algorithms are commercially available today are restricted to images with close-to-frontal views only, they are computationally efficient. In a combined approach, we have used the Morphable Model as a preprocessing tool or generating frontal views from non-frontal images which are then input to the image-based recognition systems. For generating frontal views, the Morphable Model is used to estimate 3D shape and texture of the face, and this face is rendered in a frontal pose and at a standard size and illumination. The flow chart of the algorithm is shown in Figure 5.3.

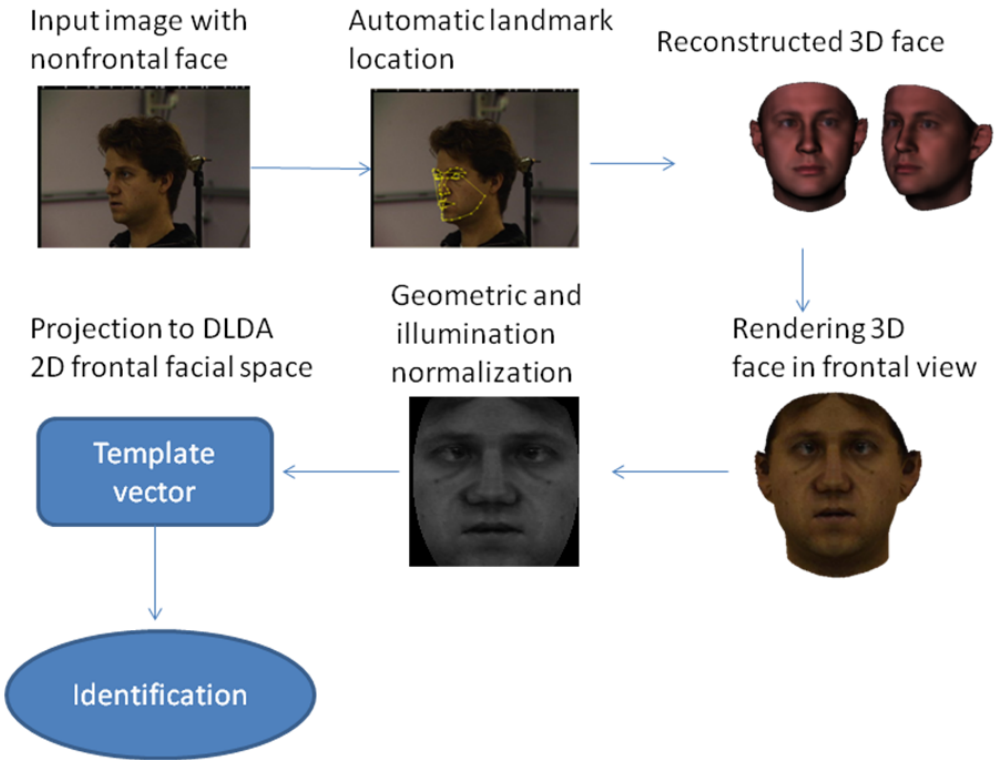

Figure 5.3: Flow charts of face identification across pose by viewpoint normalization approach.

5.6.1 Texture Extracted from Images or Synthesized Texture from 3D-MM

There are two different ways to map the texture. The first way is directly taking the texture from the input image, we name it as **image-based** texture. The second way is to synthesize the texture from the texture statistical model. We name it

model-based texture. Typical examples of those two strategies are shown in Figure 5.4.

