From Image Sequence to Frontal Image:
RECONSTRUCTION OF THE UNKNOWN FACE

Christiaan van Dam

A Forensic Case

FROM IMAGE SEQUENCE TO FRONTAL IMAGE: RECONSTRUCTION OF THE UNKNOWN FACE

A FORENSIC CASE

CHRISTIAAN VAN DAM

GRADUATION COMMITTEE:

CHAIRMAN:

PROF. DR. P.M.G. APERS UNIVERSITY OF TWENTE, EEMCS

SUPERVISOR:

PROF. DR. IR. R.N.J. VELDHUIS UNIVERSITY OF TWENTE, EEMCS

CO-SUPERVISOR:

DR. IR. L.J. SPREEUWERS UNIVERSITY OF TWENTE, EEMCS

MEMBERS:

PROF. DR. IR. C.H. SLUMP UNIVERSITY OF TWENTE, EEMCS

PROF. DR. D. MEUWLY UNIVERSITY OF TWENTE, EEMCS

PROF. DR. C. BUSCH NTNU, NORWAY

PROF. DR. IR. P.H.N. DE WITH EINDHOVEN UNIVERSITY OF TECHNOLOGY

DR. A. RUIFROK NETHERLANDS FORENSIC INSTITUTE

CTIT

CTIT PH.D. THESIS SERIES NO. 17-429

CENTRE FOR TELEMATICS AND INFORMATION TECHNOLOGY

P.O. BOX 217, 7500 AE ENSCHEDE, THE NETHERLANDS.

COPYRIGHT © 2017 CHRISTIAAN VAN DAM, ZOETERMEER, THE NETHERLANDS. ALL RIGHTS RESERVED. NO PART OF THIS BOOK MAY BE REPRODUCED OR TRANSMITTED, IN ANY FORM OR BY ANY MEANS, ELECTRONIC OR MECHANICAL, INCLUDING PHOTOCOPYING, MICROFILMING, AND RECORDING, OR BY ANY INFORMATION STORAGE OR RETRIEVAL SYSTEM, WITHOUT THE PRIOR WRITTEN PERMISSION OF THE AUTHOR.

ISBN: 978-90-365-4324-8

ISSN: 1381-3617 (CTIT PH.D. THESIS SERIES NO. 17-429)

DOI: 10.3990/1.9789036543248 ([HTTPS://DX.DOI.ORG/10.3990/1.9789036543248](https://dx.doi.org/10.3990/1.9789036543248))

FROM IMAGE SEQUENCE TO FRONTAL IMAGE: RECONSTRUCTION OF THE UNKNOWN FACE

A FORENSIC CASE

DISSERTATION

TO OBTAIN THE DEGREE OF DOCTOR
AT THE UNIVERSITY OF TWENTE,
ON THE AUTHORITY OF THE RECTOR MAGNIFICUS
PROF. DR. T.T.M. PALSTRA,
ON ACCOUNT OF THE DECISION OF THE GRADUATION COMMITTEE,
TO BE PUBLICLY DEFENDED
ON THURSDAY 30 MARCH, 2017 AT 14:45

BY

CHRISTIAAN VAN DAM
BORN ON MAY 9, 1985
IN GORINCHEM,
THE NETHERLANDS

THIS DISSERTATION HAS BEEN APPROVED BY:

PROF. DR. IR. R.N.J. VELDHUIS (SUPERVISOR)

DR. IR. L.J. SPREEUWERS (CO-SUPERVISOR)

DO ANY OF YOU NEED WISDOM?
ASK GOD FOR IT.
HE IS GENEROUS AND ENJOYS GIVING TO EVERYONE.
SO HE WILL GIVE YOU WISDOM.

JAMES 1:5 ~EASY-TO-READ VERSION

Contents

I	INTRODUCTION	I
1.1	Biometrics & Forensics	2
1.2	Forensic Use Case	3
1.3	Research Overview	5
1.4	3D Face Reconstruction Methods	6
1.5	Overview	7
2	RECONSTRUCTION METHODS	9
2.1	Introduction	9
2.2	Overview	10
2.3	Landmark Based	11
2.4	Structure from Motion	12
2.5	Shape from Shading	14
2.6	Shape from Silhouette	16
2.7	Morphable Models	17
2.8	Paper Closure	20
3	3D RECONSTRUCTION FROM VIDEO SEQUENCES: RANDOM POINT CLOUDS	21
3.1	Abstract	22
3.2	Introduction	22
3.3	Background	23
3.4	Auto-Calibration	24
3.5	Experiments	24
3.6	Conclusion	29
3.7	Paper Closure	29
4	3D RECONSTRUCTION FROM VIDEO SEQUENCES: A STYROFOAM HEAD	31
4.1	Abstract	32
4.2	Introduction	32
4.3	Background	33
4.4	Reconstruction Algorithm Overview	34
4.5	Experiments	36
4.6	Conclusion	41
4.7	Abstract	42
4.8	Introduction	42
4.9	Background	43
4.10	Reconstruction Algorithm	44

4.11	Experiments	45
4.12	Conclusion and Future Work	49
4.13	Paper Closure	50
5	3D RECONSTRUCTION FROM VIDEO SEQUENCES: REAL FACE DATA	53
5.1	Abstract	55
5.2	Introduction	55
5.3	Related Work	57
5.4	3D Face Reconstruction Method	59
5.5	Face Comparison	68
5.6	Conclusion	70
5.7	Acknowledgement	70
5.8	Paper Closure	70
6	3D RECONSTRUCTION FROM VIDEO SEQUENCES: DENSE 3D FACE MODELS	73
6.1	Abstract	73
6.2	Introduction & Background	74
6.3	Method	75
6.4	Experiments	79
6.5	Conclusion	86
6.6	Acknowledgement	87
6.7	Paper Closure	87
7	CONCLUSIONS	89
7.1	Final conclusion	89
7.2	Final Conclusion	91
7.3	Recommendations	92
A	APPENDIX A	93
A.1	Landmarks	93
B	APPENDIX B	95
B.1	Projection Models	95
	REFERENCES	104
	SUMMARY	105
	SAMENVATTING	107
	ACKNOWLEDGMENTS	109

1

Introduction

CHRIS VAN DAM

Faces, we see them every day. People recognize faces from a distance without any problem. Have you ever wondered why we are doing such a great job at recognizing faces? Or do we fool ourselves by thinking that we can recognize faces well? We consider it an easy task to recognize our relatives and friends, even when we only view them in low resolution images. But when we see our relatives and friends at an unexpected location, it may take a while before we recognize them. And how many times do we even fail to recognize them? So, the surroundings and context are definitely important for people in recognizing other people's faces. Often when we think we recognize someone, we rather recognize their clothing or haircuts, instead of their face. Our brains seem to build some model of a person, especially a person's face, based on all encounters with this person¹⁵. A model that can be updated, but which can also give a bias towards the recognition in certain surroundings or situations. One way to test our abilities to recognize faces is to perform an experiment with unfamiliar faces. How well would we recognize people that we have only seen in an image or a video? For many people it is a difficult task to recognize an unfamiliar face independently from its surroundings. And how to deal with the possible bias towards people we have encountered that we have built up over the years? In order to avoid such a bias, we would search for an automated approach to compare faces. As a result of years of research and development automated face recognition systems perform very well on frontal facial images. State-of-the-art face recognition systems make fewer than thirty errors on every thousand images⁷⁸. However, the recognition performance of face recognition systems degrades for facial images under pose⁵⁶. And how to combine the result of multiple recognitions for image sequences? Especially uncontrolled situations with faces under pose are difficult to handle for automated face comparison systems. To handle multiple images at once, 3D based reconstruction methods for pose and illumination compensation are needed. This

compensation could be acquired by creating a reconstruction of the face using 3D face models, but how much bias towards the 3D face models is introduced in such a process? In forensic casework it is even more important to avoid any bias towards specific face models that are not part of the case data. Is there a way to avoid introducing bias towards any data that is not part of the facial data during a reconstruction procedure? The case of the reconstruction of the unknown face is a difficult one.

1.1 BIOMETRICS & FORENSICS

Biometrics is defined according to Merriam-Webster as: ‘The measurement and analysis of unique physical or behavioral characteristics especially as a means of verifying personal identity’. To correctly identify a person, many biometric characteristics can be used: Fingerprints, Iris, Hand shape, Vein patterns, Faces, DNA, Gender, Gait. Some of the biometric characteristics are innate, some develop in the first years of our lives and others are based solely on behavior. All these biometric characteristics are used in the domain of biometrics to identify individuals. Some of these biometric characteristics have more discriminating power than others. DNA is for example more discriminative than facial images, but might be more difficult to obtain. Facial images are used to relate unknown facial images with facial images of which the origin is known. One of the main advantages of using face images as a biometric characteristics is the acceptance among people. Faces are already disposed to the public, and recordings can easily be made without a person’s express cooperation. Although faces are not always very discriminative, take for example the faces of twins, faces give a strong indication of a person’s identity. The downside of collecting facial images in public is that there is usually only little control over illumination and pose of the face. The variation in pose and illumination makes it difficult to perform a reliable face comparison. Other biometrics, such as Iris, do not suffer from such variation, but need people to cooperate to obtain proper data. Automatic face comparison is based on biometric characteristics of the face only. Although hair color and hair style can also be used for this purpose, these features can easily be changed. Face features that are stable under different illumination and variation in pose of the face are considered to be robust facial features for facial comparison.

1.1.1 LOW RESOLUTION DATA & BIAS

Facial images are common in crime investigation and forensic case work. When facial images are handled and processed with care, they can contribute to evidence in forensic case work. The most prominent issue with forensic facial images is the low quality of the images. Recordings are often made with the purpose of watching the behavior of persons. Therefore, the face region is small and the extracted face images are low resolution. Other issues with forensic facial images are the variation in illumination and the pose of the face. These aspects make it challenging to process facial images. The variation in view of images from the same face, makes it difficult to combine the information from multiple images. In common forensic cases there are multiple recordings of a face under pose. The challenge is to use the information of all these low quality views to reconstruct a frontal face image or 3D model of better quality and higher resolution. However, this reconstruction should not contaminate the forensic case data with data from other sources. This contamination would lead to a bias towards data that is not part of the

forensic case data. This bias, which could also be introduced using statistical facial models or average face models, should be avoided. Throughout the reconstruction procedure the source of the data, and so the reconstruction, should remain clear to perform a facial comparison reliably.

1.1.2 FORENSIC APPLICATIONS

The field of biometrics focuses on the algorithmic part of the comparison process, where performance and computational complexity are important factors. The forensic field focuses on the suitability, evidential value and the bias towards external data in forensic applications. Both fields come together in four applications in the field of forensic biometrics⁵²: In *Forensic Identification* the goal is to identify a person from an open or a closed set of persons. To make a case, a face should be matched with one of the suspects in a forensic case. Important here is the accuracy of the face comparison, evidential value of the data and to avoid bias towards certain faces. *Forensic Investigation* is where traces are matched against a database of persons. One face should be found in a huge set of faces, which leads to a focus on performance in the biometric field and on evidential value in the forensic field. In *Forensic Intelligence* traces from multiple cases are matched and linked to each other. The main idea is that for example it can be proven, with high accuracy, that two facial images on surveillance recordings originate from the same person. *Forensic Evaluation* can be used for forensic individualization. Likelihood ratios describe the ratio between the probability of comparison scores, given that the evidence originates from the suspect and the probability of comparison scores, given that the evidence originates from a unknown person. After calibrating a face comparison algorithm, likelihood ratios can be given for face comparison scores. With these ratios the evidence can be supported with statistical data. So by combining both fields and complying with the concerns of both fields, powerful applications can be implemented.

1.2 FORENSIC USE CASE

There are many forensic cases in which facial images play a role. Based on the input from the Netherlands Forensic Institute a selection of relevant forensic cases is made, where improvements of the current state of face comparison could make a difference. Camera surveillance at, for example, train stations, where cameras are positioned for observing the behavior of people, produce usually poor resolution face data. Because of the positioning of the camera the resolution of the face in the images is low. Data storage and compression might affect the quality of the face data even more. The quality of the faces in these images is currently too low to obtain a proper reconstruction of the face. In the future when the resolution of the cameras increases, the use of these images might become feasible. Currently we need to focus on cases with closeup images of a face. Entrance cameras in shops can be placed for security reasons with special focus to the face area. Due to the precise setup and setting, most of the time high quality frontal shots can be obtained. The current 2D face comparison tools can handle these images without problems, so there is no improvement needed. Another interesting case is the recordings of ATM machines. Because of the money and thus the crime involved, this case is highly relevant from a forensic perspective. Although the setup can be controlled, other aspects like illumination and position of a face are still uncontrolled. The image quality of ATM recordings is reasonable, however in most occasions only gray scale video is available, because of the

recordings during night time. The recorded data could consists of time-lapse recordings, where only a few frames are captured each second. One way to improve image quality is by obtaining a face reconstruction based on multiple images. However, the frames can not be easily combined, because of the differences in pose and illumination of the frames.

1.2.1 FORENSIC SETTING

Imagine a situation at a bank, where a stolen bank card was used to make a transaction at an ATM, see Figure 1.1. The person unaware of the camera in the ATM provided some clear video footage of a moving head in front of the camera. Looking at the screen, watching the cash coming out and keeping an eye on other people lead to some useful video footage for a forensic researcher. Although much data is available, there may be no frontal images of the face available. With the current tooling this leads to a problem, because of the lack of frontal facial images. How can the information of all frames in the video be combined, to obtain a complete face model?

Figure 1.1: Recordings of the ATM with uncontrolled illumination and pose of the face.

1.2.2 FORENSIC FACE RECOGNITION FROM IMAGE SEQUENCES

Video and image sequences are collections of multiple images. In this thesis we refer to these collections of images as sets of frames. Each frame shows a view on the face. A video differs from an image sequence by a fixed time interval between consecutive frames. An image sequence is also ordered in time, but the steps between consecutive frame in the sequence vary, which makes tracking of points impossible due to the large differences in positioning of the face. In this thesis we focus on image sequences, since the recordings could be time lapse recordings. The current procedure in forensics to process video data is to select the most frontal or the best frame from the set of frames. This procedure quickly reduces the amount of data and can be used in cases

where frontal face frames are available. In other cases where there is no best frame or where multiple frames should be chosen, the current procedure and the current face comparison tools fail most of the time. The remaining data, which was discarded, may still provide information about the identity of a person. A simple solution of fusing the comparison scores of multiple frames would not suffice. The gain of fusing low comparison scores is minor and the result might barely outperform the most frontal frame. Although there is no straightforward method to combine multiple frames, face reconstruction methods would help to maximize the use of the image data. The best performance can be achieved using the additional fact that all the data originates from a single source. This source is not a 2D, but a 3D object, which makes the reconstruction process a challenging task. All 2D frames contain partial information of the 3D face model. Therefore, multiple frames are needed to reconstruct an entire face.

1.3 RESEARCH OVERVIEW

In this PhD thesis we focus on how we can combine multiple frames of the same source to obtain higher facial comparison scores in the 2D domain. A schematic overview of the reconstruction process can be seen in Figure 1.2.

Figure 1.2: Instead of selecting the best frame, a new frontal facial image is reconstructed from an image sequence. The reconstructed face is used to perform face comparison with a known reference facial image. The comparison score expresses the similarity between the two faces.

1.3.1 RESEARCH METHOD

In the current forensic procedure for face comparison, state-of-the-art 2D face comparison software is used to automate the face comparison task. Although face comparison software saves the time of manual face comparison by forensic examiners, this approach also has a downside. The face comparison software performs best on frontal images. The performance decreases or comparison methods might even fail for faces under pose or faces with uncontrolled illumination. Due to the frontal requirement large parts of image sequences are unsuitable for face comparison with the current tooling. In most cases these images are discarded and only a small subset of the images is taken into account. The goal of this research is to develop methods for forensic face recognition using extracted 3D information from image sequences. One important aspect of forensic face recognition is that it should be unbiased. Forensic facial comparison should be based on as much information as is available in the image sequence, without using external face

information based on different facial data sources. In the end this will lead to forensic cases where all images take part of the solution, rather than a selection of facial images. From the forensic perspective it is crucial that the adaptation of the image material introduces no bias.

1.3.2 RESEARCH QUESTIONS

In this PhD thesis we investigate the following question: How can we use multiple images from a sequence which individually are considered not usable for a forensic procedure to reconstruct a face model that can be used in the forensic face comparison procedure?

In the next chapters we will look further into the following sub-questions:

- How can we extract 3D information from multiple images for forensic face recognition?
- Which reconstruction method is the most suitable in forensic cases?
- What are the requirements for landmark based 3D reconstruction of random 3D objects?
- How can we perform landmark based reconstruction on rigid face data?
- How can we obtain a coarse 3D face reconstruction using a realistic forensic case image sequence without introducing bias towards a model?
- How can we improve the coarse reconstruction based on a realistic forensic case image sequence to obtain a dense 3D reconstruction?

1.4 3D FACE RECONSTRUCTION METHODS

There are different ways to reconstruct 3D face models. In this section we focus on two types of face reconstruction methods. The first type is a reconstruction method based on existing 3D models, the second type is a model-free reconstruction method based on features extracted from multiple facial images.

1.4.1 MODEL-DRIVEN RECONSTRUCTION

The model driven reconstruction method is based on existing 3D face models, that are captured in a controlled environment. The models are aligned precisely, so that statistics of all positions in the face can be calculated. The model consists of shape information only or a combination of shape and texture information. The resolution of the model depends on the resolution of the used 3D face models. By varying the parameters of the model, different faces can be generated, based on the calculated statistics. To obtain a frontal facial image a rendering of the model can be made. To reconstruct a particular face, the parameters of the model are adjusted until the model fits a 2D face image, see Figure 1.3 for an example of a reconstructed face. Many factors such as illumination, pose and image quality can influence the shape of the fitted 3D face model. Although the face models are dense and look smooth, there is one major issue with this approach. How can the integrity of forensic data be maintained, if the reconstructed model is based on a collection of external 3D face models? The model is based on a collection of faces and can only generate new combinations of this collection. The face model might look discriminative, but contains face data from many faces and not solely the face data of the original image

sequence. Since the model-driven reconstruction method violates the forensic integrity, we can not use a model-driven reconstruction method.

Figure 1.3: Reconstruction using a model based approach. Left: Original face image. Right: Reconstructed face image.

1.4.2 MODEL-FREE RECONSTRUCTION

The model-free reconstruction method is a data driven method. The 3D face is reconstructed using features that are derived from an image sequence. These features can be pixels, landmarks or patches on the face. For each frame the correct pose and position of the face is estimated, based on the features. The density of the reconstructed face is related to the number of matching features in the image sequence. Since the number of matching features is usually low, the reconstructed 3D face is coarse. A second step, in which additional features are involved, is required to obtain a dense reconstruction. In the case of landmarks, see Appendix B, the initial reconstruction is a set of reconstructed 3D points. Patches between the landmarks need to be defined to obtain the surface for a 3D model, see Figure 1.4. In contrast to the model-based reconstruction method, the model-free reconstruction method is unbiased, and therefore suitable for forensic face comparison.

1.5 OVERVIEW

In Chapter 2 an overview of the current methods for 3D face reconstruction is given. Both model-driven and model-free reconstruction methods are reviewed. Based on the forensic context the most suitable method is selected. Next, Chapter 3 explores the possibilities of reconstructions based on landmarks only. Both the possibilities and requirements for landmark based reconstruction methods are explored using random 3D objects. Chapter 4 continues exploring the landmark based reconstruction method using a styrofoam head model. The rigid head

Figure 1.4: Left: Landmark model with defined set of patches. Right: Landmark model with texture from 2D frames.

model is a first step towards real face data. In Chapter 5 we introduce a reconstruction algorithm for image sequences of faces. The final 3D reconstruction is a coarse, textured 3D model. We minimize the possibility of introducing bias during the reconstruction process. The reconstructed model can be used to obtain a frontal reconstruction of the face. Chapter 6 introduces a method to obtain a dense textured 3D model from multiple views. Chapter 7 concludes this thesis by answering and discussing the research questions.

2

Reconstruction Methods

CHRIS VAN DAM
ROBIN VAN ROOTSELER
LUUK SPREEUWERS
RAYMOND VELDHUIS

We start by taking a closer look at the current methods for 3D face reconstruction described in literature. Both model-free and model-driven methods are taken into account. Although general reconstruction methods in other fields of research also could be of interest, we limit our scope to face reconstruction methods. In this chapter we summarize the current 3D face reconstruction methods. We use this knowledge to define the starting point of our research. Based on the forensic ATM case example, we select promising 3D face reconstruction methods suitable in a forensic context. This work was published in 2010 as an internal report.