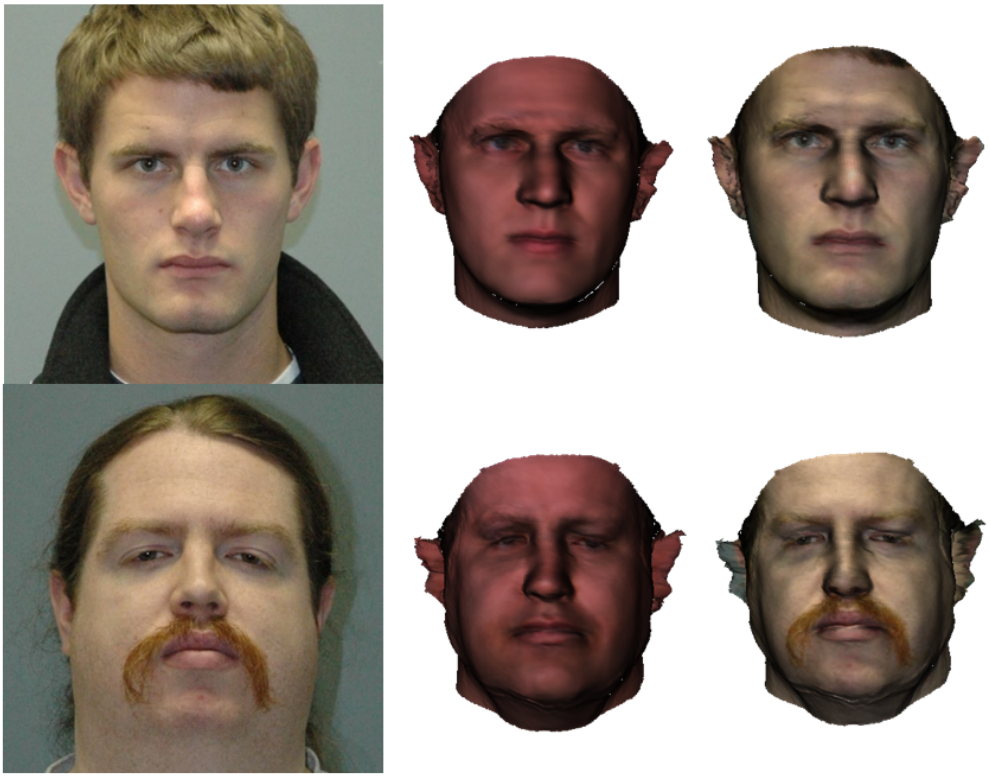

Figure 5.4: The different ways to map the texture, in the left column we give the original input image, those image are taken from the MBGCv1 database. In middle column we show the texture from the 3DMM with synthesis texture, in right column we show the mapping texture with the pixel from the input images.

For the image-based texture, the 2D image could be considered as the 3D projection on the 2D image plane, so we compute the pixel value from the color of 3D vertex in the 3D model. The 3D face reconstruction and image rendering procedure could be consider as a wrap from the input image to the target plan, as shown in Figure 5.3. The advantage of this method is that the warping processing keeps most of the information and the facial details from the input image. We can take a look at the second example in the Figure 5.4: the image-based textures recover more details of the face. The disadvantage of this algorithm is also significant, since the human face is a 3D object; one single 2D image can only contain partial data of the whole 3D face. That is to say not all of the 3D vertex on the 3D model can find the correspondence projection on the 2D image plan because of self-occlusion.

The synthesis of a virtual image is accomplished by sampling texture from

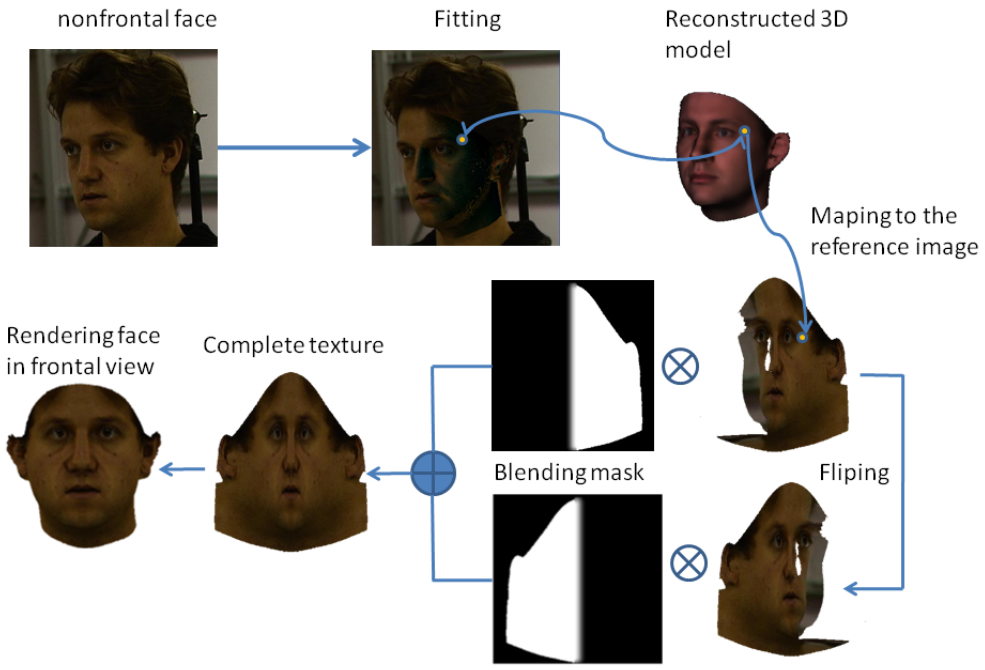

Figure 5.5: Block diagram for pose normalization using 3DMM fitting and facial symmetry. Texture from the input image is extracted by projecting the vertices of the 3DMM on the image plan. If these vertices are visible, the RGB values will be taken and copied to the texture map which is presented in cylindrical coordinates. The final texture is completed using symmetry property of faces.

the original one. The problem arises when, due to self-occlusion, some face regions become not visible, i.e., texture is not available, and hence the corresponding regions in the pose normalized image do not represent subjects appearance correctly. In order to overcome this drawback, we take advantage of the vertical symmetry of the face. For a horizontal rotation in depth of the head and once the mesh has been fitted, the parameter controlling the azimuth angle indicates whether the face is showing mostly its right or its left side. Whenever a frontal face is synthesized from a nonfrontal view, we warp the original image and its mirror version onto the cylindrical coordinate system and then blend the two virtual images, using simple masks that weigh the two sides of the face appropriately (according to the current rotation left or right of the head), as it can be seen in Figure 5.5.

In the other hand the model-based texture is generated from the statistical model of the texture. Since we have extracted the texture parameter during the face reconstruction phase, by texture PCA reconstruction. The model-based texture can be easily obtained. The advantage of this kind of texture is that it only take the importance information of the human face, this texture is only contain the color information with correspondence to the person, exclude the disturbance of the illumination. But the ability of the presentation of a new face is limited by the statistical information during the training phase. For example, if there are beard in the input image, we cant synthesized it. Because in the Human-ID database which is used for constructed our 3D Morphable model doesn't contain subjects with beards. In other word the quality of the model-based texture depends on the learning set. The experiment of evaluating this two kinds of texture mapping strategy will be given in the experimental section.

5.6.2 Experimental Result

In our experiment, we compared the model based texture and the image-base texture for recognition. For this two kind of approaches, the fitting process are the same, that is to say, the shape parameters are also the same. Frontal images (c27) are chosen for enrolment and non-frontal images (c02, c37, c05, c29, c11, c14) are chosen for the test. For all the images we have used the same preprocessing step: first we fit the 3DMM to those images and extracted the shape and texture parameters. Then for each image, a synthesized image is rendered in a frontal pose and at a standard size and

illumination. So we can change the large pose problem to the normal 2D face recognition problem. For comparison, the original images are also used for face recognition, in that case the images are geometrically normalized only by the eyes and mouth position. Figure 5.6, illustrates some examples of those two different ways to generate the images before passing them to the 2D face recognition system.

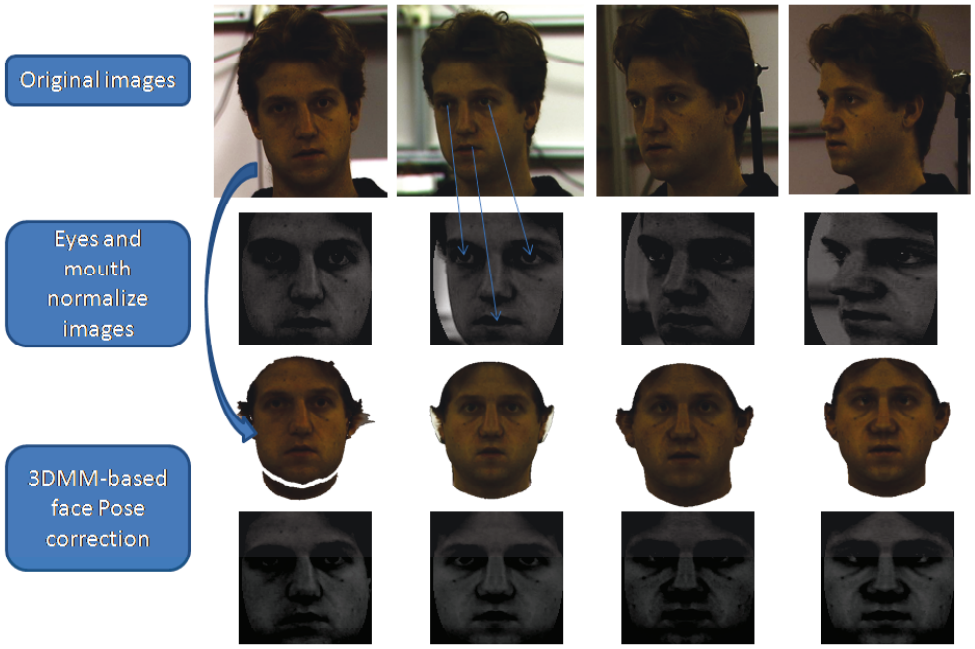

Figure 5.6: Eyes and mouth based 2D pose correction Vs 3DMM-based Face Pose correction. Images are taken from PIE database.

It can be seen that pose correction was achieved and was particularly effective in correcting pose rotations. However, texture artefacts were sometimes introduced, particularly around the areas of the nose which were hidden from view in the original images. These artefacts were most likely caused because we use the symmetry property to complete the facial texture, which makes the joint part of the face (centreline) unnatural.

For 2D face recognition system we use the open source sudfrog system, more detail about this software can be found in (svnext.it-sudparis.eu). We use Gabor filters with several resolutions and orientations (5x8) convoluted with the normalized images and only magnitude value are used in our experiment. The vector is projected in the

DLDA space to reduce the dimension of the space into 120. So the output of this step will be a vector with length 120 to represent a face image. The learning data is from FRGCv2 training set, with 120 persons with approximately 10 images for each person. We used the rank-1 recognition rate with the smallest cosine distance (1 -cosine distance) measure for this performance evaluation.