RECONSTRUCTION OF 3D FACE MODELS: AN OVERVIEW⁸⁷

2.1 INTRODUCTION

The main focus of this overview is on 3D face reconstruction from surveillance video or image sequences from surveillance video. Most of the current work in the forensic world based on surveillance video is performed by human experts. Face comparison of a suspect is done by comparing the best frame of a surveillance video with photos taken from the suspect at the police station or photos from identification documents, such as a passport or a driver license. The research is done by human expert examination, because there is currently no software available

which performs better than the human experts. Though there is a lot of software on the topic of face recognition, most video material suffers from low resolution, compression artifacts and varying lighting conditions, which are a great bottleneck on the current software applications.

Surveillance cameras are placed in many public places, but there is no standardization for the placements of the cameras. Many cameras are installed by security companies, which will use their own guidelines for the placement of cameras. The recording conditions are often not optimal with respect to proper lighting, correct camera angles or positioning. This makes automated face recognition difficult and sometimes even impossible.

This literature overview will summarize the current state of the art in 3D face reconstruction. Keep in mind that this overview is written with the desire to have a new forensic tool or method for 3D face reconstruction and face comparison. Both older and current methods to improve the 3D face reconstruction will be reviewed in this overview. The main problems of forensic 3D face reconstruction are caused by low resolution video, illumination, pose differences, uncalibrated cameras and partial occlusion of the face. Other aspect such as suspects in disguise and suspects wearing masks won't be taken into account, because in those cases even for human experts it's impossible to recognize a suspect based on the facial information only.

2.2 OVERVIEW

Multiple overviews and surveys of 3D and 2D face recognition are published before. Since this overview is only about 3D face reconstruction, the 2D surveys in this overview are omitted. Bowyer et al.¹³ give an overview and comparison between multiple 3D face recognition methods, some in combination with 2D intensity images. Scheenstra et al.⁶⁹ give a small survey on some 3D face recognition methods. In another survey of Bowyer et al.¹⁴ an extension of the survey in¹³ is given. Each algorithm is categorized and compared according to its performance. A distinction was made between 3D shape recognition and combined 3D and 2D methods. Abate et al.² give an overview of all important face databases and a large overview of 2D and 3D face recognition methods. Zhou and Chellappa¹⁰⁰ formalize and discuss the process to compare 2D image, image sequences and video with each other. A structured technical description of multiple comparison and reconstruction methods is given. Widanagamaachchi and Dharmaratne⁹³ summarize and discuss the most common 3D reconstruction approaches. A recent survey by Levine and Yu⁴⁸ focus on the 3D reconstruction methods using only one single 2D image.

There are only a few mainstream methods with the ability to reconstruct a 3D shape model from 2D image data. Some methods use *Landmark Based* feature points in 2D images to reconstruct a 3D face model. Another related method is *Structure from Motion*, which tries to estimate the 3D structure from the motion using landmarks in 2D images. Also *Shape from Shading* is used, in which the 3D structure is estimated from the shading of an object. A fourth method is called *Shape from Silhouette* which uses silhouette or contour data to build a 3D model. This overview won't deal with *Shape from Stereo*, because the focus is on normal uncalibrated video and image sequences which do not contain any stereo information. Finally the *Morphable Models* will be described which is a combination of multiple of the methods mentioned here, combined with statistical 3D information. All methods will be explained and summarized in the next sections.

2.3 LANDMARK BASED

There are many feature based tracking methods and there is not always a clear distinction between *Structure from Motion*, *Landmark Based* methods and *Morphable Models*. So some of the methods described in this section may be classified differently by others.

Kuo et al.⁴⁵ show a method to estimate the depth from a single frontal view image. They use a priori anthropometric information for this estimation. After the tracking of the face in 2D, a subset of the MPEG-4 facial feature points is automatically located in the images. They use a hierarchical structure to estimate missing or questionable feature points. Next they start estimating the relative depth of the feature points with respect to a chosen specific feature point. They use the distances between sets of 2D feature points in the neighborhood of a feature point to estimate the depth, based on the anthropometric relation between the side and front view. They used three different statistical schemes (Minimum Mean Square Error, Minimum Mean Absolute Error and Maximum A Posteriori) to estimate the relation. They report an estimation error of about 18%.

Hu et al.³⁵ use an analysis-by-synthesis manner for face recognition based on a 2D-to-3D approach. This work is closely related to the *Morphable Model* described in section 2.7. They describe a full automatic method with higher speed. Their system requires a single frontal face with normal illumination and neutral expression. With a semi-supervised ranking prior likelihood model they accurately locate 83 feature points in the frontal image. A 3D face PCA model from the USF Human ID dataset is used for reconstruction. They iteratively find the correspondence between the 3D PCA model and the 2D feature points. The texture of the image is projected orthogonally on the reconstructed 3D face model. A comparison is made with conventional PCA and LDA algorithms trained on one frontal image. In general the face recognition accuracy is higher than the accuracy of conventional algorithms. They conclude that their method is fully automatic and faster than other 3D approaches.

Jiang et al.³⁸ elaborate on the work of Hu et al.³⁵. They describe the system in some more detail and add a comparison with the *Morphable Model* described in section 2.7, see Table 2.1.

Table 2.1: Comparison with Morphable Models by Vetter et al.

	Vetter et al.	Jiang et al.
Input	Single face with arbitrary pose and illumination.	Single frontal face with homogeneous illumination and neutral expression.
Initialization	Some manual initialization.	Fully automatic.
Shape	Shape parameters are estimated by optical flow.	Recovered by 2D-3D fiducial feature points and statistic model.
Texture	Texture parameters are estimated from texture error.	Direct mapping from 2D image.
Speed	About one minute per face image.	About 5 seconds per face image.

A method to reconstruct a 3D face model from video using a generic 3D model is described in Kalinkina et al.⁴⁰. Kalinkina et al. stated that the method isn't fully automatic and doesn't work in real-time applications. The method needs at least five images as input for the reconstruction of the 3D model. The first step of the method is the calibration of the cameras used to obtain

the images. The calibration is performed by the POSIT algorithm, which needs a reasonably good estimation as a starting point. After manual specifying 38 characteristic feature points, the POSIT algorithm tries to minimize the difference between the 2D points and a projection of the associated 3D points from a generic 3D face model. After the camera calibration the 3D face model can be created using 3D stereo reconstruction. With the use of *Radial Basis Functions* other vertices of the 3D model can be found by interpolation. The 3D model can be refined using the silhouette contours of the images. The contours of the images consist of user-drawn Bezier curves. The contours of the estimated 3D model are extracted automatically. Finally the 3D model is deformed, so that the outer contours match the contours of each image.

Ishimoto and Chen³⁷ focus on pose-robust face recognition based on 3D face reconstruction. To reconstruct the 3D model they use the *Factorization Method*. This method can robustly reconstruct shape and rotation from a sequence of images under orthography. Correspondences between the images are needed for the 3D reconstruction. They use 90 manual defined feature points. These feature points can also be automatically extracted using an *Active Shape Model*. After the reconstruction of the 3D face model, new images can be generated by projections of the 3D face model. The texture is extracted from a frontal image and warped piecewise onto the model.

2.4 STRUCTURE FROM MOTION

Structure from Motion (SfM) is sometimes also called *Shape from Motion*. The motion of a 3D object in video is used to estimate the shape of the object.

Fua²² shows a method to model a head from video without calibration data. The method is applicable to any modeling problem. He describes a robust algorithm that uses a generic head model to recover the shape and the motion. Fua stated that model-free SfM is too sensitive to noise to be directly applicable for modeling. The generic head model is used to produce a regular mesh on the frames. The intrinsic camera parameters are approximated. *Bundle Adjustment* is used for reconstruction, which is a nonlinear optimization problem and needs a close initial starting value. Five landmarks in one of the frames are used as initialization. Using the *Levenberg-Marquardt* algorithm the error with respect to the camera positions and 3D coordinates is minimized. The *Bundle Adjustment* is made more robust by using regulation constraints and a recalculated weighting error function. The final 3D model is compared with laser scan data and seems to be a good approximation of the model up to an affine transformation.

Shan et al.⁷¹ present a new model-based bundle adjustment algorithm to reconstruct a 3D model from an image sequence with unknown motion. *Model-based Bundle Adjustment* does not need a prior 2D-to-3D association and has less unknowns and constraints than *Classical Bundle Adjustment*. The algorithm uses model parameters and semantic meaningful points, instead of isolated 3D feature points, for the 3D model. The model parameters and camera parameters can be estimated from the feature tracks in a sequence of images by minimizing the sum of squared error between the observed image points and the projected feature points. The model can be linearly deformed to fit the images. The neutral face and the possible deformations are designed by an artist. The user could mark some semantic point constraints. A cylindrical texture map is extracted from the images. A comparison on synthetic data was made with *Classical Bun-*

dle Adjustment. The experiments show that *Model-based Bundle Adjustment* performs better than *Classical Bundle Adjustment*.

Roy-Chowdhury et al.⁶⁶ describe a method to reconstruct a 3D face model from video using SfM in combination with a generic 3D model. The generic 3D model is used after obtaining the estimation of a standard SfM algorithm. The method only uses the generic 3D model to correct errors based on local trends, but the model isn't used to fuse the depth of both models. The correction is done using simulated annealing and a Markov Chain Monte Carlo (MCMC) sampling strategy. The paper contains only some visual results, because of the lack of a proper 3D ground truth.

In⁶⁷ Roy-Chowdhury, Chellappa and Gupta present two methods to model 3D faces from video. The first method is similar to the method in⁶⁶. The optical flow paradigm is used for 3D reconstruction. Pairs of two frames from the video are used for the reconstruction of a depth map. All depth maps are fused together using a stochastic approximation. The final depth map is fused with a generic 3D model. Again the MCMC sampling strategy is used, which is a method to solve minimization problems. Still no comparison with other methods was made, though some results on a public 3D model were published. The second method is a *Shape from Silhouette* method and is described in section 2.6.

In²⁰ Fidaleo and Medioni describe a model-assisted method for 3D reconstruction. The 3D model is reconstructed from a single consumer quality video camera. A generic model can be useful, when it is used at an appropriate point in the reconstruction process. If the reconstruction is performed from single images, the reconstruction would result in a generic-like 3D model. The method uses a sequence of images to create an accurate 3D model. A 3D face tracking algorithm gives an initial estimation of the head pose and a mask for the face. The estimation is established with the help of a generic 3D model, which can be done without biasing the final reconstruction result. Optimal views are selected and used to select a set of feature correspondences for each successive image pair. A global optimization is performed by *Bundle Adjustment* to refine the estimation of the camera pose. The result is a coarse estimation of the structure of the face. A dense model of the face is acquired by interpolation via radial basis functions. Epipolar line correspondences are used to create a disparity volume for each successive image pair. A dense 3D point cloud can be found by triangulation of the disparity volume. Outliers can be eliminated by *Tensor Voting*. A connected surface is fit on the 3D point cloud and the texture is extracted from a frontal image. Only a visual evaluation of the system is performed.

In⁵¹ Marques and Costeira explore the use of 3D reconstruction from multiple images for 3D recognition under strong pose. No prior knowledge of the cameras and images is used to create the reconstruction of the 3D face. They stated that feature points are more reliable to illumination changes, because feature points can cope with missing data. They assume that the input images have a neutral expression, the input images are well modeled by orthographic projection and the feature points are known. The method is based on the 3D relation between the different views of the same object. By using 13 feature points the motion, shape and translation can be calculated. The results can be enhanced by estimating the missing feature points in the images.

Zhang et al.⁹⁷ give an extensive overview of six major Shape from Shading (SfS) algorithms. All algorithms are grouped into minimization, propagation, local and linear methods. The essence of each algorithm is explained. The different constraints for the minimization methods are described. To compare the algorithms five images, three synthetic and two real images, were used for 3D reconstruction of a surface. Six algorithms were implemented and tested on performance, accuracy and speed. The results were visually compared. All algorithms produced poor results for the synthetic images and the results from the real images were even worse. Finally Zhang et al. conclude there is no logical connection between the results for the synthetic images and the results for the real images. Combining SfS with other methods like stereo or range data or the use of a more elaborated reflectance model would probably improve the results.

Zhao and Chellappa⁹⁹ present multiple single image based SfS methods for face recognition, that are robust to pose and illumination changes. Zhao and Chellappa use a 3D depth map to bypass the 2D-to-3D process. The illumination estimation is based on the Lambertian reflectance model. Three different cases of pose problems with varying assumptions are discussed. In the hardest case the illumination can change and there is no prior class information available. They assume only rotation around the depth axis. They try to match the depth map with an input image to estimate the pose using synthesized images of the depth map. At the same time they also introduce two more parameters to estimate the illumination.

$$(\mathcal{V}^*, \alpha^*, \tau^*) = \underset{\mathcal{V}, \alpha, \tau}{\operatorname{argmin}} (I_M^R(\mathcal{V}, \alpha, \tau) - I^R)^2 \quad (2.1)$$

Where $\mathcal{V}$ is a rotation around the depth axis and α, τ are two angles that describe the illumination direction. I^R is the input image and I_M^R is a generated image of the 3D model. They use the *Self-ratio* image to eliminate the effect of varying albedo. Based on the estimations a new frontal face image can be generated for face recognition.

Sim and Kanade⁷² propose a model- and exemplar-based approach for face recognition. Based on a model many more exemplars can be synthesized and used in the training of the face recognition system. They use a statistical 3D model to guide the SfS recovery of the depth map from a single image. They use the standard Lambertian equation augmented with an error term for the non-Lambertian shadows and reflections.

$$i(x) = n(x)^T s + e \quad (2.2)$$

Where $i(x)$ is the intensity of a pixel in point x , $n(x)$ is the normal of the surface in x , s is the direction of a single light source and e is the non-Lambertian error term. The normals and the error term are learned from a statistic model based on a face database. After the recovery new images can be generated from the 3D model under different illumination. Those new images can be used to train an exemplar-based classifier.

Dovgord and Basri¹⁹ combine the method of Zhao and Chellappa⁹⁹ with a statistical reconstruction method to reconstruct a 3D face from a single frontal image. Given an image under a known illumination direction, Dovgord and Basri provide a closed-form solution which satisfy both statistical and symmetry constraints on the facial shape and albedo. This is the first

closed-form solution for 3D face reconstruction from a single image within a few seconds. A statistical PCA face model is created from 130 3D heads. Every 3D face shape is constrained by the statistical PCA model. Combining this with the albedo free brightness constraint a least squares solution of the problem can be found and the depth can be estimated from the statistical model. The algorithm was tested on the Yale face database B. This database contains frontal face with varying illumination conditions. One of the main problems is the inaccurate result on some asymmetric faces.

Xu et al.⁹⁵ present a theory for combining the effects of motion, illumination, 3D structure, albedo and camera parameters in an image sequence of a perspective camera. Given an arbitrary image sequence it is possible to recover the 3D structure, motion and illumination simultaneously using the bilinear subspace. They show that the set of all Lambertian reflection functions of a moving object, with attached shadows at any position, illuminated by distant light sources, lies close to a bilinear subspace consisting of nine illumination variables and six motion variables. The illumination parameters can be estimated with *Spherical Harmonics*. The motion parameters are estimations of the rotation and translation of the 3D model. A detailed mathematical derivation of the bilinear space is presented in this paper.

Roy-Chowdhury et al.⁶⁸ provide a method for learning the pose and illumination conditions from video using a generic 3D model and *Spherical Harmonics*. They start by estimating the motion and illumination conditions of a video based on the differences between successive images. The motion and illumination are estimated in an iterative algorithm. These estimations are used to render images of a 3D gallery model. The rendered images can be compared with the frames of the video. The comparing metric should have the ability to integrate over all the frames, ignoring the ones with a wrong identity. Roy-Chowdhury et al. propose two distance metrics, where d_{ij} is the distance between the synthesized image S_j and the video frame P_i :

$$d_1 = \arg \min_j \min_i d_{ij} \quad (2.3)$$

$$d_2 = \arg \min_j \max_i d_{ij} \quad (2.4)$$

Both of these methods suffer from a lack of robustness. The *min* can be replaced by the 20th percentile and the *max* by the 80th percentile to make it more robust. The effectiveness of the method is showed on a private video database with both arbitrary pose and arbitrary illumination.

Boom et al.¹¹ present a method to correct the illumination variation in a single face image under uncontrolled illumination conditions. Their SfS method uses the Phong lighting model.

$$I(p) = c(p) \cdot i_a + c(p) \cdot \vec{n}(p)^T \vec{s} \cdot i_d \quad (2.5)$$

Where $I(p)$ is the illumination in point p of image I , $c(p)$ is the albedo value of the model in point p , i_a is the intensity of the ambient lighting, $\vec{n}(p)$ is the normal in point p , $\vec{s}$ is the light direction and i_d is the intensity of the diffuse light. An additional term is added to estimate the shadows in the image. Using a grid of light directions, the light intensity of a single gray scale image can be estimated by using a generic 3D face shape model and a mean albedo value. These estimations can be used to calculate an initial face shape estimation using *Lagrange Multipliers*

and the constraints from the Phong reflectance model. Then they calculate the depth map of the face image based on a PCA model of 3D range images. Finally the albedo can be estimated. Based on the evaluation of the grid of light directions, the best parameters for the illumination direction in the original image can be found. A reconstruction of the original image can be made, based on all the estimated parameters. Experiments on the FRGCv1 database show improvement over previous face recognition algorithms.

2.6 SHAPE FROM SILHOUETTE

Lee et al.⁴⁷ present a method to reconstruct a 3D face shape from multiple silhouette images. The method is independent of any color or texture information in the face. They start with applying PCA on a database of 3D faces. With this new *Eigenheads* model, new face shapes can be created based on the principle components. All new face shapes are linear combinations of the principle components.

$$H(\alpha) = h_o + \sum_{m=1}^M \alpha_m h_m \quad (2.6)$$

Where h_o is the mean head, h_m is the m -th eigenhead, α_m is the weighting of the m -th eigenhead and $H(\alpha)$ is the new head. This is similar to the model used in the *Morphable Models* in section 2.7. The point to point correspondence for the 3D PCA face model is established by aligning 26 landmark points. The silhouette images are matched with rendered 3D views of the eigenhead model by minimizing the cost of the transformation, rotation and zooming using the *Levenberg-Marquardt* algorithm. Lee et al. introduce a boundary-weighted XOR cost function, which takes the distance to the boundary of the silhouette images into account to compensate for the partial contours (missing hair area and the back of the head) of the 3D *Eigenheads* model. The whole algorithm is optimized using a downhill simplex method, which requires only function evaluation. The texture of the 3D model is obtained from the corresponding real images of the contour images. All real images are projected onto the 3D model and weighted by the angle between the normal and the viewing direction. In experiments with synthetic and real data, they demonstrate that 2D silhouette matching captures the most important 3D features of the human face. Only visual results are presented in the paper and no comparison with ground truth data has been made.

Gupta et al. present in²⁷ and in⁶⁷ a method to reconstruct a 3D face model from a video based on the outer contours of a face. They use a generic 3D face model for the reconstruction. They limit the pose to an estimation along the azimuth angle. The Kanade-Lucas tracker is used to find the tip of the nose which is used for alignment. Edge maps of the generic 3D model are calculated with steps of 5° along the azimuth angle using the *Canny Edge Detector*. Next the edges of each frame of the input video are calculated with the *Canny Edge Detector*. All frames are aligned to the nose tip of the face. The Euclidean Distance Transform is used to determine the pose of the face in each frame. A global deformation ensures that the 3D face model matches the approximated shape of each frame and that the internal features are aligned. In the next step local deformations are used to individualize the generic model. Local perturbations in the (x, y, z) direction on two different resolutions are used to deform the face model. The same cost function based on Euclidean Distance Transform is used to optimize the local perturbations.

Finally the texture of one of the frames is used for visualization of the 3D face model. Successful experiments on real video data are reported, but not described in detail.

Keller et al.⁴² try to find a robust 3D reconstruction method with a small contour reconstruction error. They also want to describe the strength of the constraints of the 3D model constructed by the contour images. There are two sub problems to construct a 3D model: feature extraction and fitting of the contours. Instead of using silhouettes they use the contours of the face, which require a robust distance measure to compensate for missing edges in a contour image. Contours are defined as edges, whose image locations are invariant to lighting, which is different from a silhouette or a normal edge image. A statistical PCA 3D model is used for the reconstruction. The contour of the face is found by determining the front and back facing polygons in the rendering. Edges between a back and front facing polygon are considered as candidates for the contour. Again the Euclidean Distance Transform is used as distance function:

$$D(I, R(p)) = \frac{1}{|S|} \sum_{(x,y) \in S} d(x,y) \quad (2.7)$$

where I is the binary contour image, S is a 'on' pixel in $R(p)$, the rendered image. And $d(x,y)$ is the Euclidean distance to the nearest contour pixel. Three modifications of this distance function were tested to make it more robust against missing contours in the input image. The Downhill Simplex algorithm was used to minimize the fitting of the contours. Tests were done on synthetic, semi-synthetic and real images. The results support the view that contours do not constrain the shape tightly. They conclude that the pose of the face can be recovered with high accuracy, but the shape can often differ greatly from the ground truth. This predestinates contour matching as a part of a system and not a system on its own.

2.7 MORPHABLE MODELS

3D Morphable models (MM) were proposed by Blanz and Vetter⁹ and are based on analysis by synthesis. The goal is to represent a novel face in an image by using model coefficients and provide a reconstruction of 3D shape and the corresponding texture. The morphable face model is based on a vector space representation of the shape and the texture of faces. The shape vector contains a fixed number of Cartesian coordinates: $S = (x_1, y_1, z_1, \dots, x_n, y_n, z_n)^T$ and the texture vector contains the corresponding RGB values: $T = (R_1, G_1, B_1, R_2, \dots, R_n, G_n, B_n)^T$. Principle Component Analysis is performed on the vectors S_i (shape) and T_i (texture, actually the albedo) of m example faces $i = 1 \dots m$. The correlation between shape and texture data is ignored. The authors assume independence between shape and texture. The eigenvectors of the covariance matrix of S and T form an orthogonal basis:

$$S = \bar{s} + \sum_{i=1}^{N_s} \alpha_i \cdot s_i, \quad T = \bar{t} + \sum_{i=1}^{N_t} \beta_i \cdot t_i \quad (2.8)$$

where N_s and N_t denote the number of PCA components of the shape and texture respectively. Using this PCA formulation the probabilities of a shape and a texture are given by their param-

eters, which will prove to be useful when minimizing the energy function:

$$p(S) \sim e^{-\frac{1}{2} \sum_i \frac{2}{s_{s,i}}}, \quad p(T) \sim e^{-\frac{1}{2} \sum_i \frac{2}{t_{t,i}}} \quad (2.9)$$

An image of a face can be rendered by projecting the 3D shape to a 2D image frame. First a rigid transformation maps the object-centered coordinates, S , to a position relative to the camera in world coordinates:

$$W = R_x R_y R_z S + t_w I_{1 \times N_v} \quad (2.10)$$

where N_v denotes the number of vertices of the shape model. After the rigid transformation a perspective projection maps a vertex i to the image plane in (x_i, y_i) :

$$x_i = t_x + f \frac{W_{1,i}}{W_{3,i}} \quad y_i = t_y + f \frac{W_{2,i}}{W_{3,i}} \quad (2.11)$$

The albedo of the face is illuminated by using the Phong reflectance model that accounts for the diffuse lighting and the specular reflection on a surface. Since input images may vary a lot with respect to the overall tone of color a color transformation is applied. The MM has in this form a total of 422 parameters (199 shape, 199 texture, 3 pose angles, 3 3D translation, 2 2D translation, 1 focal length, 3 ambient light intensities, 3 directed light intensities, 2 angles of directed light, 1 color contrast, 6 gains and offsets of color channels).

Stochastic Newton Optimization^{49, 38} is used to minimize the cost function to avoid local minima in the cost function. The convergence properties of such algorithms are however limited⁶⁴. The cost function takes into account the difference between the synthesized image and the image from which a 3D model has to be extracted. It also takes into account the reasonability of α and β using their probability density function:

$$E = \underset{\alpha, \beta, \dots}{\operatorname{argmin}} \frac{1}{\sigma_I^2} \sum x, y \|I^m(x, y; \alpha, \beta, \dots) - I(x, y)\|^2 + \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sigma_i \frac{\beta_i^2}{\sigma_{T,i}^2} \quad (2.12)$$

An important step in MM is generating the model. In order to construct a MM, a set of example 3D laser scans are put into correspondence with a reference laser scan. Using a modified optical flow algorithm a consistent labeling of all vertices across all scans can be established. More specifically: the 3D points that are equivalent across faces are put into correspondence. Facial landmarks like the center of the eyes, the corners of the mouth and the tip of the nose of different faces will all have the same index in the shape and texture vector. Optical flow algorithms are usually based on the assumption that objects in an image sequence conserve their brightness⁵ as they move across the images. Although this assumption is not valid for a pair of images taken at two discrete moments, it has been shown⁹ that optical flow algorithms may be applied successfully to the pair of images. In 2003 the algorithm has been refined⁴ by regularizing the 3D morphable models to yield fewer artifacts.²⁶ suggests to combine the algorithm of Fast-AAM with Thin Plate Splines for 3D data alignment to avoid the local minima problem of optical flow. In³⁶ another method of alignment is discussed using mesh resampling (Krishnamurtby⁴⁴). Another contribution is the multi-lights model. If you set enough number of

lights around an object and the brightness of each light can change independently, the arbitrary illumination of the object can be simulated.

In¹⁰ an extension is proposed to use five regions: the eyes, nose, mouth, surrounding area and the complete face. Using this approach the number of shape and texture parameters is multiplied by 5 which makes it possible to model more variations and more details in those regions. Using the Mahalanobis distance measure on the α and β parameters an identification experiment has been conducted on the CMU-PIE and Feret database. The average score on Feret ba-bk was 95.9% correct identification. Matching on the α and β parameters is referred to as *coefficient-based recognition*. Another approach⁸ generates a novel view of the model. A frontal view with standardized lighting is used as input to a 2D recognition algorithm. This is referred to as *viewpoint-transformed recognition*.

In⁵⁰ the model is based on the BU-3DFE database. They make a model per frame and then apply a weighted temporal fusion scheme to make recognition based on image sequences more reliable. Given the posterior probability P_i^{SVC} expressing the probability that the face in an image frame at time t belongs to class i based on a support vector classifier (SVC), the temporal fused probability is given by:

$$P_i^t = \frac{(\omega \cdot P_i^{t-1} + (1 - \omega) \cdot P_i^{SVC})}{\sum_i (\omega \cdot P_i^{t-1} + (1 - \omega) \cdot P_i^{SVC})}, i = 1, 2, \dots, n \quad (2.13)$$

in which ω is a forgetting factor.

The fitting of the MM can only succeed with a good initialization of the parameters, especially the rotation and translation parameters. These can be initialized by manually annotating a number (minimum of 5) of landmarks on an image or by automatic detection^(33, 6). In⁹¹ they propose to use silhouettes and model-based bundle adjustment to automate this initialization process. Another contribution is the usage of spherical harmonics⁶². A downside of the used Phong model is that it takes only one directed light source into account. By using 9 spherical harmonics (SH) most light patterns can be synthesized. In⁹² a link has been established between SH and the 3D statistical head model. For the optimization they split the face in two parts. The face feature area (eyebrows, curvature of the nose, eyes and mouth) is used for texture optimization and the skin area is used for lighting optimization. The SNO is replaced with Levenberg-Marquardt (LM) optimization without explanation.

In⁴⁸ several MM algorithms are compared: Stochastic Newton Optimization (SNO,^{9, 7, 10}), Inverse compositional image alignment (ICIA,⁶³), Linear shape and texture fitting (LiST,⁶¹), Shape alignment and interpolation method correction (SAIMC,³⁸). SNO was used when the MM was introduced in 1999. The ICIA + MM method uses an inverse shape projection mapping which makes the algorithm faster, but since the model does not permit shading it is not able to handle direct light sources. The LiST method uses an orthographic projection instead of a perspective one and tries to linearize the non-linear optimization of the energy function. All the methods need five or more manual selected feature points for initialization. Unfortunately there is no quantitative or comparative analysis of the reconstruction accuracy for the methods that are discussed. The main difference between the different optimization or fitting algorithms is the way in which the Jacobi (J) matrix is computed. SNO computes J at every iteration, AAM fitting assumes J to be constant. ICIA changes the energy function so that J is constant to a first

order approximation. 3D ICIA uses a constant J for a specific pose.

In⁴⁶ a 3D morphable model is used to extract facial features from range images. They fit the input range image and the 3D morphable model and use the shape coefficients of the newly synthesized 3D face as feature. Extra features can be extracted from an optional texture image.

In⁶⁴ and⁶⁰ Romdhani addresses several issues of minimizing the energy function. To avoid getting stuck in a local minimum more terms can be added to the energy function to make it convex and smooth, so that it has a single optimum. This ensures that any optimization algorithm would find this unique global optimum. Making the energy function convex is challenging, since the image formation process is nonlinear, which yields a non-convex energy function. Romdhani proposes the use of image features like edges and specular highlights as additional terms in the energy function to make it more convex, with fewer local minima.

The Basel Face Model⁵⁷ has been introduced in 2009. It provides a publicly available 3D Morphable Face Model⁵⁸ and can be used for research. The model is derived from 200 faces (100 male, 100 female) using a 3D scanner from ABW-3D. The faces are parameterized as triangular meshes with 53490 vertices and an associated color. The results on CMU-PIE and FERET using this model are an identification rate of 91.3% and 95.8% respectively.

2.8 PAPER CLOSURE

There are two promising approaches which lead to proper 3D face reconstruction results. The first one is the MM-approach. This approach is able to reconstruct fully textured 3D models of a face. The MM-approach reconstructions can be obtained from image sequences or even single images. Although the reconstruction capabilities are plenty, embedding them in the forensic context is difficult. The MM-approach is a model based approach and introduces additional statistical face information. Therefore this approach isn't suitable in a forensic context. The second method is a data driven SfM method that is based on landmarks. The main challenge using this approach is to obtain a dense 3D face reconstruction. The quality of the reconstruction depends in particular on the number and precision of the landmarks. In this PhD thesis we choose the landmark based approach as starting point, because of its suitability for forensic face comparison.

3

3D Reconstruction from Video Sequences: Random Point Clouds

CHRIS VAN DAM
LUUK SPREEUWERS
RAYMOND VELDHUIS

In this first conference publication we explore the possibilities of a landmark based reconstruction method. This work was published in 2012 in the WIC Symposium. First of all we need an error metric that indicates the quality of a reconstructed landmark model. In this publication we calculate a 2D and a 3D error for the reconstructions. The landmarks are obtained from a random 3D point cloud and do not represent a model of a face. The main issue we address in this publication is to determine the number of frames and the number of landmarks needed, to obtain an accurate reconstruction of the point cloud. On a high quality face up to 50 landmarks can be distinguished, on low resolution data only a subset of these landmarks can be found. The minimum number of frames and landmarks are an indication whether or not a landmark based method can be used for the reconstruction of faces. For enhancing the realism of the experiments, noise is added to the landmarks to model the error in the manual landmarking process of the forensic researcher. In Appendix B more information about projection models can be found.

3.1 ABSTRACT

Facial comparison in 2D is an accepted method in law enforcement and forensic investigation, but pose variations, varying light conditions and low resolution video data can reduce the evidential value of the comparison. Some of these problems might be solved by comparing 3D face models: a face model derived from CCTV camera footage and a reference face model acquired from a suspect. In our case we will assume uncalibrated CCTV footage, because the original camera setup may be destroyed or replaced after the incident, so precise camera information is no longer available. In contrast to other statistical methods, like Morphable Models, we would like to use no additional statistical information at all. Our method is based on a projective reconstruction of landmarks on the face and an auto-calibration step to obtain a 3D face model in a Euclidean space. In our experiment the effect of the number of frames and noise on the landmarks is explored for 3D face reconstruction based on landmarks. An estimation of the 3D face shape can already be obtained using 25 points in 30 frames.

3.2 INTRODUCTION

In forensic research anno 2012 most of the law enforcement services still use 2D frontal facial comparison. Although this can give good results for frontal or near-frontal reference faces, many problems still arise due to pose variations, varying light conditions and low resolution video data. One way to improve facial comparison would be to compare 3D facial models instead of 2D models, since in most cases there is much more information available in CCTV (Closed-Circuit Television) camera footage. The use of 3D face models requires a change in the technical infrastructure of the law enforcement services and their current methods, but we think that this method can improve the facial comparison results by taking advantage of more information available in the original evidence.

Next to eye witnesses, the most common source of evidence in street crime, burglary and robbery cases is CCTV camera footage. In this paper we will take a specific case into account: fraud at an ATM with an uncalibrated camera installed. The suspect is close to the camera and therefore there is much perspective distortion in the frames of the camera footage. We assume that there is no information available of the original camera, because in many cases the original camera setup may be destroyed or replaced after an incident. So the only data available is CCTV camera footage of the suspect, mainly containing footage of the suspect's face. Our goal in the Person Verification 3D project is to create a 3D facial reconstruction of the suspect, which can be used for 3D facial comparison. In this paper we will use landmarks in multiple 2D frames to obtain an initial estimation of the camera parameters and the 3D shape of the face. In our experiments we determine the minimum number of points and frames needed to obtain an accurate reconstruction of a simulated face model. Next we determine the maximum noise that will still allow us to obtain a precise reconstruction. Finally we do some experiments with auto-calibration of the reconstruction to validate if the methods described in this paper can be applied on face models.

3.3 BACKGROUND

Our problem, where the face of the suspect is moving in front of a static camera, is equivalent to a problem where the camera is moving and the suspect is static. So for each frame $[i = 1..M]$ we have to find the internal and external camera parameters of that specific frame. The static shape of the face can be described by $[j = 1..N]$ 3D landmarks. We will use N 2D landmarks with known correspondences to 3D landmarks in all M frames to obtain a 3D reconstruction of the face. Sturm and Triggs provide a method to obtain a projective structure X and projective motion P by factorization of the projections x of all frames⁷⁵:

$$\lambda_{ij}x_{ij} = \hat{P}_i \cdot \hat{X}_j = P_i \mathcal{H} \cdot \mathcal{H}^{-1} X_j \quad (3.3.1)$$

Where $\hat{P}_i$ is a 3×4 projection matrix of frame i , $\hat{X}_j$ is a 4×1 homogeneous 3D vector of point j , x_{ij} is a homogeneous 2D vector of the projection i of landmark j and λ_{ij} is a scalar representing the projective depth of x_{ij} . If the projective depths λ_{ij} are known, the system of equations is of rank 4. The projective depths can be estimated using epipolar geometry on pairs of frames, see⁷⁵ for details. A rank 4 approximation of the system can be found using the Singular Value Decomposition (SVD) of the system. For details about the linear algebra or SVD see⁵⁴. Noise or imprecise measurements on the landmarks can lead to a system with a higher rank. The error minimized by Sturm and Triggs in equation 3.3.2 is based on both the estimated projective depths and the image coordinates, but has no geometric meaning, see⁹⁰:

$$\sum_{i=1}^N \sum_{j=1}^M \|\lambda_{ij}x_{ij} - \hat{P}_i \cdot \hat{X}_j\|_F^2 \quad (3.3.2)$$

Where x_{ij} are the image coordinates (which might include noise) and λ_{ij} the estimated projective depths corresponding to these points. The reconstruction we have now is a projective reconstruction of the cameras and shape. Before we can do any measurements of length or angles of the projective structure X , we need to find the 4×4 projective ambiguity $\mathcal{H}$, which is independent of the number of frames or the number of points, to update the projective space to Euclidean space, see Figure 3.1.

Figure 3.1: Three projective reconstructions of a cube with different ambiguities ($\mathcal{H}$).

The calibration can be achieved by adding extra information about the shape or internal parameters of the cameras, but since the intrinsics of the camera and the 3D shape of the face are unknown, there is no additional information available. A second method would be auto-calibration (self-calibration), in which case (almost) no additional information is needed for the

calibration. The auto-calibration estimates the shape and camera parameters simultaneously. Two available methods for auto-calibration are the absolute dual quadric as described in⁸¹ and Kruppa equations which can be found in³⁰. According to Hartley and Zisserman³¹: ‘*The application of the Kruppa equations to three or more views provides weaker constraints than those obtained by other methods such as the modulus constraint or the absolute dual quadric*’. Since our purpose is to use as many frames (data) as possible, we choose the absolute dual quadric method for auto-calibration.

3.4 AUTO-CALIBRATION

Auto-calibration is a method to estimate the internal camera parameters from uncalibrated CCTV footage. The object itself is used to perform the calibration. Auto-calibration is based on the dual image of the absolute conic (DIAC), which is fixed under similarity transformations, so the internal camera parameters can be estimated despite of the unknown external parameters. The goal of the auto-calibration is to locate the plane at infinity and the absolute conic ω . For a projective reconstruction where the first frame contains no rotation and translation, $\mathcal{H}$ can be expressed in terms of the calibration matrix K and the plane at infinity v ²⁴. In our case K is the same for all frames.

$$\mathcal{H} = \begin{bmatrix} K & \mathbf{0} \\ v^\top & \lambda \end{bmatrix} \quad (3.4.1)$$

Since we can’t determine the scale of the reconstruction without using additional input data, the scale factor λ can be chosen as $\lambda = 1$. The absolute dual quadric Q_∞^* encodes both K and v in one mathematical entity. The null space of Q_∞^* encodes the plane at infinity v . Without proof the following equation is given:

$$\omega^* = KK^\top = P_i Q_\infty^* P_i^\top \quad (3.4.2)$$

Equation 3.4.2 shows the relation between the projection of the absolute dual quadric and the calibration matrix K . Constraints on K can be transferred to the absolute dual quadric. The assumption of square pixels and a principle point close to the center of the camera are sufficient conditions to obtain linear equations for Q_∞^* , see³¹ for more details.

3.5 EXPERIMENTS

In our experiments we first obtain a projective reconstruction from the projections and compare the reconstruction to a known ground truth. Our goal is to see if the quality of the reconstruction and the method are suitable for the reconstruction of facial models. The second step is the auto-calibration, which is completely separated from the projective reconstruction. To express the quality of the projective factorization, we use the 2D RMS reprojection error. The 2D

reprojection error E_{2D} is defined as:

$$E_{2D} = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N \|x_{ij} - \hat{P}_i \cdot \hat{X}_j\|^2} \quad (3.5.1)$$

In all experiments the internal camera parameters K are fixed. The generated projections are comparable with realistic face images. Therefore the camera rotations vary between -40 and 40 degrees and the translations vary between -10 and 10 units. All camera parameters are randomly chosen within their respective bounds. All projections fit within an image of 400×600 pixels. The 3D ground truth point cloud contains uniformly distributed random points within a bounding box of $100 \times 100 \times 10$ units.

3.5.1 NUMBER OF FRAMES AND NUMBER OF POINTS

In the first experiment we try to find the minimal number of frames and points needed to obtain a projective reconstruction. In theory 4 landmarks in 3 frames are enough to obtain a projective reconstruction, but if the image coordinates contain noise, more points and/or frames are necessary to average out the noise. The projection of each point in each frame is known, but we add Gaussian noise with a standard deviation of $\sigma = 1$ to both the x- and y-coordinates. For each combination of number of points and number of frames the reprojection error E_{2D} is calculated two times: with respect to the projections with noise and with respect to the ground truth image points. The experiment was repeated 1 000 times, with independent instances of noise for every combination of points and frames to get more stable results. The curves show the average value over all repetitions.

Notice that in the left graph at least 50 points are needed to approximate the value of the expected asymptote $\sqrt{2}$. Since the most consistent reconstruction is the reconstruction of the noise-free image points, the reprojection error approximates the $\sqrt{2}$ value for Gaussian noise of $\sigma = 1$. More points still improve the results, but a number of points around 50 seems to be the lower bound on an approximation of the asymptote. According to the number of frames at least 30 frames are needed to stabilize the lines. Adding more frames doesn't seem to offer a drastic improvement of the results.

In the right graph the error seems to be decreasing for the first time for around 25 points. Using more points leads to an even faster decreasing function. Using more frames seems to have less effect on the error for a given number of points. So adding more points has a stronger effect than adding more frames and can even lead to a switch from an increasing to a decreasing error function. Finding the lower bounds allows us to make an approximation of the number of points and frames needed for a 3D facial reconstruction. We would like to find the lowest number of points possible, because in CCTV footage usually plenty frames are available, but determining more landmarks is difficult. We choose 25 points as an acceptable lower bound on the number of points, since it provides a decreasing function when more frames are added.

Figure 3.2: Variable number of point and frames. In the left graph the reprojection error is calculated with respect to the projections with noise and in the right graph with respect to the ground truth.

3.5.2 EFFECT OF THE NUMBER OF FRAMES ON 2D AND 3D ERROR

In the following experiment we use a fixed number of points and explore the effect of the number of frames in 2D and in 3D. The 3D error is calculated based on the 3D ground truth landmarks, see Equation 3.5.2.

$$E_{3D} = \operatorname{argmin}_{\mathcal{H}} \sqrt{\frac{1}{N} \sum_{j=1}^N \|X_j - \mathcal{H}\hat{X}_j\|^2} \quad (3.5.2)$$

Where X_j is a normalized known 3D ground truth point of the shape and $\hat{X}_j$ is a normalized

reconstructed 3D point. E_{3D} in Equation 3.5.2 can only be calculated for a known ground truth, but mostly finds the best auto-calibration possible independent of the auto-calibration method used.

Projective reconstructions are made from the projections of 25 points in a variable number of frames. Each subset of frames is randomly taken out of 100 000 frames. In each set zero-mean Gaussian noise with $\sigma = 1$ is added to both x- and y-dimension. The reconstructions are made with an increasing number of frames, ranging from 4 to 50 frames. The experiment is repeated 1 000 times with different subsets of frames. Outliers of more than 4 times the standard deviation were removed from the results. The results are shown in Figure 3.3.