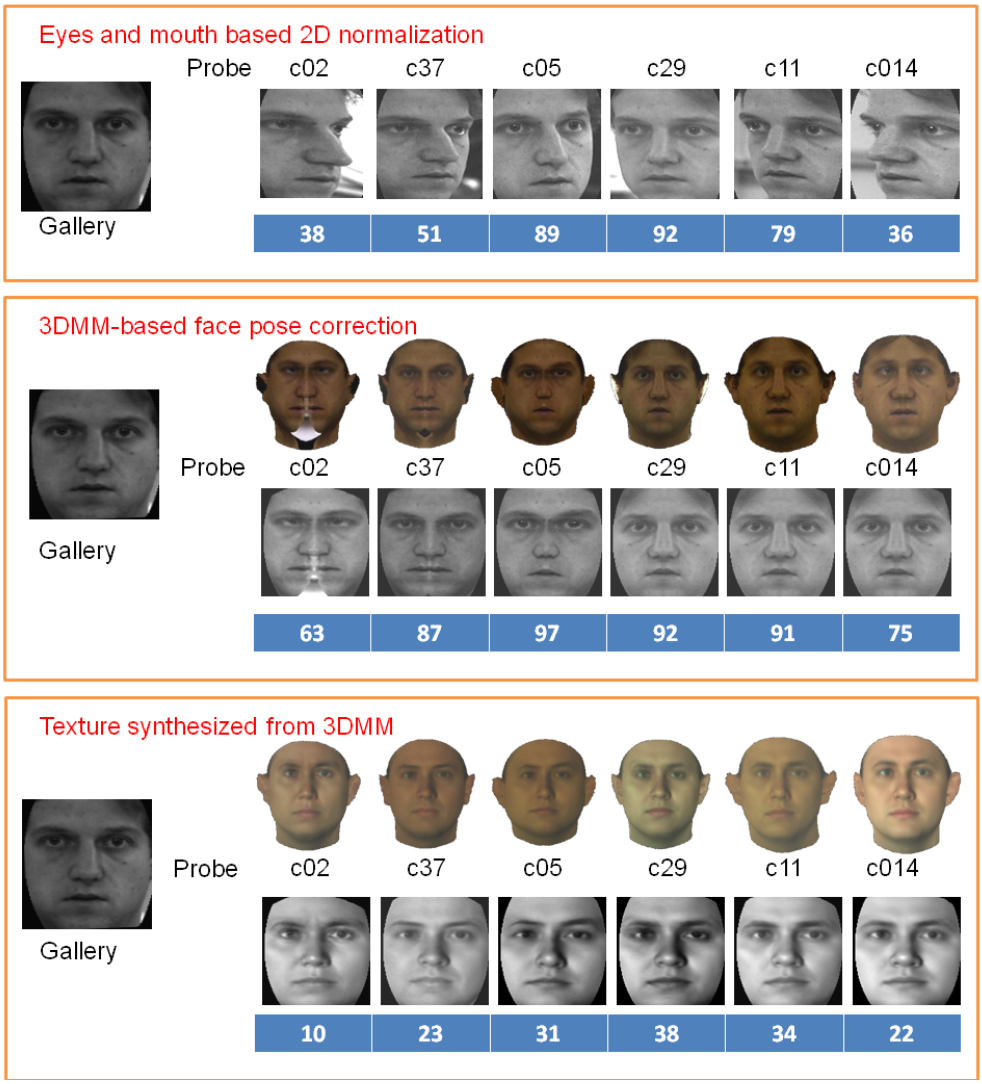

Figure 5.7: Face recognition performance comparison on PIE database. The first row is the face identification rates using eyes and mouth based 2D normalization. While the second row is the face identification rates using 3DMM-based face pose correction. The third row we are using the same reconstruction step as the second row, the only difference is that texture for face reconstruction is synthesized from 3DMM instead of taken from input image.

As shown in Figure 5.7, the first row is the face identification rates using eyes

and mouth based 2D normalization. While the second row is the face identification rates using 3DMM-based face pose correction. The third row we are using the same reconstruction step as the second row, the only difference is that texture for face reconstruction is synthesized from 3DMM instead taken from input image. Our 3DMM-based face pose correction methods performs robustly well across pose changes against the eyes and mouth coordinate-based normalized method. In that case, the face recognition results could be considered as some indirect way to evaluation the 3D face reconstruction algorithm.

For the reconstructed 2D frontal images, using the texture taken from image has much better face recognition result than synthesized texture from 3DMM. Taken texture from the input images keeps most information and the facial details from the input image, although some time it bring artificiality to the 2D frontal images. This kind of artificiality may not break the facial structure for face recognition even it looks unnatural.

Since the viewpoint normalization approach gives our best face identification performance, in Figure 5.8, we compared the results with some published works. We have better face recognition performance than the [32] which is based on 3DMM and LDA, but it is still worse than the method based on stereo-matching by [14].

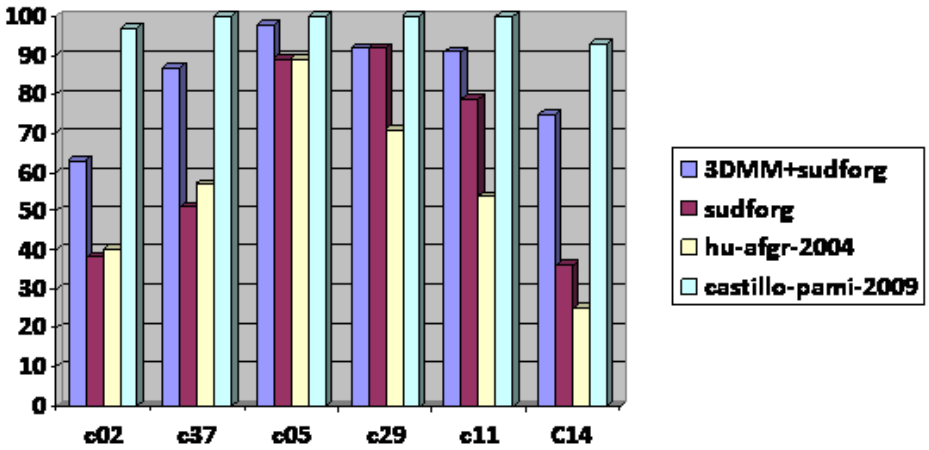

Figure 5.8: Recognition accuracy comparison. In this figure, we compared our face recognition result with some published works [32, 14].