Figure 3.3: 2D RMS and 3D RMS error of projective reconstruction with a variable number of frames.

The 2D reprojection error with respect to the ground truth image points shows only a slight

improvement of the quality between 10 and 50 frames, but the 3D error is reduced by 40% when comparing 10 to 40 frames (0.27 compared to 0.16). So more frames indeed improve the 3D reconstruction.

3.5.3 GAUSSIAN NOISE

To validate our choice that 25 points are sufficient for a projective reconstruction, we add Gaussian noise with a higher σ to the projections in the next experiment. The Gaussian noise varies within the range of 0 to 2.5 times the standard deviation for both the x- and y-coordinates. The noise experiment is repeated 1 000 times. The results are shown in Figure 3.4.

Figure 3.4: Variable Gaussian noise used for reconstruction.

For noise $\sigma \leq 1$ the reprojection error with respect to the image points with noise seems to be

linear with the Gaussian noise. Also the 3D error (for a perfect calibration) seems to be linear. If more noise is added, both errors rise above linear, see the blue line in Figure 3.4. This might be a problem of the depth estimation in the projective reconstruction method or more frames may be needed to get a consistent projective reconstruction.

3.5.4 AUTO-CALIBRATION

The auto-calibration method of Section 3.4 is also calculated for a variable number of frames and for increasing Gaussian noise. The 3D error of the auto-calibration is calculated by finding the best similarity transformation $\mathcal{H}$, because after the auto-calibration the 3D points and the cameras should no longer be affected by an affine or a projective transformation, see Equation 3.5.2. The auto-calibration uses a projective reconstruction as a starting point. The results are shown in Figure 3.5.

The left graph in Figure 3.5 shows that a minimum number of frames is required to perform the auto-calibration. Adding more frames improves the 3D result of the auto-calibration, but as can be seen in the right graph of Figure 3.5 adding noise ($\sigma > 1.5$) leads to an explosion of the 3D error function. So some work needs to be done on the robustness of the auto-calibration under Gaussian noise or more information needs to be added to the problem.

3.6 CONCLUSION

At least 25 image points in 30 frames are needed to obtain a precise perspective reconstruction. This is still valid for Gaussian noise with $\sigma \leq 1$. Accurate automatic or even manual labeling of the landmarks is needed to find such landmarks on a face. Though the number of frames certainly influences the quality of the reconstruction, the number of points seems more important to obtain an accurate reconstruction. Our experiment shows that there is no use adding additional frames for a small number of points. The number of points even has influence on the increasing or decreasing behavior of the 2D error function. If more noise ($\sigma \geq 1.5$) is involved, the projective reconstruction deteriorates and auto-calibration might fail. The auto-calibration method or conditions need to be improved to obtain better results. Also more work needs to be done to determine if additional available information of the camera or the background scene can improve the 3D reconstruction.

3.7 PAPER CLOSURE

This publication shows that it is possible to obtain accurate reconstructions of point clouds using around 25 landmarks. The number of frames is of less importance, around 30 frames is sufficient. After publishing this paper, we discovered that the auto-calibration results were too optimistic, because the incorrect positioning of the cameras was not taken into account in the 3D error calculation. In the following publications we use around 20 landmarks on a face, because this number seems to be realistic for manual annotation in forensic cases and still has a decent reconstruction error.

Figure 3.5: 3D error after auto-calibration for a variable number of frames and increasing Gaussian noise.

4

3D Reconstruction from Video Sequences: A Styrofoam Head

CHRIS VAN DAM
LUUK SPREEUWERS
RAYMOND VELDHUIS

Continuing our previous work, we take a next step towards the reconstruction of actual video sequences of faces. In the next two publications we still use a random point cloud, but also a styrofoam head and real video data. The styrofoam head we use, is a simplified version of a real head: The shape of the face is smoothed and no skin color is available. The landmarks on the styrofoam face are mimicked by pins with colored heads. We defined only 22 landmarks on a styrofoam face, because of two reasons: First of all the previous experiments showed that a number of landmarks between 20 and 25 is the minimum number of landmarks needed to find a reconstruction. Secondly, 22 landmarks is in our opinion a reasonable number of landmarks for low quality forensic video data. So, instead of taking a huge number of landmarks, we try to stay close to the forensic context, and use a realistic number of landmarks. In this chapter, we propose an improved version of the reconstruction algorithm. Most experiments in the first publication are performed on a random point cloud. We show only a few examples of reconstructions using a styrofoam head. This work was published in 2013 in the WIC Symposium. In the second publication, we extended the experiment using a styrofoam head and included statistics of the error in these reconstructions. The second publication, published in the proceedings of the BIOSIG conference in 2013, also includes an example of a reconstruction of the shape of a face from real video data.

4.1 ABSTRACT

In forensic comparison of facial video data, often only the best quality frontal face frames are selected, and hence much video data is ignored. To improve 2D facial comparison for law enforcement and forensic investigation, we introduce a model-free 3D shape reconstruction algorithm based on 2D landmarks. The algorithm uses around 20 landmarks on the face and combines the structure information of multiple frames. Model based 3D reconstruction methods, such as Morphable Models, reconstruct a 3D face shape model that is strongly biased towards the average face. Therefore, we don't use statistical face shape models in our model-free approach. The 3D landmark reconstruction algorithm simultaneously estimates the shape, pose and position of the face, based only on the fact that all images in the sequence are recorded using a single calibrated camera. The algorithm iteratively updates the reconstruction by including new frames, while maintaining the consistency of the reconstruction. We demonstrate the convergence properties of the method reflected in the 2D reprojection error and the 3D error with respect to a ground truth model. We show that the quality of the reconstruction depends on the noise on the landmarks. In a second experiment we show that the method can be used on realistic face shape data with a styrofoam head model.

4.2 INTRODUCTION

One of the unsolved issues in forensic comparison of facial data is the comparison with 'wild photo' or video data. Law enforcement services are constrained to work with the case material provided, and unlike researchers, they are not able to use recordings from a controlled environment. Among the most difficult problems of 'wild photo' materials are the non-frontal pose of faces and low resolution faces, because often material of overview cameras is used for facial comparison. Automatic face recognition software can only handle 2D facial data under a small pose angle. Faces with larger pose angles result in low recognition scores. At the moment there is no automatic face comparison software available for faces under pose. As a consequence often only the best quality frontal face frames are selected, and hence much video data is ignored.

In the future the pose problem may be solved using 3D video techniques, but the changes in both hardware and procedures will be slow. In the meantime the old 2D equipment will still be used in most cases. So for now law enforcement services are still in need of the 'tools' to compare non frontal faces. However, these 'tools' should treat the video data in such a way that no supplementary information is added to the video data. Reconstruction methods, such as Morphable Models⁹, reconstruct a 3D face shape model that is strongly biased towards the average face. Such reconstructions could lead to weird forensic cases, where innocent persons are suspected of a crime because of their average-looking faces. In our method we try to avoid this situation, caused by facial models.

The experiments we perform are comparable to a case where a single ATM-camera has caught the face of a suspect in several frames. Each frame contains a different view on the face of a suspect. We assume that the calibration parameters of the camera, such as focal length, principal

point and skew, are available. The recording is assumed to contain different views of a face without variation in facial expression. Our final goal is to reconstruct the face in 3D.

In this paper we introduce a model-free 3D shape reconstruction algorithm based on 2D landmarks, so no additional statistic face models or average face models will be used. We use around 20 landmarks on the face to estimate the shape of the face along with the pose and the position of the face for each view. In our first experiment we use simulated data to demonstrate the convergence properties of the method reflected in the 2D reprojection error and the 3D error with respect to a ground truth model. In the second experiment we show that the method can be used on realistic face shape data with a styrofoam head model. The reconstructed 3D models contain only shape information and no texture information.

4.3 BACKGROUND

Our problem, of the face of the suspect moving in front of a static camera, is equivalent to a problem where the camera is moving and the suspect is static. So for each view $i = 1..N$ we have to find the external camera parameters of that specific view. The static shape of the face can be described by $j = 1..M$ 3D landmarks. We will use M 2D landmarks with known correspondences to the 3D landmarks in all N views to obtain a 3D reconstruction of the landmarks on the face. Our camera is described by the pinhole camera model³¹, see Figure 4.1:

Figure 4.1: Pinhole camera model.

where a 3D point X is projected on the image plane in 2D point x .

Sturm and Triggs provide a method to obtain a projective structure $\hat{X}$ and projective camera $\hat{P}$ by factorization of the projections x of all views⁷⁵:

$$\lambda_{ij}x_{ij} = \hat{P}_i \cdot \hat{X}_j = P_i \mathcal{H} \cdot \mathcal{H}^{-1} X_j \quad (4.3.1)$$

where P_i is a 3×4 camera matrix, containing the calibration, rotation and translation parameters of view i , X_j is a 4×1 homogeneous 3D vector of point j , x_{ij} is a known homogeneous 2D vector of the projection i of landmark j and λ_{ij} is a scale factor representing the projective depth of x_{ij} . The projective transformation $\mathcal{H}$ is a 4×4 transformation matrix, which defines an unknown

projective ambiguity between the points X_j and camera matrices P_i . If the projective depths λ_{ij} are known, the system of equations is of rank 4. The projective depths can be estimated using epipolar geometry on pairs of views, see^{75,90} for details. Singular Value Decomposition (SVD) can factorize the projections, using a rank 4 approximation of the projections, in $\hat{P}_i$ and $\hat{X}_j$. For details and properties of SVD see, for example,⁵⁴. The main problem with this factorization method is that all landmarks should be visible in all views, but in realistic cases around 30% of the landmarks on the face are hidden due to the pose of the face. Another disadvantage of this method, for us, is the ambiguity $\mathcal{H}$, which makes it difficult to merge, for example, multiple partial solutions.

With the calibration parameters available, we can use a method described in²⁹ to estimate the rotation and translation parameters for one pair of views. This method provides four solutions for the rotation and translation parameters, see 4.3.2, but only one of these solutions consists of points in front of our camera:

$$\begin{aligned} \hat{P}_1 &= [UWV^\top \mid +u_3] & \hat{P}_2 &= [UW^\top V^\top \mid +u_3] \\ \hat{P}_3 &= [UWV^\top \mid -u_3] & \hat{P}_4 &= [UW^\top V^\top \mid -u_3] \end{aligned} \quad (4.3.2)$$

where the rotation matrix defined by U and V is the result of a singular value decomposition of the essential matrix, the translation u_3 is the last column of U and W is a matrix that mirrors one of the axis, see^{31,29}. This solution has 5 degrees of freedom, 3 for the rotation and only 2 for the translation, because the equation is determined up to an unknown scale. The rotation and translation parameters are extracted directly from the essential matrix of one pair of views. After estimating the rotation and translation, we can estimate the structure by linear triangulation of the pair of views³¹. This method reconstructs only all visible points in 2 views, but can be extended to more than 2 views. In our case we have a low number of landmarks, so the reconstruction based on two views gives a poor estimation of the shape. Therefore, we extend the algorithm using multiple views to overcome the problems of noise and the low number of landmarks. We introduce an algorithm that iteratively updates the reconstruction by including new views, while maintaining the consistency of the reconstruction for a low number of landmarks.

4.4 RECONSTRUCTION ALGORITHM OVERVIEW

STEP I: ESTIMATE TRANSLATION AND ROTATION PARAMETERS

Determine the relation between all pairs of views by calculating the essential matrices for each pair. This leads to $N \times N$ essential matrices, where N is the number of views. For each pair of views the relative rotation and translation (except for the scale) can be calculated from the essential matrices, see 4.3.2. The essential matrices have 5 degrees of freedom and therefore require at least an overlap of 5 landmarks for each pair of views in a noise free case. We will use 8 landmarks as a minimum to be able to handle some noise. The essential matrices are estimated with a robust MSAC method (M-estimator SAmple Consensus)⁷⁹. If less than 8 common landmarks are visible between a pair of views, the rotation and translation of this pair are marked as undefined.

STEP 2: ESTIMATE MISSING ROTATION PARAMETERS

Find a solution for missing rotations between a pair of views. Some of the missing rotations can be estimated based on views defined via other views. A new $N \times N \times N$ array is created with rotations defined via intermediate views. So for each reference view, all views via other views are calculated by multiplying the rotations of these views. If one of the three views is not defined, the rotation is undefined. If, for example, the essential matrix between view 1 and 4 was not available, but the essential matrices between view $1 \rightarrow 2$ and $2 \rightarrow 4$ were available, a new rotation can be calculated.

STEP 3: SELECTING ROBUST ROTATION PARAMETERS

The estimations of the rotations for each pair of views are made more robust, based on the largest cluster of points in the 3D angle space of all possible estimated rotations. The average of the largest cluster is used as a new estimation of the rotation between the pair of views. The cluster should at least contain 5 points to be considered as a robust estimation. Otherwise the original estimation will be used, if available. A new $N \times N$ array of rotations can now be calculated, which contains at least as many estimated rotations as before. Finally the translations of the original views are added to the corresponding rotations. For newly estimated rotations, the translations are set to zero.

STEP 4: SELECTION OF REFERENCE VIEW

We select view k with the lowest number of undefined points as reference view for our starting pair of views. We calculate the relative rotations and relative translations to view k for the total $N \times N$ set of rotation and translation estimations.

STEP 5: OPTIMIZATION OF TRANSLATION PARAMETERS

The translation estimations are optimized by minimizing the reprojection error of all camera estimations. The shape for this minimization is calculated based on view k and the current view. All estimations are relative to view k and can therefore be exchanged to obtain a total set of views with minimum reprojection errors.

STEP 6: SELECTION OF INITIAL VIEWS AND INITIAL STRUCTURE

For each available pair of views with view k as reference, an estimation of the structure is obtained. The pair with the most defined views and the lowest total reprojection error is chosen as the start for our iterative optimization. All views with the lowest reprojection errors based on the estimated structure are selected to be the initial estimations of the other views. We have now only N estimations of the rotation and the translation left. Our iterative algorithm starts with the two views used for estimation the structure.

STEP 7: ITERATIVE OPTIMIZATION OF ALL PARAMETERS AND STRUCTURE

In each round of this step, one new view is added to the selection set. A new estimation of the structure is made based on the current selection set. The reprojection error of all views in the selection is minimized by optimizing the rotation, the translation and the structure. The next candidate view for the selection set is chosen by checking the convergence behavior of each view. The one with the quickest convergence in 10 iterations will be chosen as the next candidate. This process continues until all views are added and optimized. To prevent overfitting, only 30 iterations are used for each round. Before adding a next view we do a quick optimization of the rotation and translation parameters of the views, which are not in the selection set, using the current structure estimation.

4.5 EXPERIMENTS

4.5.1 EXPERIMENT ON SIMULATED DATA: SETUP

The goal of the first experiment is to determine the influence of the number of views on the reconstruction, and to investigate the convergence properties of our algorithm. We create a random point cloud of 25 3D points and obtain a set of 100 projections of this point cloud with variation in rotation and translation. Later on, the point cloud is used as ground truth for the 3D error. The calibration information and a random selection of the projections are used in the reconstruction algorithm. We performed two experiments in which we add a different level of Gaussian zero-mean noise, with a standard deviation of 1.0 and 2.0 pixels respectively, to the projections. The noise is added independently to the x- and y-coordinates of the projections. Finally we use a random mask to hide 30% of the data to imitate the hidden landmarks on a face. We use our reconstruction algorithm to estimate the 3D point cloud and the external camera parameters of the views. The quality of the reconstruction will be determined by the 2D RMS reprojection error E_{2D} :

$$E_{2D} = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N \|x_{ij} - \hat{P}_i \cdot \hat{X}_j\|^2} \quad (4.5.1)$$

where index i represents a view $1..N$ and j represents a point $1..M$, $\hat{P}$ contains the external camera parameters of each view and $\hat{X}$ contains a collection of homogeneous 3D points. All projections that were not visible, were left out of the equation, so MN is defined as the total number of visible landmarks summed over all views. The parameter x_{ij} describes the known projections with Gaussian zero-mean noise.

After reconstruction the 3D RMS error E_{3D} between the reconstruction and the ground truth point cloud can be calculated with:

$$E_{3D} = \underset{\mathcal{H}}{\operatorname{argmin}} \sqrt{\frac{1}{M} \sum_{j=1}^M \|X_j - \mathcal{H}\hat{X}_j\|^2} \quad (4.5.2)$$

where $\mathcal{H}$ is a rigid 3D transformation which aligns the ground truth point cloud X with the reconstruction $\hat{X}$ and j is the index of a point. Note that the point cloud is not aligned to a frontal face position, because of the rigid transformation. The 3D RMS error is only calculated if all points were reconstructed by our algorithm, which might not be the case if one of the landmarks is hidden in all views. The experiment is repeated a 100 times with different instances of noise to see whether our algorithm is robust.

4.5.2 EXPERIMENT ON SIMULATED DATA: RESULTS

For Gaussian noise on both the x- and y-coordinates with a standard deviation of 1.0 pixels the graphs in Figure 4.2 show the expected behavior. The more views are added, the more robust the reconstruction is. If the shape is estimated perfectly, then we would expect the 2D reprojection error to converge to the level of noise added. The 2D reprojection error converges to an asymptote of $\sqrt{2} \approx 1.41$, which is the expected level of noise, see the left graph of Figure 4.2. Another observation we make is that a number of views above 30 has little influence on both the 2D and the 3D average error. The robustness of the algorithm is only slowly increasing for more than 30 views, see the right graph of Figure 4.2. So adding more than 30 views seems to have a small impact on both the quality and robustness of the algorithm. Therefore, also computation time and computation cost become interesting at that point.

If the level of Gaussian noise is doubled to a standard deviation of 2.0 pixels, the behavior is similar to the experiment with a standard deviation of 1.0 pixels. The asymptote here is $\sqrt{8} \approx 2.83$, see the left graph in Figure 4.3. Adding more views has less effect on the robustness of the reconstruction algorithm, but it still has a decreasing effect on the average reprojection error. When more views are added, the average 3D error also decreases slowly, though the robustness of the algorithm seems not to increase. For more than 35 views, the system shows even more variation in the 3D errors than for 35 views, see Figure 4.3. This could be explained by the fact that the more views are added, the higher the change for heavy outliers in the projections. From previous experiments we learned that the algorithm is known to be sensitive to heavy outliers. Since none of the selected views are skipped, outliers might severely decrease the result of the reconstruction.

These last graphs of this first experiment, see Figure 4.4, show the number of times the reconstruction has failed. The reconstruction is assumed to be failed, if the reprojection error is above 5.0 pixels. For the experiment with Gaussian noise with a standard deviation of 1.0 pixels and more than 30 views, the algorithm converges to a solution in about 99% of the cases. In the case of a standard deviation of 2.0 pixels the algorithm only converges in 75% of the cases. So the algorithm seems stable for Gaussian noise with a standard deviation of 1.0 pixels, but becomes less stable for Gaussian noise with a standard deviation of 2.0 pixels or above.

4.5.3 EXPERIMENT ON STYROFOAM HEAD: SETUP

The goal of the second experiment is to determine if the algorithm is capable of working with manually labeled face data. We acquired a 3D model of a styrofoam head with 22 markers located on the face. We created 50 renderings of the model with different rotation and translation parameters. The calibration of the camera is known. The landmarks of the projections were labeled manually. In contrast to the previous experiment, no noise was added to the projections,

Figure 4.2: 2D and 3D error of the reconstruction with Gaussian noise with a standard deviation of 1.0 pixels.

leaving us with only the noise of the manual landmarking. The reconstruction is based on the calibration data and all of the renderings. Similar to the previous experiment, the quality of the reconstruction will be expressed by the 2D RMS error and the 3D RMS error. The ground truth 3D points are manually labeled in the 3D data of the styrofoam head.

Figure 4.3: 2D and 3D error of the reconstruction with Gaussian noise with a standard deviation of 2.0 pixels.

4.5.4 EXPERIMENT ON STYROFOAM HEAD: RESULTS

The 2D reprojection error of the reconstructed styrofoam head, see the right picture in Figure 4.5, is 1.96 pixels, which is probably the accuracy of the manual landmarking of the 2D data set. This noise level is similar to a 1.4 pixels error in both x- and y-coordinates. A rough estimation gives us a head size of 300 mm and the size of the head in the frames is around 500 pixels. So each pixel represents 0.6 mm. Our method is able to estimate the landmarks with $(1.96 \times 0.6 =)$ 1.2 mm precision on average. The 3D error is 1.10, which is around 0.6% of the size of the head.

Figure 4.4: Number of failures out of 100 times for both types of noise. Left → Gaussian noise with a standard deviation of 1.0 pixels and right → Gaussian noise with a standard deviation of 2.0 pixels.

The results are in line with the results of the first experiment on simulated data. Our algorithm has similar convergence properties and errors, and can therefore be applied on manually labeled realistic face data.

Figure 4.5: Left: Orthogonal view of the 3D styrofoam head model. Middle: One of the rendered 2D views of the model. Right: The reconstructed 3D landmarks with added edges for visibility.

4.6 CONCLUSION

The experiment on the simulated point cloud shows that the quality of our reconstruction depends on the level of noise in the projections. For small level of noise, around 1.0 pixels, the convergence and robustness of the algorithm seem sufficient. For a larger level of noise the system might become unstable, and even not converge to a useful solution. For Gaussian noise with a standard deviation of 2.0 pixels the algorithm only converges in 75% of the cases. The minimum number of views needed to get sufficient quality for the reconstruction is around 30 views. More views can improve the reconstruction, but this will only give a small improvement. In the second experiment, we showed that manual landmarking leads to an error comparable to a Gaussian noise with a standard deviation of 1.4 pixels. This value converted to an accuracy of 1.2 mm for an average face, indicates that the algorithm is capable of working with manually labeled realistic face data.

4.6.1 FUTURE WORK

Currently, we are working on more extensive experiments on the styrofoam data. The next step will contain experiments on real video data. In future work we will include the texture information in the reconstruction to get a full 3D model of the face. We need to find a method to blend the textures of all the visible views. Another improvement would be to get a more dense reconstruction of the face, based on our estimated reconstruction. This can be done by searching for image similarities in local regions around the known landmarks. A more dense reconstruction provides a more accurate description of the facial structure.

4.7 ABSTRACT

In forensic comparison of facial video data, often only the best quality frontal face frames are selected, and hence potentially useful video data is ignored. To improve 2D facial comparison for law enforcement and forensic investigation, we introduce a model-free 3D shape reconstruction algorithm based on 2D landmarks. The algorithm uses around 20 landmarks on the face and combines the structure information of multiple frames. Model based 3D reconstruction methods, such as Morphable Models, reconstruct a 3D face shape model that is strongly biased towards the average face. Therefore, we don't use statistical face shape models in our model-free approach. The 3D landmark reconstruction algorithm simultaneously estimates the shape, pose and position of the face, based only on the fact that all images in the sequence are recorded using a single calibrated camera. The algorithm iteratively updates the reconstruction by including new frames, while maintaining the consistency of the reconstruction. We demonstrate the convergence properties of the method reflected in the 2D reprojection error and the 3D error with respect to a ground truth model. We show that the quality of the reconstruction depends on the level of noise in the landmarks. In follow-up experiments we show that our method is able to reconstruct the 3D structure of a face, using a styrofoam head and real video data. The results of the real face data show the same behavior as the results of the simulated data, which indicates that our method is capable of reconstructing real facial structures, depending on the noise of the landmarks.

4.8 INTRODUCTION

One of the unsolved issues in forensic comparison of facial data is the comparison with ‘wild’ photo or video data. Law enforcement services are constrained to work with the case material provided, and unlike researchers, they are not able to use recordings from a controlled environment. Among the most difficult problems of ‘wild’ photo materials are the non-frontal poses of faces and low resolution facial images, because often material of overview cameras is used for facial comparison. Automatic face recognition software can only handle 2D facial data under a small pose angle. At the moment the accuracy of automatic face comparison algorithms degrades quickly for faces under large pose. As a consequence often only the best quality frontal face frames are selected, and hence much video data is ignored. Law enforcement services are still in search of the ‘tools’ to compare non frontal faces. However, these ‘tools’ should treat the video data in such a way that no supplementary information is added to the video data. Reconstruction methods, such as Morphable Models⁹, reconstruct a 3D face shape model that is strongly biased towards the average face. Such reconstructions could lead to unacceptable forensic conclusions. In the proposed method we try to avoid this situation caused by facial models.

In this paper we introduce a model-free 3D shape reconstruction algorithm based on 2D landmarks, so no additional statistical face models or average face models will be used. We assume that the calibration parameters of the camera, such as focal length, principal point and skew, are available. Any recording is assumed to contain a subset of frames with different views of a face without variation in facial expression. Our final goal is to reconstruct the face in 3D. We use

around 20 landmarks on the face to estimate the shape of the face together with the pose and the position of the face for each view. We present three different experiments. In our first experiment we use simulated data to demonstrate the convergence properties of the method reflected in the 2D reprojection error and the 3D error with respect to a ground truth model. In the second experiment we continue our work in⁸⁴ and we explore the strength of our method more extensively on realistic face shape data with a styrofoam head model. In our last experiment we use real video sequences for our reconstruction. Note that our reconstructed 3D models only contain shape information and no texture information. This paper continues with section 4.9 where we give a background on the methods and notations used in this paper. In section 4.10 we introduce and explain our proposed algorithm. In section 4.11 we show the performance of our algorithm in several experiments. Then we end up with the conclusion in section 4.12.

4.9 BACKGROUND

Our problem, in which the face of the suspect is moving in front of a static camera, is equivalent to a problem where the camera is moving and the suspect is static. So for each view $i = 1..N$ we have to find the external camera parameters of that specific view. The static shape of the face can be described by $j = 1..M$ 3D landmarks. We will use M 2D landmarks with known correspondences to the 3D landmarks in all N views to obtain a 3D reconstruction of the landmarks on the face. Our camera is described by the pinhole camera model³¹, where a 3D point Q is projected on the image plane in 2D point q . The point projection equation is usually written as $q = P \cdot Q$, where P contains both the calibration parameters of the camera and the rotation and translation of a view.

We prefer a method in which we can add additional views to the current solution to improve the reconstruction. To be able to find such a method, we should search for a method that starts with one pair of views and then provides an iterative solution or a solution that merges groups of views. The method described in²⁹ is able to estimate the rotation and translation parameters for one pair of views. This method expresses the relation between calibrated views in the essential matrix. The essential matrix can be estimated from corresponding landmarks in two views using a robust MSAC method (M-estimator SAMple Consensus) method⁷⁹. Once we determined the relation between two views, the relative rotations and translation parameters can be estimated for both views. This method provides four solutions for the rotation and translation parameters, see Equation 4.9.1, but only one of these solutions is posing the points in front of the camera:

$$\begin{aligned} \hat{P}_1 &= [UWV^T \mid +u_3] & \hat{P}_2 &= [UW^T V^T \mid +u_3] \\ \hat{P}_3 &= [UWV^T \mid -u_3] & \hat{P}_4 &= [UW^T V^T \mid -u_3] \end{aligned} \quad (4.9.1)$$

where the rotation matrix defined by U , W and V is based on the result of a singular value decomposition of the essential matrix. The matrix W is a matrix that mirrors one of the axes. The translation u_3 is the last column of U , see^{31,29}. This solution has 5 degrees of freedom, 3 for the rotation and only 2 for the translation, because the equation is determined up to an unknown scale. The rotation and translation parameters are extracted directly from the essential

matrix of one pair of views. Then, we can estimate the structure by linear triangulation of one pair of views³¹:

$$A = \begin{bmatrix} xp^3{}^\top - p^1{}^\top \\ yp^3{}^\top - p^2{}^\top \\ x'p'^3{}^\top - p'^1{}^\top \\ y'p'^3{}^\top - p'^2{}^\top \end{bmatrix} \quad (4.9.2)$$

where $p^{i\top}$ are the rows of P in the first view and x, y are the x - and y -values of the projection of point Q in the first view. The other parameters are the corresponding values of the second view. The point Q can be found by solving $AQ = 0$. This method reconstructs only the visible points in one pair of views. The method can be extended to more than 2 views by including more equations from additional views in A . In our case we have a low number of landmarks, so the reconstruction based on two views gives a poor estimation of the shape. Therefore, we extend the algorithm using multiple views to overcome the problems of noise and the low number of landmarks. We introduce an algorithm that iteratively updates the reconstruction by including new views, while maintaining the consistency of the reconstruction for a low number of landmarks. The quality of the reconstruction can be determined by the 2D RMS reprojection error E_{2D} :

$$E_{2D} = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N \|q_{ij} - \hat{P}_i \cdot \hat{Q}_j\|^2} \quad (4.9.3)$$

where index i represents a view $1..N$ and j represents a point $1..M$, $\hat{P}$ contains the external camera parameters of each view and $\hat{Q}$ contains a collection of homogeneous 3D points. The homogeneous 2D vector q_{ij} represents the known projections including the noise on the landmarks.

4.10 RECONSTRUCTION ALGORITHM

In this section we describe the proposed algorithm for the reconstruction of the structure of the face based on 2D projections. In short the algorithm finds an initial pair of views with a low reprojection error. Based on this pair of views we obtain a linear estimation of the structure. Then we start an iterative procedure in which we add one new view in every step of the procedure. After adding the new view, the current selection of views and the current structure estimation are optimized. The result of the reconstruction algorithm is an estimation of the 3D positions of the landmarks and an estimation of the rotation and translation parameters of each view.

The best initial estimate for the structure is found by calculating the reprojection error for every possible pair of views in the dataset and to select the pair with the minimum reprojection error. To calculate the reprojection error we need to know the rotation and translation parameters of each view. These values (except for the scale) can be extracted from the Essential Matrix, see Equation 4.9.1. The essential matrix can be estimated, in turn, from the projections using a robust MSAC method (M-estimator SAmple Consensus)⁷⁹. Knowing the rotation and translation parameters of a pair of views, allows us to estimate the structure for this pair of views. Based on this structure we can calculate the reprojection error for this pair of views. However,

also the reprojection errors of the other views are important for consistency during the optimization. So, to find the best pair of views we choose a reference view and calculate all rotations and translation relative to the reference view. Then we calculate the reprojection error of the total set of views for every view as reference. A second criterion for the selection of the best pair of views is the number of landmarks that could be reconstructed, because not only the reprojection error is important, but also the number of visible corresponding landmarks in the initial pair of views. Our selection criterion is now to find the pair of views with the maximum number of corresponding landmarks in two views and a minimal reprojection error for the total set of views. We choose to obtain the subset of 25% of the solutions containing the most reconstructed points over all views. From this subset we select the pair with the lowest reprojection error over all views. This solution provides us a solution that is sufficient for initialization of our iterative optimization. We calculate an initial linear estimation of the structure based on the selected pair of views.

In the optimization step one new view is added in each iteration to keep all views in our current estimation consistent with the estimated structure. The selection of the new candidate view is based on the convergence behavior of the candidate view. The view with the lowest reprojection error after 10 optimization iterations, is chosen as the next view. This candidate selection is necessary to prevent the algorithm from failing in the first few iterations. Based on the new selection of views, a linear estimation of the structure is obtained, see Equation 4.9.2. Then the reprojection error of both the rotation and translation parameters and the structure are minimized using the Levenberg-Marquardt algorithm. To prevent overfitting, we used only 30 Levenberg-Marquardt iterations for each optimization step, which performs properly for the minimization. Finally, the rotation and translation parameters of the views that were not in the selection set are optimized to maintain the consistency of the total set of views. The iterative optimization procedure continues until all views are added and optimized.

4.11 EXPERIMENTS

The goal of the first experiment on simulated data is to determine the influence of the number of views on the reconstruction, and to investigate the convergence properties of our algorithm. We create a random point cloud of 25 3D points and obtain a set of 100 projections of this point cloud with variation in rotation and translation. The calibration information and a random selection of the projections are used in the reconstruction algorithm. We performed two experiments in which we added a different level of Gaussian zero-mean noise to the projections, with a standard deviation of 1.0 and 2.0 pixels respectively. The size of the face in each frame is around 250-350 pixels. The noise is added independently to the x- and y-coordinates of the projections. Finally we used a random mask to hide 30% of the data to imitate the hidden landmarks on a face. We use our reconstruction algorithm to estimate the 3D structure. The quality of the reconstruction will be determined by the 2D RMS reprojection error E_{2D} , see Equation 4.9.3. All landmarks that were not visible, were left out of the equation, so MN is defined as the total number of visible landmarks summed over all views. After reconstruction the 3D RMS error

E_{3D} between the reconstruction and the ground truth point cloud can be calculated with:

$$E_{3D} = \operatorname{argmin}_{\mathcal{H}} \sqrt{\frac{1}{M} \sum_{j=1}^M \|Q_j - \mathcal{H}\hat{Q}_j\|^2} \quad (4.11.1)$$

where $\mathcal{H}$ is a rigid 3D transformation which aligns the ground truth point cloud Q with the reconstruction $\hat{Q}$ and j is the index of a point. The experiment is repeated 100 times with different instances of noise to investigate the robustness of the algorithm.

The graphs in Figure 4.6 show the expected behavior for Gaussian noise with a standard deviation of 1.0 pixels. The more views are added, the more robust the reconstruction is. If the shape is estimated perfectly, then we would expect the 2D reprojection error to converge to the level of noise added. The 2D reprojection error converges to an asymptote of $\sqrt{2} \approx 1.41$, which is the expected level of noise, see the left graph of Figure 4.6. Another observation we make is that the number of views above 30 has little influence on both the 2D and the 3D average error. The robustness of the algorithm is only slowly increasing for more than 30 views, see the right graph of Figure 4.6. So adding more than 30 views seems to have only a small impact on both the quality and robustness of the algorithm.

If the level of Gaussian noise is doubled to a standard deviation of 2.0 pixels, the behavior is similar to the previous experiment. The asymptote here is $\sqrt{8} \approx 2.83$, see the left graph in Figure 4.7. Adding more views has less effect on the robustness of the reconstruction algorithm, but it still has a decreasing effect on the average reprojection error. When more views are added, the average 3D error also decreases slowly, though the robustness of the algorithm seems not to increase. For more than 35 views, the system shows even more variation in the 3D errors than for 35 views, see Figure 4.7. This can be explained by the fact that the more views are added, the higher the change for heavy outliers in the projections. Since none of the selected views are skipped, outliers might severely decrease the result of the reconstruction. The reconstruction is assumed to be failed, if the reprojection error is above 5.0 pixels. For the experiment with Gaussian noise with a standard deviation of 1.0 pixels and more than 30 views, the algorithm converges to a solution in about 99% of the cases. In the case of a standard deviation of 2.0 pixels, the algorithm only converges in 75% of the cases. So the algorithm seems stable for Gaussian noise with a standard deviation of 1.0 pixels, but becomes less stable for Gaussian noise with a standard deviation of 2.0 pixels or above.

The goal of the second experiment with the styrofoam head is to determine whether the algorithm is capable of working with manually labeled face data. We acquired a 3D model of a styrofoam head with 22 colored pins located on the face. An orthogonal view of the styrofoam model can be seen in the left image in Figure 4.8. We choose a virtual camera and we extract the calibration data from this camera. We created 51 renderings of the model with different rotation and translation parameters, see Figure 4.8. All visible landmarks are labeled manually in all renderings. In contrast to the previous experiment, no noise was added to the projections, leaving us with only the noise of the manual landmarking. The reconstruction is based on the calibration data and subsets of the renderings. The 3D points of the ground truth model are also manually labeled on the 3D model of the styrofoam head, which, in contrast to the previous experiment, could influence the 3D error. The experiment is repeated 100 times for each number of views to determine the robustness of the algorithm.

Figure 4.6: 2D and 3D error of the reconstructions using noise with a standard deviation of 1.0 pixels.

The second experiment shows the same behavior as the experiment with Gaussian noise with a standard deviation of 1.0 pixels. Adding more views increases the quality of the 3D reconstruction, but for more than 40 views, in this case, the gain is very low for both the 2D and 3D error. The asymptote for the 2D error is around 2.0 pixels, which is somewhere between the results of the first two experiments. This noise level is similar to a $\sqrt{2} \approx 1.41$ pixels error in both x- and y-coordinates, which is probably the accuracy of the manual landmarking of the 2D dataset. A rough estimation gives us a head size of 300 mm and the size of the head in the frames is around 500 pixels. So each pixel represents 0.6 mm. Our method is able to estimate the landmarks with

Figure 4.7: 2D and 3D error of the reconstructions using noise with a standard deviation of 2.0 pixels.

$(2.16 \cdot 0.6 \Rightarrow) 1.3$ mm precision on average. The average 3D error is 1.22, which is around 0.7% of the size of the head. The results are in line with the results of the first experiment on simulated data. This second experiment shows that our algorithm has similar convergence properties and errors to the experiment on the simulated data, and can therefore be applied on manually labeled realistic face data.

In this last experiment we show that our algorithm can handle real video data using a calibrated camera. We acquired 100 frames of several volunteers in which they slowly moved and rotated

Figure 4.8: Left: Orthogonal view of the 3D styrofoam head model. Middle: One of the rendered 2D views of the model. Right: The reconstructed 3D landmarks with added edges for visibility.

their heads in front of a camera. We annotated 20 landmarks in each frame in a semi-automatic manner. Finally we calibrated our camera with 20 frames of a planar calibration board, which provided us the camera calibration data. In the next experiment we use a selection of 50 frames to reconstruct the structure of the face. Since we don't have 3D ground truth data of our landmarks, we will only use the 2D reprojection error and visual inspection to express the quality of the reconstruction. We ran the experiment two times, with different subsets of views: one using the 50 even frames and another using the 50 odd frames of the first volunteer.

The 2D reprojection error for the even set was 2.13 and the reprojection error of the odd set was 2.54, where the size of the frames is similar to the styrofoam experiment. These results are completely in line with the results of the styrofoam experiment, see the left graph in Figure 4.9. There is a small variation in the 2D error, but nevertheless the variation seems acceptable compared to the previous results. Visual inspection of the 3D structure shows that both 3D structures are close to each other, see Figure 4.10. So even with real video data, including calibration, and landmarking, our algorithm is able to reconstruct the position of the 3D landmarks of the face.

4.12 CONCLUSION AND FUTURE WORK

The experiment on the simulated point cloud shows that the quality of our reconstruction depends on the level of noise in the projections. For a small level of noise, around 1.0 pixels, the convergence and robustness of the algorithm seem sufficient. For a larger level of noise the system might become unstable, and even not converge to a useful solution. For Gaussian noise with a standard deviation of 2.0 pixels the algorithm only converges in 75% of the cases. The minimum number of views needed to get sufficient quality for the reconstruction is around 30 views. More views can improve the reconstruction, but this will only give a small improvement. In the second experiment, we showed that manual landmarking leads to an error comparable to a Gaussian noise with a standard deviation of 1.4 pixels. The results of the styrofoam experiment were in line with the simulated reconstructions with Gaussian noise. The third experiment with real video data shows results similar to the styrofoam experiment. The visual inspection of the 3D structure and the 2D reprojection errors indicate that the algorithm is capable of reconstructing real facial structures. In future work we will include the texture information in the reconstruction to get a full 3D model of the face. The full reconstruction allows us to perform facial recognition experiments on 2D faces under pose.

Figure 4.9: 2D and 3D error of the styrofoam face reconstructions.

4.13 PAPER CLOSURE

In these two publications the proposed reconstruction algorithm is tested with more realistic data. The distribution and the number of the landmarks are chosen in line with the minimum requirements for face reconstruction and real forensic video data. The landmarks in the frames are manually annotated, which brings us closer to a real forensic scenario. Therefore we did not add any noise to the landmarks like we did in our previous experiments. The distribution of the noise is no longer Gaussian, but is now depending on the manual landmarking. For faces

Figure 4.10: Left: Two 3D reconstructions of the first volunteer based on different subsets of views. Right: Aligned 3D models of the two reconstruction.

under large pose it is sometimes hard to decide if a landmark is still visible or not. In the case of the pins with the colored heads this is a relatively easy task, but for real facial data this decision becomes more challenging. Based on the initial results in the second publication⁸³ we conclude that the stability and robustness of the proposed algorithm is sufficient to handle realistic face data.

5

3D Reconstruction from Video Sequences: Real Face Data

CHRIS VAN DAM
LUUK SPREEUWERS
RAYMOND VELDHUIS

We arrived at the point where we switch to real facial video data. In this chapter we introduce a complete reconstruction method including both shape and texture reconstruction. The data we acquired consist of video data of persons that move their face in front of a camera. In Figure 5.1 the setup for the data acquisition is shown. We used a special triggering device for these recordings to semi-automate the landmarking procedure. The triggering device will trigger the camera, but also change the illumination for each frame. The persons were landmarked by hand with an UV-marker. The landmarks are not visible under normal illumination. However, if we use UV-illumination the landmarks become visible, mainly in the blue channel of the images, see Figure 5.2. We used a triggering device to acquire consecutive images with UV-illumination and images with normal illumination. The landmarks are manually annotated in the UV-images. This gives the opportunity to estimate the position of the landmarks in the normal images. By interpolation of the landmark positions of the UV-image before and after the normal image, we obtain an estimate of the landmarks in the normal images.

Although this speeds up the process of landmarking, new noise is introduced due to the interpolation. Both the pose and the position of the face could be different, which causes a deviation from the real landmark position. We tried to minimize this error by instructing the persons to slowly move their heads in front of the camera.

Figure 5.1: Setup with mounted lights and camera in the center.

Figure 5.2: Image with normal illumination, UV-illumination and extracted blue channel from UV image.

During the experiments with real image data, we noticed that the number of landmarks we use, is too low to estimate the camera calibration parameters and shape unambiguously, see Figure 5.3, where multiple renderings of the same 3D shape are made using different parameters for the focal distance. To resolve this ambiguity, we assume that the camera calibration matrix is known or can be estimated.