Using the 3DMM to correct the face pose in the image can significantly im-

proved the performance of face recognition across pose. But the pose is still a big challenge for the face recognition problem, as we see in the Figure 5.8, the bigger face pose is presented in the examples the worse face recognition result we obtain.

5.7 Conclusions

In this chapter, we have studied how to use 3D Morphable Model was used as a tool for correcting the pose of 2D images prior to presenting them to a face recognition algorithm. Experiments on the PIE database showed that the approaches proposed for pose correction improved the performance original 2D face recognition system when non frontal images were used on a system trained with near frontal images only. During the experiment, we found that the facial texture is more important for the face recognition. And to achieve automatic frontal-profile face recognition is still a challenge work. Although the experiment results are not out performance of the state of the art algorithm, but we have demonstrated in this chapter that we have studied in detail a version of an automated 3D Morphable Model based face recognition algorithm and discussed the issues related to its success and failure.

Chapter 6

Conclusions and Future Work

In the beginning of our work, we wanted to find a automatic 2D facial landmark location algorithm (Combined Active Shape Model) for 2D the face recognition in near frontal images where the landmarks can not by manually located. It can be used to improve the normalization step of global 2D frontal face recognition systems. The Combined Active Shape Model can also be used for initialization for 3D face reconstruction from near frontal images. But as we are also interested in 3D recongnition from nonfrontal facial images, we developed the 3D Acitve Shape Model. And we evaluate our 3D face reconstruction system in two different ways: (1) quantitatively evaluation on biometric databases, and (2) 2D face recognition.

This chapter concludes the work of the 3D statistical face reconstruction and its application for face precessing, and presents future direction. The first section gives significant achievements and conclusion throughout the thesis. Followed by some directions for future research aimed at solving the remaining problems.

6.1 Achievements and Conclusion

In this thesis, the two kind of automatic facial landmark location algorithms: Combined Active Shape Model and 3D Acitve Shape Model are proposed, developed, tested and compared. Combined Active Shape Model and 3D Acitve Shape Model algorithms are both based on the original Active Shape Model and exploit scale-invariant feature transform descriptor as local texture feature which increased the robustness to the scale and rotation variability and challenging lighting conditions. While the Combined Active Shape Model uses 2D images as training data, we split the facial landmarks in facial internal region and facial contour landmarks to deal with small pose variation. The experimental results show that it has good performance on near frontal view face images in the degraded condition such as video surveillance. In another hand, the 3D Active Shape Model extents the original Active Shape Model by using 3D scan data. By using the view-based local texture model, it can deal with large pose variations. Because we use a 3D Morpable Model to generate the training data, few manual operations are need for the training and it also gives better initialization the 3D Morphable Model reconstruction.

By exploiting the 3D Active Shape Model, we present a fully automated al-

gorithm for reconstructing 3D models of a face from single photograph with nonfrontal faces. The algorithm is based on a combination of 3D-ASM and 3D Morphable Model face reconstruction.

Another contribution of the work is to use the biometric data to quantitatively evaluate the 3D face reconstruction precision. We evaluate the automated 3D face reconstruction results quantitatively on the IV^2 Biometric Database. Different automatic landmark location algorithms for initialization the system are compared. The influence of the precision of landmark location due to the 3D face reconstruction is evaluated. The results show that the 3D-ASM provides excellent initialization for the 3D face reconstruction with nonfrontal faces. It seems to be even better compared to the manual annotation for nonfrontal images.

Finally, we studied how to use the 3D Morphable Model as a tool for correcting the pose of 2D images prior to presenting them to a face recognition algorithm. Experiments on the PIE database showed that the approaches proposed for pose correction improved the performance of global 2D face recognition algorithm when nonfrontal images were used on a system trained with near frontal images only. During the experiment, we found that facial texture is more important for the face recognition. And to achieve automatic frontal-profile face recognition is still a challenge work.

6.2 Future work

Future work could be extended in many aspects, such as face pre-processing, investigation on the training dataset, exploration of more types of features, using more input images, experiments on other databases.

In Chapter 2, for the 2D face verification we only use the global based method (Gabor feature and DLDA classifier). Since the landmarks around the facial region are located by Combined Active Shape Model detector, local based approach could be adopted or fusion to the global based method, such as Elastic Bunch Graph Matching(EBGM) into the algorithm to improve its performance at the feature selection stage.

In Chapter 3, the 3D Active Shape Model can't deal with the expression because the expression variability is not present in the 3D Morphable Model. The expression can be handled by the same framework by increasing expression variability in

3DMM.

In Chapter 4, the 3D Morphable Model reconstruction algorithm use only pixels value in the analysis-by-synthesis loops, more feature cloud be also used: edge, contour et al.. And also using more input data (stereo images and video sequence which contain in IV^2 database) for the 3D face reconstruction can also be tested.

Bibliography

[1] *Face Processing: Advanced Modelling and Methods*. Academic Press, 2006.

[2] Brian Amberg, Sami Romdhani, Andrew Blake, Thomas Vetter, and Andrew Fitzgibbon. Reconstructing high quality face surfaces using model based stereo. In *Proc. ICCV (11)*, 2007.