In this paper we obtain a reconstruction in multiple steps. First we reconstruct a 3D shape model of the landmarks only. Then we define patches between the 3D landmarks and reconstruct the

Figure 5.3: Multiple renderings of the same 3D model using different parameters for the focal point. Renderings have a field of view from left to right of 0, 30, 60 and 90 degrees.

texture of these patches by combining the texture information from the 2D frames.

MODEL-FREE FACE RECONSTRUCTION FROM IMAGE SEQUENCES FOR FRONTAL FORENSIC FACIAL COMPARISON⁸⁶

5.1 ABSTRACT

Correctly handling surveillance video data is a major challenge in forensic face comparison. Often only the frames with frontal faces are used, because other frames cannot be processed properly. Although 3D model-based reconstruction methods do exist, they provide reconstructions with a bias towards the average face, which is unacceptable for face comparison in a forensic context. We investigate the possibilities of a model-free facial 3D reconstruction method which uses information from multiple non-frontal frames. The proposed 3D reconstruction method does not introduce supplementary data or statistical face models that could compromise the evidential value of the original video data. This qualifies the proposed reconstruction method for application in a forensic context. We introduce a semi-automatic facial reconstruction method, based on 20 landmarks on the face. The method consists of two stages in which we first reconstruct the shape of the face and then, in the second stage, reconstruct the texture from multiple images. We are able to reconstruct frontal views from image sequences without the use of frontal face images. Recognition experiments with frontal faces show that the reconstructions substantially surpass the face comparison scores of the 2D frames.

5.2 INTRODUCTION

In forensic face comparison there are still many unsolved issues, such as varying illumination conditions, low resolution data and non-frontal pose. These issues arise due to the uncontrolled nature of the photo or video data, which are more difficult to handle than controlled lab recordings. Since forensic researchers have to deal with the provided case data, they need to overcome these problems and find solutions in order to be able to use the data as forensic evidence. Most of the facial comparison cases are still done manually by forensic examiners on 2D photographic data. Although automatic 2D comparison methods are available for facial comparison, the ac-

curacy of automatic face comparison algorithms degrades quickly for faces under large pose⁵⁶. As a consequence, often only the best quality frontal face frames are selected, and hence a large part of the video data is ignored in forensic face comparison. Another important concern is the introduction of supplementary data that compromises the evidential value of the original photo or video data. For instance, model based 3D reconstruction methods, such as Morphable Models⁹, reconstruct a visually pleasing 3D face shape model which is biased towards the average face⁸⁸. The result is that 3D reconstructions of dissimilar faces tend to become more similar, which is undesirable behavior. In a forensic context the original data should not be contaminated with data from statistical face models or average face data. To provide a transparent way of handling the forensic evidence, statistical face models, in our opinion, should not be used in forensic face reconstructions.

In this article we introduce a model-free landmark-based face reconstruction method based on image sequences. A schematic overview of the proposed method can be found in Fig 5.4. Our goal is to use model-free 3D facial reconstruction methods to improve or complement the current 2D facial comparison process, resulting in a representative evidential value of the forensic case data. We show that without using prior information in the form of facial models, we can still realize a significant improvement in face recognition performance, compared to the individual images. Although other methods might claim a better reconstruction performance, these methods often exploit statistical knowledge of faces and are therefore, by definition, biased, which is unacceptable in a forensic context.

Figure 5.4: Reconstruction method overview. The blue boxes describe the input and output data, the green boxes describe the appropriate actions to obtain the frontal image and the orange box indicates manual interaction.

Forensic surveillance data may consist of relatively low resolution video data which is prone to errors for automatic landmarking algorithms as described in^{76,94,3}. Since the current process of preparing the data and checking the validity of the results by a forensic examiner is time consuming, we assume that the forensic examiners spend their time on manually annotating the landmarks. This gives us the opportunity to focus on other aspects of the reconstruction process, for example, the handling of time-lapse recordings, which still can be commonly found in forensic cases. The time-lapse recordings prevent us from using information of prior frames to determine or track the landmarks in image sequences. Based on the forensic setting we assume relatively low resolution video data and limit the number of landmarks to 20, which is a realistic number for medium to low quality faces. The number of visible landmarks in each image de-

depends on the view and is however much lower. We propose a model-free reconstruction method to prevent contamination of the original 2D data. After the reconstruction of the 3D shape, we apply illumination correction on the frames, based on the reconstructed shape. Subsequently, the texture information of the frames is merged into one textured 3D face model. Finally, in a 2D face comparison experiment we compare the scores of the original 2D frames with the scores of the reconstructed frontal views that are based on multiple non-frontal views.

We focus on an example case of fraud at an ATM machine. There might have been a transaction with a stolen bank card or some skimming devices might be placed at the ATM. The internal camera of the ATM recorded the incident and we have relatively good quality image sequences of a face. We assume that the person was not in disguise or wearing a fake mask. The sequence of images contains different views of the face of a person. The police arrested a suspect and needs evidence whether or not the suspect is the same person as the one that is visible in the ATM recordings. We propose a 3D reconstruction method to obtain a frontal view of the face of a person, which can be compared with a reference photo of the suspect.

The remainder of this paper is structured as follows: In the related work section we provide an overview of the literature on 3D face reconstruction. Subsequently we describe the proposed shape reconstruction method, the illumination correction and texture reconstruction. Next, we perform a face comparison experiment to illustrate the strength of the presented reconstruction algorithm. In the last section we conclude this paper.

5.3 RELATED WORK

Based on the work of pioneers, such as Triggs, Hartley, Zisserman and Pollefeys, various 3D reconstruction methods are described in literature. We provide an overview of *Structure from Motion* methods, that use the *Factorization Method*, see the work of Sturm and Triggs⁷⁵, or *Bundle Adjustment*, see the description by Hartley and Zisserman³¹, to reconstruct a 3D facial shape. All of the following *Structure from Motion* methods incorporate a generic face model or a statistical face model to find or improve the 3D reconstruction, see^{22,71,66,41,59,67,20,51,55,37}. These methods are by design unfit for forensic face comparison, since they are biased towards the generic or statistical face model. Chowdhury¹⁶ removed local outliers in the reconstruction by using a generic face model afterwards. Although such reconstructions have less tendency to look similar to the average face model, they still introduce supplemental data and therefore a bias towards the generic face model. Also *Nonrigid Structure from Motion* methods are described in literature^{23,80}, but they only seem to present results on motion capture data or rendered 3D model data. Synthetic data are very precise and differ on several aspects from real data, such as constant brightness, rigidity of the shape and visibility of the data. Also the proposed tracking methods cannot be used on forensic time-lapse recordings. Other reconstruction methods mostly based on the work of Blanz and Vetter⁹ use a *Morphable Model*, a PCA-based statistical model of the texture and shape of a face, to reconstruct faces. These methods certainly introduce supplementary data, since the reconstruction is a combination of the faces from the *Morphable Model*. See for example the work of Kemelmacher-Shlizerman on face reconstruction in the wild^{43,96} and³⁹, where a *Morphable Model* is incorporated in all of these reconstruction methods, resulting in a biased reconstruction. Another approach involving a *Learning Model* for the optimal warping of non-frontal views to frontal views, generates a frontal view, but does not combine multiple views³⁴ and is therefore not suitable in a forensic context. Hamsici²⁸ introduces a face

reconstruction method that learns a 2D to 3D function based on a studio recorded face model. However, this function is not general and is therefore biased towards the original model data. In^{32, 21} and⁶⁵ face data is mapped onto a generic 3D model, which again causes a bias and is not usable in a forensic context. The recent work of Suwajanakorn et al.⁷⁷ shows a method based on optical flow and shape from shading which is able to reconstruct even small details, but this method cannot be applied on time-lapse recordings and the authors state that they seek to produce a reasonably convincing model, instead of a physically correct or non-biased model. None of the references in literature provides a complete bias-free reconstruction method that can be used in a forensic context.

5.3.1 PREVIOUS WORK

In previous work⁸³ we showed the possibility of face shape reconstruction using only 20 landmarks on simulated data. The simulated data originated from a styrofoam 3D head model. We used 2D renderings and manual landmarks to illustrate the strength of our shape reconstruction algorithm. In this paper we introduce a first bias-free reconstruction method which supports forensic data. The proposed 3D face reconstruction method consists of two parts. In the first part we obtain an estimate of the 3D shape and the parameters of the views. Secondly, we perform illumination correction on the 2D frames based on the 3D shape and the viewing directions, which enables us to merge the texture information from multiple views. The result is a coarse 3D model with texture information. We perform recognition experiments on frontal reconstructions from image sequences recorded with a real camera.

5.3.2 DATASET

The problem we address in this paper is the reconstruction of a frontal face image from multiple frames recorded by a single camera, where the camera and illumination remain fixed and the face is moving. The result is that in every frame the face has a different pose and different illumination. This is different from current benchmark datasets such as CMU Multi-PIE and FRGC, where the data is taken at the same moment in time with multiple cameras. These datasets address a simplified problem, because of the constant illumination and the lack of deformations of the face over time. Secondly, as shown in our previous work⁸³, we need a minimum of 30 views of a face to obtain a 3D reconstruction, which is more than the number of views available in for example CMU Multi-PIE, which only provides up to 15 views of each person. Although there are multiple recording sessions, we are unable to use the data from multiple sessions without compromising the link with the ATM scenario, where only a single recording is captured. The current benchmark datasets are not applicable for reconstruction experiments in a forensic context. Therefore, we collected our own dataset consisting of face recordings of 48 people. Each recording contains 100 frames with different views of the face. In each frame the rotation and position of the face are different. The faces have mainly neutral looks, although some persons might show minor expressions. Fig 5.5 shows some examples of multiple persons in the dataset. From each person a reference 3D model was recorded with a 3dMD structured light scanner¹. The 3D model can be used for comparison with both 3D data and 2D data. Fig 5.5 shows some examples of frontal renderings of the 3D reference models. The dataset will be available to other researchers via the SCS website⁷⁰. Additionally, the dataset includes 20 frames with different views of a 2D checkerboard, which could be used for camera calibration. Both the values of

the calibration matrix and the lens distortion parameters can be estimated using the camera calibration data. The input frames for the experiments are undistorted using the lens distortion parameters. The calibration matrix is used in the proposed shape reconstruction method.

Figure 5.5: Samples from the recorded dataset. Images in the first row: Examples of the dataset from multiple persons with different views. Images in the second row: Frontal reference images used for face comparison from different persons.

5.4 3D FACE RECONSTRUCTION METHOD

5.4.1 BACKGROUND

We model the face as a rigid object which translates and rotates in front of a static camera. For each frame we have to find the rotation and translation parameters of that specific view to be able to estimate the shape of the face. The rigid shape of the face can be described by one set of 3D landmarks. We will use 2D landmarks with known correspondences to the 3D landmarks in all views to obtain a 3D reconstruction of the landmarks on the face. We use the pinhole camera model described in³¹, which projects 3D points on a 2D image plane. The projection of a point is usually written as $q = K(PQ)$, where q is a 3×1 homogeneous vector describing a position on the image plane, P is a 3×4 matrix containing the rotation and translation parameters of a view and Q is a single 4×1 homogeneous 3D point, K is a 3×3 matrix describing the internal camera calibration parameters, see³¹. The number of landmarks we use, is too low to estimate camera

parameters and shape unambiguously, therefore we assume that the camera calibration matrix K is known or can be obtained separately. Both sides of the projection equation are premultiplied with K^{-1} , leading to the point equation for a calibrated camera, see Eq 5.4.1:

$$w = K^{-1}q = K^{-1}K(PQ) = PQ \quad (5.4.1)$$

where w is now a homogeneous vector describing a point on the image plane in world coordinates.

5.4.2 SHAPE RECONSTRUCTION

Our goal for the 3D shape reconstruction is to obtain an estimate of the 3D shape of a face. This estimate can only be obtained simultaneously with an estimate of the rotation and translation parameters of all views. First we assign a label $i \in \mathcal{I}$ to each view and a label $j \in \mathcal{J}$ to each landmark point, see Eq 5.4.2:

$$\begin{aligned} \mathcal{I} &= \{1 \cdots N\} \\ \mathcal{J} &= \{1 \cdots M\} \end{aligned} \quad (5.4.2)$$

We denote the estimated 3D points $\hat{W}_j$ and we denote the estimated rotation and translation parameters $\hat{P}_i$. The shape is estimated using manually labeled 2D landmarks. We call these 2D landmarks w_{ij} with $i \in \mathcal{I}$ and $j \in \mathcal{J}$. The 2D Root Mean Square (RMS) reprojection error, see Eq 5.4.9, is used to minimize the sum of squared distances between the landmarks w_{ij} and the reprojected 3D points $\hat{W}_j$, based on the rotation and translation parameters estimated in $\hat{P}_i$. We define a set ω_i of landmarks for view $i \in \mathcal{I}$, see Eq 5.4.3. Some landmarks might not be visible, but still have an empty entry in the set, we refer to them as invisible.

$$\omega_i = \{w_{ij}\}, j \in \mathcal{J} \quad (5.4.3)$$

The reprojections of the corresponding 3D points are defined by $\hat{w}_i$, see Eq 5.4.4. The reprojections are calculated from $\hat{P}_i$ and $\hat{W}_j$ which are reconstructed from a set of two or more views.

$$\hat{w}_i = \{\hat{P}_i \hat{W}_j\}, j \in \mathcal{J} \quad (5.4.4)$$

To calculate the reprojection error over multiple views, we define the set Ω of all views and the set $\hat{\Omega}$ of all reprojected views, see Eq 5.4.5:

$$\begin{aligned} \Omega &= \{\omega_i\}, i \in \mathcal{I} \\ \hat{\Omega} &= \{\hat{w}_i\}, i \in \mathcal{I} \end{aligned} \quad (5.4.5)$$

The mask matrix $\mathcal{M}_{ij}$ is introduced to deal with invisible landmarks w_{ij} , undefined estimated 3D points $\hat{W}_j$ and undefined view parameters $\hat{P}_i$. If either the landmarks, the view parameters

or the 3D points are undefined, the mask value is 0, otherwise 1, see Eq 5.4.6:

$$\mathcal{M}_{ij} = \begin{cases} 1 & \text{where } \hat{\mathbf{P}}_i, \hat{\mathbf{W}}_j \text{ and } \mathbf{w}_{ij} \text{ are defined} \\ 0 & \text{otherwise} \end{cases} \quad (5.4.6)$$

The reprojection error for one view $e(\omega_i, \hat{\omega}_i)$ is defined in Eq 5.4.7.

$$e(\omega_i, \hat{\omega}_i) = \sum_{j \in \mathcal{J}} \mathcal{M}_{ij} \|\mathbf{w}_{ij} - \hat{\mathbf{P}}_i \hat{\mathbf{W}}_j\|^2 \quad (5.4.7)$$

The total sum of entries $\mathcal{N}$ in matrix $\mathcal{M}_{ij}$ is used to normalize the RMS reprojection error for a set of views Ω , see Eq 5.4.8.

$$\mathcal{N} = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \mathcal{M}_{ij} \quad (5.4.8)$$

Finally we can define the RMS 2D reprojection error E_{2D} for a set of views Ω and reprojected views $\hat{\Omega}$, see Eq 5.4.9. The 2D reprojection error expresses the accuracy of the shape reconstruction.

$$E_{2D}(\Omega, \hat{\Omega}) = \sqrt{\frac{1}{\mathcal{N}} \sum_{i \in \mathcal{I}} e(\omega_i, \hat{\omega}_i)} \quad (5.4.9)$$

After defining the 2D reprojection error we describe the proposed algorithm for the minimization of the reprojection error. First we give an impression of the manually annotated landmarks in Fig 5.6. Almost none of the landmarks are visible in all views due to the pose variation in each view. The correspondence between the landmarks in every frame is assumed to be known. Since both the views and the shape are unknown, we have to start by obtaining an initial estimate of a subset of the 3D points of the shape. Starting with this initial shape estimate, more views can be added and the shape can be re-estimated based on these views. However, only small changes, both in shape and views, can be handled in the optimization method without getting an unstable estimate of either the shape or the views. Therefore we propose an iterative optimization method. In short the proposed algorithm selects an initial pair of views with a sufficiently small reprojection error. Based on this pair of views we obtain a linear estimate of the shape. Then we start an iterative procedure in which we add one new view in every iteration. After adding the new view, the current selection of views and the current shape estimate are optimized. We continue this iterative procedure, until all views are added and optimized. The result of the reconstruction algorithm is an estimate of the 3D positions of the landmarks and an estimate of the rotation and translation parameters of each view. In the next subsections we describe the algorithm in more detail.

INITIALIZATION: FINDING AN INITIAL PAIR OF VIEWS

To avoid getting stuck in local minima, it is important to find a good initialization point. We use the reprojection error calculated on pairs of views to compare views with each other. We start by

Figure 5.6: Manually defined landmarks. Left and right view of a face with the defined landmarks.

calculating the reprojection error for every possible pair of views in the dataset. To calculate the reprojection error we need to know the rotation and translation parameters of each view. These values are extracted from the Essential Matrix using a robust MSAC (M-estimator Sample Consensus) method⁷⁹. When we have extracted an estimate of the rotation R and translation t , we can reconstruct an estimate of the 3D shape using the triangulation method described in³⁴.

To select an initial pair of views, there are some considerations:

- The reprojection errors from other views based on the structure of the initial pair are important to maintain the consistency within the set of views.
- The number of reconstructed points, which is equal to the number of corresponding landmarks in the initial pair of views, should be as high as possible to prevent getting stuck in a local minimum.
- The reprojection error should be as low as possible, because this indicates a more accurate reconstruction.

Unfortunately, not all these considerations can be fully satisfied with the same pair of views. In the best case we will find a sufficiently good pair of views. Eq 5.4.10 describes the selection criterion for finding a sufficient good initialization, where Ω is the set of all views, E_{2D} is the reprojection error based on different reconstructions in $\hat{\Omega}$. $\mathcal{V}$ is a set of different $\hat{\Omega}$ based on the triangulation of the views ω_r and a second view. We select the subset $\mathcal{V}_{25}$ containing sets of reconstructed views $\hat{\Omega}$ with a reprojection error below the threshold of the 25th percentile t_{25} value of all reprojection errors:

$$\mathcal{V}_{25} = \left\{ \mathcal{V} \mid E_{2D}(\Omega, \hat{\Omega}) < t_{25} \right\} \quad (5.4.10)$$

From the selected subset $\mathcal{V}_{25}$ we take the set that reconstructs the highest number of 3D points $\mathcal{C}(\hat{\Omega})$:

$$\hat{\Omega}^* = \underset{\hat{\Omega} \in \mathcal{V}_{25}}{\operatorname{argmax}} (\mathcal{C}(\hat{\Omega})) \quad (5.4.11)$$

From this set $\hat{\Omega}^*$ we pick the view $\hat{\omega}^*$ that was used for the triangulation of $\hat{\Omega}^*$. The associ-

ated view ω^* forms the initial pair together with the reference view ω_r . The initial pair of views provides us with a solution that is sufficient to initialize our iterative optimization. The initial linear estimate of the shape is calculated based on the selected pair of views.

OPTIMIZATION: ITERATIVE OPTIMIZATION OF ALL VIEWS

The optimization step is initialized with the selected pair of views and the associated shape estimate. Outliers of the 3D shape estimate are removed based on the relative distance of a point to all other points. In the optimization step, one new view, called candidate, is added to the selection of views in each iteration. Adding more views at once is possible, but would increase the chance of getting stuck with an inconsistent set of views. A candidate is selected based on the reprojection error of the candidate using the current estimate of the shape. Fig 5.7 illustrates the process of shape reconstruction, where after several iterations an increasing number of 3D points is reconstructed.

Figure 5.7: Shape reconstructions of each iteration. The number of reconstructed landmarks increases during the reconstruction process. The alignment of the 3D model is only performed if all points are reconstructed.

Based on the previous selection and the new candidate view, a new linear estimate of the shape is obtained. We use *Bundle Adjustment*³⁴ to optimize both the shape and the rotation and translation parameters simultaneously. The reprojection error is minimized using the Levenberg-Marquardt (LM) algorithm. To prevent wasting time on the current selection, we terminate the LM optimization process after 30 iterations. The result of the minimization process is new rotation and translation parameters for every view and a new estimate of the shape, that is used to select the next view candidate. The iterative optimization procedure continues to add new views until all views are added and optimized. Finally, we have an estimate of the 3D positions of the landmarks and an estimate of the rotation and translation parameters of each view.

5.4.3 TEXTURE RECONSTRUCTION

After reconstruction of the 3D shape, the next issue is combining the texture information from multiple views. Due to the difference in pose of each view, the observed intensity varies. The first step is to define patches on the surface of the reconstructed 3D shape. Later on these patches

are used for the illumination correction of the 2D images and re-illumination of the reconstructed face. In the second step we remove the illumination effect due to the viewing angle of the 2D frames. The third step is to combine all the texture information from the different views. Fourth, the frontal alignment step, which can be useful for the rendering of frontal views from the 3D model.

CHOOSING AN ILLUMINATION MODEL

We choose an appropriate illumination model based on the ability to remove illumination effects from the 2D images. Therefore, we assume a single dominant directional light source that is constant in all views. Next, we choose a reflection model, which describes how the light is reflected from a 3D surface. For the surface of the reconstructed 3D model, we choose a configuration of patches which splits at the symmetry axis of the face. This configuration is useful when many left and right side views are available. Fig 5.8 shows the defined patches of a face.

Figure 5.8: Selected areas with skin texture information. Other areas on the face often include details of the nose, mouth or eyes.

We use the Lambertian reflection model as basis for the illumination correction. The Lambertian reflection model includes two types of reflections: ambient reflection and diffuse reflection. These types of reflections depend on the properties of the surface, in this case the skin, hair, mouth and eyes. Following the generalized ambient reflection model described by Zhang and Gao⁹⁸, we use the same reflection coefficient κ for both the ambient reflections i_a and the diffuse reflections i_d . This results in the following illumination model, where I is the observed intensity:

$$I = \kappa i_a + \kappa i_d (\mathbf{n}^\top \mathbf{l}) \quad (5.4.12)$$

To apply the illumination correction we need to invert the reflection model and find the unknown parameters i_a , i_d and κ .

FINDING THE ILLUMINATION PARAMETERS

The first problem we encounter is the ambiguity between the coefficient κ and the intensities of the ambient and diffuse reflections. Since both factors of the product are unknown, there is no unique solution for i_a , i_d and κ , without having more knowledge about either the face or the light source. To resolve this ambiguity we assume a fixed reflection coefficient κ_s for patches of the face with only skin texture information. Fig 5.8 shows the patches with the regions in a face containing mainly skin texture. Based on the average intensities of the patches with skin texture information in multiple views, we can construct a system of equations to determine the combined ambient term $\kappa_s i_a$, that we denote as m_a , and the combined diffuse term $\kappa_s i_d$, that we denote as m_d . From each visible patch we take the average gray value and the average normal, based on the normals of the vertices of the patch, to calculate n_q and l_q . The system we construct has only two unknowns, so two or more patches are enough to estimate the ambient and diffuse terms m_a and m_d , see Eq 5.4.13, where q is the total number of visible patches and g_q is the average gray value in patch q . The system of equations becomes underdetermined if all the estimated dot products in the system are close to each other. The system is solved using multiple patches in multiple frames of the same person.

$$\begin{aligned} g_1 &= m_a + m_d(n_1^T l_1) \\ \vdots & \quad \quad \quad \vdots \\ g_q &= m_a + m_d(n_q^T l_q) \end{aligned} \tag{5.4.13}$$

We perform the illumination parameter optimization on a database of 48 persons, who move their heads in front of a camera. Each recording consists of 100 frames. If we solve the linear system of equations, on average the combined ambient term is 0.02 and the combined diffuse term is 0.71, the standard deviations are 0.02 and 0.08 respectively. The intensities for the ambient and the diffuse reflections can now be written as:

$$i_a = \frac{m_a}{\kappa_s} \quad i_d = \frac{m_d}{\kappa_s} \tag{5.4.14}$$

With the estimate of the ambient and diffuse terms, we can perform the illumination correction on all frames.

APPLYING THE ILLUMINATION CORRECTION

The goal of the illumination correction is to obtain the reflection coefficient κ , which is independent of illumination. With the found ambient and diffuse intensities we can calculate the illumination correction in every pixel, see Eq 5.4.15 and Eq 5.4.16, where I_p is the observed inten-

sity in pixel p and κ_p is the reflection coefficient in pixel p .

$$\begin{aligned} I_p &= \kappa_p i_a + \kappa_p i_d (\mathbf{n}_p^\top \mathbf{l}_p) \\ \kappa_p &= \frac{I_p}{i_a + i_d (\mathbf{n}_p^\top \mathbf{l}_p)} \end{aligned} \quad (5.4.15)$$

The normals of the pixels in the patches are calculated by interpolation of the normals in the vertices of the patches. We use Gouraud shading²⁵ and barycentric interpolation to calculate the interpolated values in every point on the surface. By substituting the intensities by the intensities of Eq 5.4.14, we obtain the following equations:

$$\begin{aligned} \kappa_p &= \frac{I_p}{\frac{m_a}{s} + \frac{m_d}{s} (\mathbf{n}_p^\top \mathbf{l}_p)} \\ \frac{\kappa_p}{\kappa_s} &= \frac{I_p}{m_a + m_d (\mathbf{n}_p^\top \mathbf{l}_p)} \end{aligned} \quad (5.4.16)$$

The reflection coefficients κ_p are linearly scaled by the constant κ_s , the reflection coefficient of the skin. The resulting $\frac{\kappa_p}{\kappa_s}$ is independent of a view and characterizes the texture in pixel p and the corresponding point on the 3D model. Fig 5.9 shows an example of the illumination correction, where patches before and after the correction are shown. The corrected intensities of the patches are more similar and the differences due to illumination are resolved.

Figure 5.9: Comparison of normal and corrected patches. The top row shows the upper left region of the forehead in six arbitrary views. The bottom row shows the same patches in the same views after illumination correction. The textures are transformed to equilateral triangles for comparison.

COMBINING TEXTURES

After the illumination correction, combining the texture information from all frames is still not straightforward. For views with a large viewing angle compared to the normal of a point on a patch, details of the texture are no longer visible. Therefore, we restrict the viewing angle compared to the normal of a point to a maximum of 45 degrees. Points with a larger viewing angle contain low resolution texture information, which is superseded by the other views. For each point on a patch and for each view we select only the texture information from points with

an angle smaller than 45 degrees. We call this set Δ_p . The final texture $\hat{\kappa}_p$ in point p is constructed by taking the median of all selected texture values δ_p in each point p :

$$\hat{\kappa}_p = \text{median}_{p \in \Delta_p} (\delta_p) \quad (5.4.17)$$

The median is applied to avoid the influence of points p with extreme texture values. Fig 5.10 shows some examples of a 3D model with combined textures. Finally, the illumination model, see Eq 5.4.12, is applied again to render shaded face images. We choose $i_d = 0$ and i_a such that the intensities of the images fall within a normal range. Some frontal reconstructions can be seen in Fig 5.10. Notice that the reconstruction of the area of the eyes is the most difficult part due to the variation between all views. The alignment of the 3D model is conducted by detecting the symmetry plane of the head and by rotating the landmark on the tip of the nose, around the symmetry plane, to a frontal position.

Figure 5.10: Reconstructed face examples. Images in the first row: Multiple views of the same reconstructed 3D face model. Images in the second row: Examples of frontal reconstructed views after alignment of 3D models from two different persons.

5.5 FACE COMPARISON

In this paper our purpose is to investigate if a frontal rendering of the reconstructed textured 3D model could surpass the face comparison scores of the original 2D frames without introducing a bias. Since we propose a data driven approach, the potential of introducing a bias is minimal. The reconstructions are compared to the same frontal reference images, see Fig 5.5, as the input images to validate whether the comparison scores have been improved or not. We do not perform a direct comparison with model based reconstruction methods, which are unsuitable for forensic face comparison due to their bias.

5.5.1 EXPERIMENT

We use state of the art facial comparison software, namely FaceVACS¹⁷, to perform the face comparisons. FaceVACS compares two face images and provides comparison scores in the range of 0.0 to 1.0. Genuine scores are usually between 0.4-1.0, where full frontal neutral faces, for example, have score values above 0.95. From preliminary experiments we have learned that frontal reconstructions never outperformed the full frontal face images in our dataset. Therefore we take a selection without any high quality frontal frames to perform the proposed face reconstruction method. This experiment is similar to a forensic case where, for example, only top and side views of the face are available, but not any frontal views. If frontal views are available there is no need for a reconstruction. We obtain a selection from our dataset based on the recognition score of the individual frames. All frames with a recognition score above a threshold t are removed from the set, prior to the shape reconstruction. The threshold t changes with steps of 0.1 in each experiment and varies between 0.2 to 0.5. A threshold below 0.2 doesn't provide enough subsets with at least 30 frames to obtain reconstructions. In the experiments we use 20 landmarks in 30 frames to reconstruct the shape and the texture information of the face. The proposed reconstruction method needs at least 30 frames, so when the subsets contain less than 30 frames, no reconstructions are obtained for those subsets. In the case of more frames, 30 frames are automatically selected based on the sharpness of the images and maximization of the visibility of each landmark in the subset, since the performance decreases with lack of visibility of the landmarks. The reconstructed 3D models with texture information are transformed to a frontal face position using the symmetry plane of the face and the landmark on the tip of the nose. Once we have the aligned frontal 3D model we can render a 2D frontal face image using the reconstructed 3D model. For each reconstruction we calculate the FaceVACS comparison scores using the reference images. Since all subsets of higher thresholds include the subsets from lower thresholds, we take the maximum over the current and all lower comparison scores for each person. This is valid, because all the required frames are included in the supersets with a higher threshold t .

5.5.2 VISUAL RECONSTRUCTION RESULTS

A few samples of good quality reconstructions are already shown in Fig 5.10. Some of the quality issues with the reconstructions can be explained by visual inspection of the frontal views and the 3D shape reconstructions. Three examples of different quality reconstructions are shown in Fig 5.11. The reconstruction on the left in Fig 5.11 is a high quality reconstruction with minor

artifacts around the eyes and nose. The reconstruction in the middle has more severe texture artifacts around the eyebrows and the chin, which are usually caused by deformations of the face, such as expressions. Although the artifacts look severe to the human eye, the reconstruction score is above the threshold value. The reconstruction on the right shows problems with merging the texture information from both sides of the face. FaceVACS is quite sensitive to such kind of artifacts and the resulting score is below the threshold value. Including more frames or selecting the input frames manually can solve this problem. Although some of the artifacts of the reconstructions look severe, the artifacts are all intrinsic to the input data and not caused by a facial model, which makes the reconstructions applicable in a forensic context.

Figure 5.11: Multiple frontal reconstruction results. Frontal reconstructions from different persons using a selection of frames with a score threshold of 0.50.

5.5.3 COMPARISON RESULTS

The performance of the reconstructions based on the different thresholds is visualized in Fig 5.12, where the FaceVACS scores of the reconstructions are compared to the scores of the set of input frames. The horizontal axis shows the FaceVACS scores, where the vertical axis shows the percentage of reconstructions above these FaceVACS scores. The gray area of the graph shows the comparison scores of the input frames. These scores are bounded by the best input frames of the subset below a threshold t . The blue region shows the reconstructions that didn't exceed the threshold score of the input set of frames. In these cases selecting the highest scoring frame of the input set is the best option. The green area shows the reconstructions that exceeded the threshold value t . The results show a significant gain over the original 2D comparison scores. These comparison scores can only be obtained using the proposed reconstruction method on the input sets of 2D frames. However, the experiment shows clearly that it becomes more and more difficult to improve the comparison scores for a higher threshold t . For reconstructions based on lower quality data, frame pose is more important than the introduced artifacts, but this changes when the image quality increases. The best results are obtained with a threshold of 0.5, where some high quality reconstructions are obtained from the selection of frames. Also the threshold of 0.4 provides a considerable improvement, but never provides comparison scores above 0.8. For lower thresholds fewer reconstructions are obtained in total, because of the minimum number of frames needed for the shape reconstruction.

5.6 CONCLUSION

In this research we addressed forensic face comparison, one of the major issues in forensics. We introduced, in contrast to the Morphable Model, a model-free reconstruction method that does not compromise the evidential value of the facial reconstruction by introducing supplementary data. The proposed method is based on 20 landmarks in 30 frames to obtain a 3D reconstructed face model. We developed a two stage algorithm in which we first reconstructed the shape of the face and in the second stage reconstructed the texture. The texture illumination correction is an important step in the texture reconstruction process, because it allows us to merge the textures from multiple views. Especially in the case where only low-quality and non-frontal images are available, for example only side and top views, our proposed method can surpass the FaceVACS comparison scores of the original frames. Our method can be used in a forensic context, because no supplemental data or models are used in our reconstruction algorithm.

5.7 ACKNOWLEDGEMENT

This work was supported by BZK 5.50: ‘Gezichtsvergelijking op basis van niet gekalibreerde Camerabeelden’ in a cooperation between the Netherlands Forensic Institute and University of Twente. We would like to thank Cognitec Systems GmbH. for supporting our research by providing the FaceVACS software. Results obtained for FaceVACS were produced in experiments conducted by the University of Twente, and should therefore not be construed as a vendor’s maximum effort full capability result.

5.8 PAPER CLOSURE

Based on the landmarks, a complete coarse reconstruction of the face is obtained. This is a first step towards a forensic method for face reconstruction. Although there are constraints on the input data and visibility of landmarks in the frames, the proposed method is able to improve the face comparison scores, compared to the original images. Because no statistical models of faces are used, the proposed method is usable in a forensic context.

Figure 5.12: Histogram of the face comparison results. Cumulative histograms of the percentage of reconstructions above a certain FaceVACS score. The gray area of the graph shows the scores of the input frames. The blue area indicates the gap between the threshold t and the highest scoring frame of the input set. The green area shows the gain of the reconstruction method in comparison to the best original 2D frame.

6

3D Reconstruction from Video Sequences: Dense 3D Face Models

CHRIS VAN DAM
LUUK SPREEUWERS
RAYMOND VELDHUIS

In this chapter we extend the proposed reconstruction method to a dense 3D reconstruction method. We show on the same PV_{3D} dataset, using the coarse landmark reconstruction as a starting point, that we can extract 3D information from multiple images for forensic face recognition. We introduce a completely new texture reconstruction method based on the albedo of the face. This texture reconstruction method is incorporated in the optimization process. We make a step forward in improving the forensic face comparison process, without compromising the integrity of the original frames. The work in this chapter was published in 2015 in IET Biometrics.

FACE RECONSTRUCTION FROM IMAGE SEQUENCES FOR FORENSIC FACE COMPARISON⁸⁵

6.1 ABSTRACT

We explore the possibilities of a dense model-free 3D face reconstruction method, based on image sequences from a single camera, to improve the current state of forensic face comparison. We propose a new model-free 3D reconstruction method for faces, based on the Lambertian reflectance model to estimate the albedo and to refine the 3D shape of the face. This method

avoids any form of bias towards face models and is therefore suitable in a forensic face comparison process. The proposed method can reconstruct frontal albedo images, from multiple non-frontal images. Also a dense 3D shape model of the face is reconstructed, which can be used to generate faces under pose. In our experiments the proposed method is able to improve the face recognition scores in more than 90% of the cases. Using the likelihood ratio framework, we show for the same experiment that for data initially unsuitable for forensic use, the reconstructions become meaningful in a forensic context in more than 60% of the cases.

6.2 INTRODUCTION & BACKGROUND

Face comparison in a forensic context is a challenging task. Reconstructing faces with a model driven approach^{9,22,71,66,41,16,67,20,55,51,80,37,28} has been addressed by several researchers, but always suffers from a bias towards the underlying face model. For example a face model trained on multiple look-a-like faces of a person is far more likely to match with that person than a general face model. In a forensic context avoiding any form of bias towards other faces is crucial to create better opportunities for evidence in court. In this paper we explore the possibilities of a dense model-free 3D face reconstruction method, based on an image sequence from a single camera. In the first stage we use landmarks in multiple images to obtain a coarse estimation of the shape of a face and the rotation and translation parameters of the face in each frame. In the second stage we enhance the reconstruction by estimating the reflection coefficient and surface normals simultaneously for every point in the 3D shape. Based on the estimate of the normals we enhance the 3D shape of the face. Applying these steps iteratively leads to a dense 3D reconstruction of a face including texture information based on the reflection coefficients of the face. The proposed method is model-free and has therefore no bias towards an underlying face model. We further refer to this as the model-free requirement. As a consequence of this requirement, the reconstruction process becomes more difficult, since no average model can be used as a starting point for the 3D reconstruction. And where, for example, the Morphable Model method can provide reconstructions based on a single image⁹, we need multiple images. Our goal is to increase the performance of face recognition, while still maintaining the model-free requirement of the forensic context.

Although some parts of our work are related to stereo and multi-camera reconstruction approaches, see for example⁷³, our problem is more complex. First of all in a stereo setup the exact positions and viewing directions of the cameras are known, while we need to estimate the position and direction of each view. Estimations are less precise and depend on the quality of the data. Secondly, there is deformation of the faces due to expression and motion artefacts. In a multi-camera setup this could never happen, because in such a case all images are recorded at the same moment in time. Thirdly, the constant brightness assumption⁷⁴ is not applicable in a single camera situation, because the object is moving, and so the illumination changes in each frame. Therefore, there is more uncertainty during the reconstruction process. Garg et al.²³ introduce a variational nonrigid structure from motion (NRSfM) approach, but they use a massive number of tracking points, which is not realistic using forensic data. In their experiments they use synthetic face sequences, where the level of difficulty for reconstruction is far less than with real image sequences. Delaunoy and Pollefeys¹⁸ present a photometric bundle adjustment method for multi-view 3D modelling which shows low reprojection errors, but all their

experiments are performed on feature rich objects including unrealistic views in comparison to faces under pose.

In our reconstruction process we handle a moving face in front of a static camera, so the constant brightness assumption is not valid in our case. We introduce a reconstruction process based on image sequences which avoids any bias towards face models. Our reconstruction method provides dense 3D reconstructions of the face sequences.

The remainder of this paper is structured as follows: First our proposed method is introduced in Section 6.3, including a dense reconstruction algorithm. In section 6.4 we included face comparison experiments on frontal reconstructions, the forensic implications of the proposed reconstruction method and visual inspection of the dense 3D reconstructions. Section 6.5 concludes our paper.

6.3 METHOD

The proposed method is designed to work with real image data, comparable to image sequences obtained from an ATM camera. Multiple non-frontal images under varying pose are available, captured from a moving face with a single camera. Due to this setup the illumination is different in every image. Our goal is to obtain a dense 3D reconstruction of the face without introducing any bias towards face models. The reconstruction will be used for forensic facial comparison.

In previous work⁸³ we presented a coarse 3D shape reconstruction method for faces. This coarse shape reconstruction method is model-free, which is important for usage in a forensic context. The reconstruction method is based on 20 manually labelled landmarks in multiple views to support forensic data with low quality images. First one pair of images is selected for the start of an iterative reconstruction process. A 3D shape estimate is obtained based on the initial pair of images. In each iteration one new image is added and both the 3D shape and the view of the face in each frame are optimized simultaneously. The reconstruction method is a structure from motion approach that minimizes the 2D reprojection error, the error between the reprojected landmarks and the landmarks in the input frames. Several steps were taken to provide a robust reconstruction method for a low number of landmarks. The output of the 3D shape reconstruction algorithm is an estimate of the 3D points and an estimate of the rotation and translation parameters of the face in each view. The method is tested on simulated data: a random point clouds and a set of views rendered from a styrofoam head. An analysis on the 2D reprojection error and the 3D error using a varying number of views is provided, for more details we refer to⁸³. Finally our previous work shows that the shape reconstruction algorithm is capable of handling real image data.

In this article we propose a 3D reconstruction method based on a coarse 3D shape and a powerful texture reconstruction method to obtain dense 3D reconstructions and frontal reconstructions of high quality. The coarse 3D shape reconstruction is obtained using the shape reconstruction algorithm in⁸³, including the rotation and translation estimations of the views. The rotation and translation parameters for a view v define a rigid transformation T_v , that we use for illumi-

nation estimation. Since the reconstructed coarse shape only consists of 3D points, we define 28 triangular patches on the reconstructed coarse shape, see Figure 6.1, to obtain a 3D surface. The patches are chosen in such a way that the surface has a seam running along the symmetry plane of the face. The patches are used for initialization of the dense shape reconstruction.

Figure 6.1: Coarse 3D shape reconstruction with added surface.