[3] G. Antonini, V. Popovici, and J. Thiran. Independent Component Analysis and Support Vector Machine for Face Feature Extraction. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, Guildford, UK*, volume 2688 of *Lecture Notes in Computer Science*, pages 111–118, Berlin, 2003. IEEE.

[4] Stefano Arca, Paola Campadelli, and Raffaella Lanzarotti. A face recognition system based on automatically determined facial fiducial points. *Pattern Recogn.*, 39(3):432–443, 2006.

[5] A.B. Ashraf, S. Lucey, and Tsuhan Chen. Learning patch correspondences for improved viewpoint invariant face recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008.

[6] Joseph J. Atick, Paul A. Griffin, and A. Norman Redlich. Statistical approach to shape from shading: Reconstruction of 3d face surfaces from single 2d images. *Neural Computation*, 8:1321–1340, 1997.

[7] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):239–256, 1992.

[8] Volker Blanz, Patrick Grother, P. Jonathon Phillips, and Thomas Vetter. Face recognition based on frontal views generated from nonfrontal images. In *Proc.*

IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005, volume 2, pages 454–461, 2005.

[9] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In ACM SIGGRAPH, editor, *Proceedings of SIGGRAPH 99*, pages 187–194, August 1999.

[10] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[11] G. Bradski, A. Kaehler, and V. Pisarevski. Learning-based computer vision with intel's open source computer vision library. *Intel Technology Journal*, 9(2):119–130, May 2005.

[12] P. Breuer, K. I. Kim, W. Kienzle, B. Scholkopf, and V. Blanz. Automatic 3d face reconstruction from single images or video. In *FG*, 2008.

[13] Bernard F. Buxton and M. Benjamin Dias. Implicit, view invariant, linear flexible shape modelling. *Pattern Recogn. Lett.*, 26:433–447, March 2005.

[14] Carlos D Castillo and David W Jacobs. Using stereo matching for 2d face recognition across pose. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, june 2007.

[15] Carlos D Castillo and David W Jacobs. Using stereo matching with general epipolar geometry for 2D face recognition across pose. *IEEE transactions on pattern analysis and machine intelligence*, 31(12):2298–304, December 2009.

[16] Angela Caunce, Chris Taylor, and Tim Cootes. Improved 3d model search for facial feature location and pose estimation in 2d images. In *Proceedings of the British Machine Vision Conference*, pages 81.1–81.10. BMVA Press, 2010. doi:10.5244/C.24.81.

[17] Angela Caunce, Chris Taylor, and Tim Cootes. Improved 3d model search for facial feature location and pose estimation in 2d images. In *Proceedings of the British Machine Vision Conference*, pages 81.1–81.10. BMVA Press, 2010. doi:10.5244/C.24.81.

[18] Chun Chen, Ming Zhao, Ming Stan Z. Li, and Jiajun Bu. Parameter optimization for active shape models. In *Asian Conference on Computer Vision*, 2004.

[19] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. In *Computer Vision and Image Understanding*. Elsevier Science Inc., 1995.

[20] T. F. Cootes and C.J. Taylor. Statistical models of appearance for computer vision, 2004.

[21] Chunhua Du, Qiang Wu, Jie Yang, and Zheng Wu. Svm based asm for facial landmarks location. In *Proc. 8th IEEE International Conference on Computer and Information Technology*, pages 321–326, 8–11 July 2008.

[22] B.A. Efraty, E. Ismailov, I.A. Kakadiaris, and S. Shah. Towards 3d-aided profile-based face recognition. In *in Proc. BTAS (3)*, 2009.

[23] Lixin Fan and Kah-Kay Sung. Model-based varying pose face detection and facial feature registration in colour images. *Pattern Recogn. Lett.*, 24:237–249, January 2003.

[24] I. R. Fasel, M. S. Bartlett, and J. R. Movellan. A comparison of gabor filter methods for automatic detection of facial landmarks. In *Proc. Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 242–246, 21–21 May 2002.

[25] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:643–660, 2001.

[26] N. Gourier, D. Hall, and J. L. Crowley. Facial features detection robust to pose, illumination and identity. In *Proc. IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 617–622, 2004.

[27] Ralph Gross. http://ralphgross.com/facelabels.

[28] Ralph Gross, Iain Matthews, and Simon Baker. Appearance-based face recognition and light-fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:449–465, 2002.

[29] Lie Gu and T. Kanade. 3d alignment of face in a single image. In *CVPR*, volume 1, pages 1305–1312, 17–22 June 2006.

[30] M. Hamouz, J. Kittler, J. K. Kamarainen, P. Paalanen, H. Kalviainen, and J. Matas. Feature-based affine-invariant localization of faces. 27(9):1490–1495, Sept. 2005.

[31] Berthold K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America. A*, 4(4):629–642, April 1987.

[32] Yuxiao Hu, Dalong Jiang, Shuicheng Yan, Lei Zhang, and Hongjiang Zhang. Automatic 3D reconstruction for face recognition. In *Proc. Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 843–848, 2004.

[33] Yuxiao Hu, Dalong Jiang, Shuicheng Yan, Lei Zhang, and Hongjiang zhang. Automatic 3d reconstruction for face recognition. In *FG*, pages 843–848, 17–19 May 2004.

[34] Oliver Jesorsky, Klaus J. Kirchberg, and Robert Frischholz. Robust face detection using the hausdorff distance. pages 90–95, 2001.

[35] D.L. Jiang, Y.X. Hu, S.C. Yan, L. Zhang, H.J. Zhang, and W. Gao. Efficient 3d reconstruction for face recognition. 38(6):787–798, June 2005.

[36] Takeo Kanade and Akihiko Yamada. Multi-subregion based probabilistic approach toward pose-invariant face recognition. In *Proceedings of 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, pages 954 – 959, July 2003.

[37] A. Koschan, S. K. Kang, J.K. Paik, M. Abidi, and M. Abidi. Video object tracking based on extended active shape models with color information. In *1st European Conf. on Color in Graphics, Imaging, and Vision*, pages 126–131, 2002.

[38] Vuong Le, Yuxiao Hu, and Thomas S. Huang. A quantitative evaluation for 3d face reconstruction algorithms. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 0:1269–1272, 2009.

[39] Jinho Lee, Hanspeter Pfister, Baback Moghaddam, and Raghu Machiraju. Estimation of 3d faces and illumination from single photographs using a bilinear illumination model. In *Rendering Techniques*, pages 73–82, 2005.

[40] Martin D. Levine and Yingfeng (Chris) Yu. State-of-the-art of 3d facial reconstruction methods for face recognition based on a single 2d training image per person. *Pattern Recogn. Lett.*, 30(10):908–913, 2009.

[41] Yongmin Li, Shaogang Gong, and Heather Liddell. Modelling faces dynamically across views and over time. In *Proceedings. Eighth IEEE International Conference on Computer Vision*, pages 554–559, 2001.

[42] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[43] Iain Matthews, Jing Xiao, and Simon Baker. 2d vs. 3d deformable face models: Representational power, construction, and real-time fitting. *nternational Journal of Computer Vision*, 75:93–113, October 2007.

[44] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *ECCV*, 2008.

[45] D. Monzo, A. Albiol, and J. Sastre. Hog-ebgm vs. gabor-ebgm. In *Proc. 15th IEEE International Conference on Image Processing ICIP 2008*, pages 1636–1639, 12–15 Oct. 2008.

[46] National Institute of Standards and Technology (NIST). Multiple biometric grand challenge (mbgc), http://face.nist.gov/mbgc, 2008.

[47] S. Ordas, L. Boisrobert, M. Huguet, and A. F. Frangi. Active shape models with invariant optimal features (iof-asm) application to cardiac mri segmentation. In *Proc. Computers in Cardiology*, pages 633–636, 21–24 Sept. 2003.

[48] Aurelien Mayoue & Anouar Mohamed Mellakh & Dianle Zhou & Dijana Petrovska-Delacrétaz and Bernadette Dorizzi. Utilisation de séquences vidéo avec critres de qualité pour la reconnaissance faciale. In *TRAITEMENT ET ANALYSE DE L'INFORMATION : Méthodes et Applications*, 2009.

[49] Dijana Petrovska-Delacrétaz, Gérard Chollet, and Bernadette Dorizzi, editors. *Guide to Biometric Reference Systems and Performance Evaluation.* Springer-Verlag, 2009.

[50] Dijana Petrovska-Delacrétaz, Sylvie Lelandais, Joseph Colineau, Liming Chen, Bernadette Dorizzi, Emine Krichen, Mohamed Anouar-Mellakh, Anis Chaari, Souhila Guerfi, Mohsen Ardabilian, Johan D Hose, and Boulbaba Ben Amor. The IV2 Multimodal Biometric Database (Including Iris, 2D, 3D, Stereoscopic and Talking Face Data) and the IV2-2007 Evaluation Campaign. In IEEE, editor, *IEEE Second International Conference on Biometrics: Theory, Applications and Systems (BTAS 08)*, pages 1–7, September 2008.

[51] P. Jonathon Phillips, Patrick J. Flynn, Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:947–954, 2005.

[52] P. Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J. Rauss. The feret database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16(5):295 – 306, 1998.

[53] Bui Tuong Phong. Illumination for computer generated pictures. *Commun. ACM*, 18:311–317, June 1975.

[54] Sami Romdhani, Shaogang Gong, and Ahaogang Psarrou. A multi-view nonlinear active shape model using kernel pca. In *Proceedings of the British Machine Vision Conference*, pages –1–1. BMVA Press, 1999.

[55] Albert Ali Salah, Hatice inar, Lale Akarun, and Bulent Sankur. Robust facial landmarking for registration. *Annals of Telecommunications*, 62:1–2, 2006.

[56] Conrad Sanderson, Samy Bengio, and Yongsheng Gao. On transforming statistical models for non-frontal face verification. *Pattern Recognition*, 39(2):288 – 302, 2006. Part Special Issue: Complexity Reduction.

[57] Sudeep Sarkar. http://www.csee.usf.edu/ sarkar/.

[58] T. Shakunaga, K. Ogawa, and S. Oki. Integration of eigentemplate and structure matching for automatic facial feature detection. In *Proc. Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 94–99, 14–16 April 1998.

[59] Yan ShuiCheng and QianSheng Cheng. Multi-view face alignment using direct appearance models. In *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, FGR '02, pages 324–, Washington, DC, USA, 2002. IEEE Computer Society.

[60] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *FG*, 2002.

[61] L. Sirovich. Turbulence and the Dynamics of Coherent Structures, Part I: Coherent Structures. *Quarterly of Appl. Math.*, XLV:561–571, 1987.

[62] P. D. Sozou, T. F. Cootes, C. J. Taylor, and E. C. Di Mauro. Non-linear point distribution modelling using a multi-layer perceptron. In *6 th British Machine Vison Conference*, pages 107–116. BMVA Press, 1995.

[63] M. B. Stegmann, B. K. Ersbøll, and R. Larsen. FAME – a flexible appearance modelling environment. *IEEE Trans. on Medical Imaging*, 22(10):1319–1331, 2003.

[64] Yanchao Su, Haizhou Ai, and Shihong Lao. Multi-view face alignment using 3d shape model for view estimation. In *Proceedings of the Third International Conference on Advances in Biometrics*, ICB '09, pages 179–188, Berlin, Heidelberg, 2009. Springer-Verlag.

[65] Federico M. Sukno. *Invariance and Reliability in Statistical Shape Models.* PhD thesis, Universidad de Zaragoza and Universitat Politcnica de Cataluny, 2007.

[66] Xiaoyang Tan, Songcan Chen, and Zhi hua Zhou Fuyan Zhang. Face recognition from a single image per person: A survey. *Pattern Recognition*, 39:1725–1745, 2006.

[67] Yan Tong, Yang Wang, Zhiwei Zhu, and Qiang Ji. Robust facial feature tracking under varying face pose and facial expression. *Pattern Recogn.*, 40:3195–3208, November 2007.

[68] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, pages 586 –591, jun 1991.

[69] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[70] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *Proc. IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1692–1698, 10–12 Oct. 2005.

[71] Wei Wang, Shiguang Shan, Wen Gao, Bo Cao, and Baocai Yin. An improved active shape model for face alignment. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, pages 523–, Washington, DC, USA, 2002. IEEE Computer Society.

[72] W.N. Widanagamaachchi and A.T Dharmaratne. 3D Face Reconstruction from 2D Images. *2008 Digital Image Computing: Techniques and Applications*, pages 365–371, December 2008.

[73] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. In *Proc. International Conference on Image Processing*, volume 1, pages 129–132, 26–29 Oct. 1997.

[74] Jing Xiao, Simon Baker, Iain Matthews, and Takeo Kanade. Real-time combined 2d+3d active appearance models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 535 – 542, June 2004.

[75] S. Xin and H. Ai. Face alignment under various poses and expressions. In *Proc. 1st Int. Conf. on Affective Computing and Intelligent Interaction, Beijing, China*, page 40C47, 2005.

[76] Pengfei Xiong, Lei Huang, and Changping Liu. Initialization and pose alignment in active shape model. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3971 –3974, aug. 2010.

[77] Shuicheng Yan, Xinwen Hou, Stan Z. Li, Hongjiang Zhang, and Qiansheng Cheng. Face alignment using view-based direct appearance models. *Special Issue on Facial Image Processing, Analysis and Synthesis. International Journal of Imaging Systems and Technology. 2003*, 13:106–112, 2003.

[78] Y. Yan and J. Zhang. Rotation-invariant 3d reconstruction. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, volume 1, pages 156 –160 vol.1, oct 1998.

[79] Xiaozheng Zhang and Yongsheng Gao. Face recognition across pose: A review. *Pattern Recognition*, 42(11):2876 – 2896, 2009.

[80] Zhenqiu Zhang, Long Zhu, Stan Z. Li, and Hongjiang Zhang. Real-time multi-view face detection. In *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, FGR '02, pages 149–, Washington, DC, USA, 2002. IEEE Computer Society.

[81] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35:399–458, December 2003.

Appendix A

List of Publications

1. Chollet, G., Esposito, A., Gentes, A., Horain, P., Karam, W., Li, Z., Pelachaud, C., Perrot, P., Petrovska-Delacrétaz, D., Zhou, D., and Zouari, L. 2009. Multimodal Human Machine Interactions in Virtual and Augmented Reality. In Multimodal Signals: Cognitive and Algorithmic. Issues: COST Action 2102 and Eucognition international School Vietri Sul Mare, Italy, April 21-26, 2008 Revised Selected and invited Papers, A. Esposito, A. Hussain, M. Marinaro, and R. Martone, Eds. Lecture Notes In Artificial Intelligence, vol. 5398. Springer-Verlag, Berlin, Heidelberg, 1-23.

2. A. Mayoue, A. M. Mellakh, D. Zhou, D. Petrovska-Delacrétaz,and B. Dorizzi :Utilisation de squences vidéo avec critres de qualité pour la reconnaissance faciale. TRAITEMENT ET ANALYSE DE LINFORMATION : Methodes et Applications, 2009.

3. Dianle Zhou, Dijana Petrovska-Delacrétaz and Bernadette Dorizzi. Automatic landmark location with a combined active shape model. In Proceedings of the 3rd IEEE international Conference on Biometrics: theory, Applications and Systems (Washington, DC, USA, September 28 - 30, 2009). IEEE Press, Piscataway, NJ, 49-55.

4. Dianle Zhou, Dijana Petrovska-Delacrétaz and Bernadette Dorizzi. 3D Active Shape Model for Automatic Facial Landmark Location, Proc. International Conference on Pattern Recognition (ICPR 2010), 2010.