The proposed dense reconstruction method is based on the Lambertian reflectance model, where we assume ambient light l and a single directional light source l_d . The light source l_d is posed perpendicular to the frontal face. The illumination model we describe is object oriented and therefore we transform the constant directional light source l_d to object coordinates and make the light direction dependent on view v . We use a rigid transformation T_v , based on the estimates of the rotation and translation parameters, to calculate the light direction in each view, see Equation (6.3.1).

$$l_v = T_v(l_d) \quad (6.3.1)$$

We define our illumination model in Equation (6.3.2):

$$\hat{G}_{xv} = \alpha_x l + \rho_x (n_x^\top l_v) \quad (6.3.2)$$

where α_x is the ambient reflection coefficient, l the ambient light intensity, ρ_x is the diffuse reflection coefficient (albedo) in point x , n_x is the 3D normal in point x , l_v is the directional light source, including the light intensity, dependent on view v , see Equation (6.3.1), and $\hat{G}_{xv}$ is the predicted intensity in point x using view v . The Lambertian reflectance model connects the 3D shape, via the normals of an object, to the observed intensities in images. We use this connection to optimize both the texture and the 3D shape. From the coarse reconstruction of the shape, estimates of the normals in each 3D point are available, but need to be optimized. Each 3D point x can be projected to multiple views, which we use to optimize the albedo and the normal of

each point. We define the following optimization, see Equation (6.3.3):

$$(\xi_x^*, \mathbf{n}_x^*) = \operatorname{argmin}_{\xi_x, \mathbf{n}_x} \sum_{v \in V} |G_{xv} - \hat{G}_{xv}| \quad (6.3.3)$$

where ξ_x^* is the optimized albedo, $\mathbf{n}_x^*$ is the optimized normal, V is the set of frames used for the reconstruction, G_{xv} is the observed intensity of point x in view v and $\hat{G}_{xv}$ the predicted intensity given Equation (6.3.2), where the term $\alpha_x l$ is constant and has no influence on the optimization. Because there is an ambiguity between the intensity of l_v and the albedo ξ_x for optimization purpose we assume the L2-norm of l_v to be one. The direction of the light in l_v , however, is still depending on view v . The albedo ξ_x and the normals $\mathbf{n}_x$ of the 3D reconstruction are both optimized by minimization of the absolute difference with the observed intensities. For each point x that is visible in at least 4 views, we are able to optimize the albedo and the normal. The normals are calculated from the shape for every point x using its neighbours, see Figure 6.2, where x is the index of the central point and 1–4 are the indices of the neighbouring points. The 3D points are represented in a depth map, where each point in the grid contains the Z-coordinate as depth.

Figure 6.2: A 3D point x on a grid and its neighbours form the triangular patches A-D. These patches are used to calculate the normal in point x .

The normals of the four patches A-D are calculated with Equation (6.3.4), where p_x is the central point, p_i are the neighbours counted clockwise, see Figure 6.2, starting with the leftmost point and $\|\cdot\|_2$ denotes the Euclidean norm.

$$\mathbf{n}_{xi} = \frac{(p_x - p_i) \times (p_x - p_{(i \bmod 4)+1})}{\|(p_x - p_i) \times (p_x - p_{(i \bmod 4)+1})\|_2} \quad (6.3.4)$$

The normal in point x is then calculated based on the triangular patches A-D, see Equation

(6.3.5), where we define the depth of p_x as: $d_x = [p_x]_z$, the Z-coordinate of p_x .

$$\hat{n}_x(d_x) = \frac{n_{x1} + n_{x2} + n_{x3} + n_{x4}}{\|n_{x1} + n_{x2} + n_{x3} + n_{x4}\|_2} \quad (6.3.5)$$

Based on the optimized normals we adapt the depth of the 3D shape (and the associated normals) to fit the optimized normals, see Equation (6.3.6), where d_x^* is the optimized depth in point x .

$$d_x^* = \underset{d_x}{\operatorname{argmin}} \|n_x^* - \hat{n}_x(d_x)\|_1 \quad (6.3.6)$$

We perform the full process multiple times using different resolutions of the 3D shape to obtain our final reconstruction and albedo estimations. The proposed reconstruction method is described in Algorithm 1.

Algorithm 1 Dense Reconstruction Algorithm

Input: V is a set of rotation and translation parameters
 S is a coarse shape description with surface
 I is a sequence of non-frontal face images
Output: G is a grid with depth values
 $\mathcal{A}$ is a set of albedo values with the same resolution as G

```

1: procedure ITERATIVEDENSERECONSTRUCTION( $V, S, I$ )
2:    $(V, S) \leftarrow \text{TransformToFrontalShape}(V, S)$   $\triangleright$  Transformation, see Equation (6.3.1)
3:    $G \leftarrow \text{SampleShapeToGrid}(S)$   $\triangleright$  Sample 3D depth values to grid
4:   while Resolution( $G$ )  $\neq$  MaxResolution do
5:      $G \leftarrow \text{SmoothDepthGaussian}(G)$   $\triangleright$  Smoothing needed due to upsampling
6:      $N \leftarrow \text{CalculateNormals}(G)$   $\triangleright$  Set of normals, see Equation (6.3.5)
7:      $(\mathcal{A}, N) \leftarrow \text{OptimizeNormalsAndAlbedo}(V, N, I)$   $\triangleright$  Optimization, see Equation (6.3.3)
8:      $(G, N) \leftarrow \text{AdaptDepthMatchNormals}(G, N)$   $\triangleright$  Adapt the depth, see Equation (6.3.6)
9:      $\mathcal{A} \leftarrow \text{RecalculateAlbedo}(G, N, I)$   $\triangleright$  Calculate albedo using adapted depth
10:     $(G, N) \leftarrow \text{UpsampleGrid}(G, N)$   $\triangleright$  Obtain 3D shape with higher resolution
11:  end while
12:  return  $G, \mathcal{A}$   $\triangleright$  Return optimized albedo and dense 3D shape description
13: end procedure

```

The shape and the albedo of the face are both optimized and provide a more detailed and dense 3D shape description of the face. In each iteration of the algorithm we obtain an estimate of the albedo of the face and a description of the 3D shape. Note that the proposed method does not optimize or improve the rotation and translation parameters and the internal camera calibration values of the input data, that are defined in V . These parameters are considered precise enough to perform the optimization.

The coarse 3D shape is aligned in Step 2 using the eigenvectors of the symmetry plane of the reconstructed shape and by rotating the vector, defined by the tip of the nose and the center of gravity of the symmetry plane, around the axis through the center of gravity of the symmetry

plane, to a frontal position. The face is now in an object oriented coordinate system based on the shape. In Step 3 a grid search is used to sample the nearest neighbours on the surface of the coarse 3D shape. A grid of 100×100 points is an appropriate starting point for the proposed method. The size of the grid is important, because the smaller the grid the more 3D information is described by each point in the grid. So, with a smaller size of the grid larger shape deformations occur, for a large size of the grid only small details of the shape are changed. We stop at a size of 400×400 points, because the alterations are barely visible at this stage. The factor for upsampling in Step 10 in the algorithm is also of importance, because of the 3D information described by each point in the grid. The upsampled depth values of the grid are calculated by bilinear interpolation. We set the factor for upsampling to $\sqrt{2}$, which is appropriate according to the details that need to be reconstructed. Throughout an iteration only the depths of the points on the grid are adapted. In the optimization in Step 7 we use a selection of normals with an angle smaller than 60° compared to the direction of the light, since these normals are more accurate than normals with larger angles. Although this angle could be smaller, this would lead to points in the grid where we are unable to minimize Equation (6.3.3) due to the low number of normals. The minimization of Equation (6.3.3) is attained by finding the optimum values using an exhaustive search within small variations of the angle, at most 5 degrees in both direction in multiple steps, for the estimated normals and by calculating the albedo values in 256 steps within the range $[0..1]$. During the shape adaptation in Step 8 we test three options for each point in the grid: We increase the depth, we decrease the depth or we keep the current depth. For each of these options we calculate the absolute difference between the optimized normals in Step 7 and the normals calculated from the current 3D shape, see Equation (6.3.6). We choose the depth with the smallest absolute difference. The altered depth is directly applied on each point. We continue altering the depth until there is no more change of depth. We used depth alterations of 0.5mm at the time. The small steps lead to a slower convergence, but prevent extreme changes in the normals of the shape. Secondly, the small steps support a continuous 3D shape reconstruction.

In Figure 6.5 on the right, an example of a frontal albedo reconstruction is shown. Figure 6.10 shows multiple renderings of a reconstructed 3D shape with the reconstructed albedo as texture.

6.4 EXPERIMENTS

To evaluate the performance of the proposed reconstruction method, we performed a 2D face comparison experiment on the albedo reconstructions. We compared the reconstructed albedo images with the ground truth images of the corresponding persons. Secondly, we evaluate the implications for forensic face recognition. In the last part of this section we also visually inspect the reconstructed 3D models to indicate the quality of the 3D reconstruction.

6.4.1 DATASET

The dataset required for our experiments cannot be extracted from benchmark datasets such as CMU Multi-PIE, because in those datasets the data is taken at the same moment in time with multiple cameras. For these recordings the illumination estimation and depth estimations become much easier, because of the constant brightness and lack of expression and deformation

of the face. Therefore, we had no choice but to record our own dataset. Our dataset consists of recordings of 48 people. Each recording contains 101 frames with different views of the face of a person, recorded with a single camera. In the experiments we use subsets of this dataset, based on the availability of a set of frames suitable for reconstruction and a coarse shape estimate with a reasonable reprojection error. The selection criterion of the subsets is explained later in this section. Figure 6.3 shows some samples of our dataset. The persons in the data set were asked to rotate their heads to the left and to the right, while looking up or down at the same time. As result a variety of frontal, near-frontal and frames under pose are captured in the dataset. The rotation and movement of the head simulate the actions performed at an ATM. All the frames were captured in about 45 seconds, similar to forensic time lapse recordings. Ground truth frontal views were taken afterwards with a different camera setup, during the same session.

Figure 6.3: Examples of the dataset from multiple subjects with different views, showing minor expressions and some image artefacts.

6.4.2 2D FACE COMPARISON

In the first experiment we use the albedo reconstructions for comparison with ground truth images. The albedo reconstructions are stored in each iteration of the proposed method with increasing resolutions. Each albedo reconstruction is compared to the corresponding ground truth frontal view. We use the B7 algorithm of FaceVACS for face comparison, which is part of a commercial face comparison SDK¹⁷. FaceVACS is known in the face comparison community to be an excellent 2D face comparison algorithm. The comparison scores of FaceVACS are in the range $[0..1]$. Usually a threshold of 0.4 or above is taken for genuine scores. We use 0.5 in our experiment as genuine threshold. To show the ability of the proposed reconstruction method we select a subset of 2D frames from the dataset with FaceVACS scores below 0.5. These frames, which have non-frontal pose, expression or motion artefacts, are used in the reconstruction process, see Figure 6.4 for some examples. We refer to this set as selection Σ . Improvement of the frames in selection Σ is important in forensic cases where no frontal images of the face are available. We used a subset of 30 frames from the selection Σ to ensure that every reconstruction

is based on the same number of frames. This is also an appropriate number of frames for shape reconstruction according to earlier experiments, see⁸³, and for a reliable estimation of the albedo of the face using the optimization in Equation (6.3.3). The subset of 30 frames from selection Σ is selected based on their variation in view and image sharpness.

Figure 6.4: Examples of 4 frames from an input set of 30 frames.

In Figure 6.5 an example of one of the reconstructions is shown together with its corresponding ground truth image. In most of the cases the reconstructions seems to be slightly squint-eyed, because of the different origin of the views in the eyes. We notice that, although, humans are sensitive to the squint reconstruction artefacts, the performance of the FaceVACS algorithm is not affected by these artefacts.

Figure 6.5: Right: Example of reconstruction in the highest resolution. Left: The corresponding gallery image.

The total result of the experiment is shown in Figure 6.6. This graph shows the cumulative percentage of scores above a certain FaceVACS score for all reconstructions based on the selection Σ . For example 20% of the reconstructions have a FaceVACS score above 0.8. The scores are the maximum scores taken over all albedo reconstructions of different resolutions of each person. The green area visualizes the gain of the proposed reconstruction method. The grey area indicates the maximum scores of the input frames from selection Σ . The blue area indicates reconstruction scores below the threshold of 0.5. In that case taking the best input frame would probably give better results. The graph shows that in 91% of the reconstructions we surpassed the threshold score 0.5 of the input set of frames. In quite some cases we have a considerable gain, in some cases even up to 0.99. Such scores are in the same range as high quality frontal face images. Only in 9% of the reconstructions the proposed method did not improve the recognition results.

Figure 6.6: FaceVACS reconstruction results exceeding the threshold in 91% of the cases.

6.4.3 IMPLICATIONS FOR FORENSIC FACE COMPARISON

To give an indication of how FaceVACS scores should be interpreted in a forensic context, we provide likelihood ratios based on the statistics in the FaceVACS SDK¹⁷. Here, likelihood ratios describe the ratio between the probability of a comparison score, given that that faces compared originate from the same person and the probability of a comparison score, given that that faces compared originate from different persons, see⁵³. This is the standard approach when evaluating the meaningfulness of forensic evidence. If we calculate the likelihood ratios for all FaceVACS scores, we can compare the result of the best 2D input frame with the reconstruction to see how the reconstruction increases the likelihood ratio. The FaceVACS scores can be converted into likelihood ratios based on the FAR and FRR, derived from a large test database¹⁷. In Figure 6.7 a recreated FRR and FAR graph is shown. The likelihood ratios can be calculated from this graph by applying Equation (6.4.1) to each score s , see⁸⁹.

$$LR(s) = \frac{1 - \frac{\partial(FRR)}{\partial s}}{\frac{\partial(FAR)}{\partial s}} \quad (6.4.1)$$

Score	Likelihood Ratio
0.4	0.4
0.5	3.5
0.6	4.0×10^1
0.7	4.9×10^2
0.8	6.2×10^3
0.9	7.2×10^4
1.0	7.5×10^5

Table 6.1: FaceVACS likelihood ratio per score of the FaceVACS algorithm.

The statistics provide a good indication of the likelihood ratios of the FaceVACS scores. As can be seen in Table 6.1 the likelihood ratio can increase by a factor of up to 1.0×10^5 due to the reconstructed albedo images. In more than 75% of the reconstructions the likelihood ratio increases by a factor of 10 or higher, which is a considerable gain. Since the statistics in Figure 6.7 are valid for both the genuine and non-genuine comparisons, there is no additional gain for the non-genuine reconstructions.

Figure 6.7: Expected FAR and FRR from FaceVACS algorithm.

If we convert the FaceVACS scores of the results to likelihood ratios, the reconstructions can be reviewed in a forensic context, see Figure 6.8. The maximum likelihood ratio for each set of input frames of selection Σ is 3.5, based on the score threshold of 0.5. If a forensic researcher takes, for example, a likelihood ratio of 100 as a minimum, the proposed method is able to provide meaningful results in more than 60% of the cases.

In Figure 6.9 on the top we show the results for all 23 reconstructions individually. The same

Figure 6.8: Likelihood ratios of the reconstructions. The grey area corresponds to the maximum likelihood ratio of the frames in selection Σ for each person. The green area shows the likelihood ratio of the reconstructions using selection Σ as input frames.

threshold of 0.5 was used for the selection of the input frames. For the other 25 persons in the dataset there was not enough low quality data available to obtain appropriate reconstructions. Increasing the quality threshold of 0.5 for the frame selection would enable the availability of more and higher quality frames in the selection, however, this would be a less challenging problem. Continuing the reconstruction procedure with less than 30 frames, sometimes even less than 10, would lead to less accurate frontal face reconstructions. The green bars in Figure 6.9 indicate that we were able to exceed the threshold scores for that current person. The red bars indicate the opposite. We are able to surpass the threshold score for 21 persons in the dataset.

In some occasions the reconstruction scores can be further improved by changing the field of view of the frontal view. We included an experiment where we changed the field of view of the frontal view to 30 degrees, which approximately fits the field of view of the camera of the input images. This was done by rendering the 3D shape including the albedo texturing of the 400×400 grid with a fixed field of view of 30 degrees to a frontal view. In some cases this gave an improvement of the results, see the bottom of Figure 6.9, but in other cases the performance was worse. We can take the maximum FaceVACS score over both experiments to get the optimal result. In some rare occasions there are strong motion artefacts and the proposed method is not able to surpass the FaceVACS scores. The only way to improve the scores in such a case would be to manually select the set of input frames to minimize motion artefacts and other abnormalities. We decided to not further pre-process or post-process the results to give a clear demonstration of the performance of the proposed algorithm.

We expect the quality of the albedo images to increase along with the resolution. If we look into the scores for all resolutions we notice that the bigger the resolution of the grid the higher the percentage which surpasses the comparison threshold, see Table 6.2. The gain here represents the average increase over all 23 reconstructions of the FaceVACS scores compared to the threshold score. The maximum gain possible is 0.5. We decided to stop at 400×400 points

Figure 6.9: Top: FaceVACS scores per person. Bottom: FaceVACS scores per person using a fixed field of view of 30 degrees.

for the grid. Although higher resolution reconstructions still have a positive effect on the gain, the calculation time increases quadratically and the gain becomes smaller each time. The last row shows the average increase for the resolution with maximum score, because sometimes an increase of resolution doesn't result in an increase of the FaceVACS scores.

6.4.4 3D VISUAL INSPECTION

Apart from the albedo, also the 3D shape is reconstructed with the proposed algorithm. Some examples of these full 3D reconstructions can be seen in Figure 6.10. The 3D models look quite realistic, but in most cases the nose is a bit flattened by the smoothing in the algorithm. For frontal views this effect is minimal, but for views under pose the effect is stronger. Small details of the 3D shape, like the shape of the lips and the shape of the eyes are visible on the reconstructed 3D models. Although, the 3D shape seems quite accurate, experiments with shape

Size	Gain
100 × 100	0.04
141 × 141	0.10
200 × 200	0.11
283 × 283	0.15
400 × 400	0.17
Maximum	0.19

Table 6.2: FaceVACS recognition score gain for multiple reconstruction sizes.

only comparison showed that the shape is not near the quality of laser scans and structured light models. The reconstructed 3D models are of high enough quality to correct the pose of the face, but cannot be used for shape comparison.

Figure 6.10: Multiple renderings of a reconstructed 3D face shape.

6.5 CONCLUSION

We explored the possibilities of a dense model-free 3D face reconstruction method, based on image sequences from a single camera, to improve the current state of forensic face comparison. The forensic context implies that models based on facial data cannot be used, because they cause a bias towards those faces. We proposed a dense 3D reconstruction method with two stages. In the first stage we obtain a coarse 3D landmark shape and an estimate of the rotations and translation in each frame. In the second stage we use the Lambertian reflectance model to estimate the albedo and to refine the 3D shape of the face. In our experiments the proposed multi-resolution approach is able to deliver frontal face reconstructions that improve the face comparison scores in more than 90% of the cases. Therefore we reached our goal of improving face recognition performance without introducing any bias towards a face model. Using the likelihood ratio framework, we show for the same experiment that for data initially unsuitable for forensic use, the reconstructions become meaningful in a forensic context in more than 60% of the cases. Visual inspection of the 3D shape shows that the reconstructed 3D shape with albedo texture can be used to generate faces under pose, but not for shape-based 3D face recognition.

6.6 ACKNOWLEDGEMENT

This work was supported by BZK 5.50: ‘Gezichtsvergelijking op basis van niet gekalibreerde Camerabeelden’ in a cooperation between the Netherlands Forensic Institute and University of Twente. We would like to thank Cognitec Systems GmbH. for supporting our research by providing the FaceVACS software. Results obtained for FaceVACS were produced in experiments conducted by the University of Twente, and should therefore not be construed as a vendor’s maximum effort full capability result.

6.7 PAPER CLOSURE

Finally, we managed to obtain a dense 3D reconstruction of the face including albedo texture. Although the quality of the shape estimation is not suitable for 3D shape comparison, we can still use this method to obtain frontal images of a face. The reconstruction process is model-free and is therefore suitable for embedding in a forensic context.

7

Conclusions

CHRIS VAN DAM

In this thesis we propose a complete method to reconstruct frontal faces from sequences of non-frontal images. The proposed reconstruction method consists of two parts, where we first obtain a coarse shape reconstruction, based on multiple landmarks in multiple images. In the second part we estimate the albedo of the face and obtain a dense 3D reconstruction of the shape simultaneously. No statistical facial models are used in the reconstruction process, the reconstruction is based on the input data only. The proposed reconstruction method is designed for usage in a forensic setting. In the next section we give our final conclusion on all the research questions, see Chapter 1, separately.

7.1 FINAL CONCLUSION

How can we extract 3D information from multiple images for forensic face recognition?

We reviewed the current procedure in forensic face recognition. Most of the forensic case work on face comparison is manually processed by a forensic researcher. The available face comparison software is optimized for frontal images and the performance degrades for face images under pose. This has led to a standard procedure where only one image from a whole image sequence is used for face comparison. After selection of the best image, the rest of the data is discarded, because there is no accepted procedure for merging face comparison result from multiple images. We start our research with the observation that all the images originate from the same 3D source and therefore are useful to obtain a maximum result. The best way to improve the forensic face comparison procedure is to incorporate multiple images in the reconstruction process to reconstruct a high quality frontal view of the face. There are, however, a few things to take into account, when working with forensic data. First of all the image quality of the data is lower than the current standard, because often old recording systems are not replaced and image compression techniques are used to reduce the space for data storage. Secondly, it is important that

the source of the data remains clear throughout the whole process. If no additional facial data is added in the reconstruction process, the reconstruction is unbiased and suitable for forensic facial recognition.

Which reconstruction method is the most suitable in forensic cases?

We reviewed the existing literature on 3D face reconstruction. There are many reconstruction methods around to reconstruct the shape of the face based on landmarks, motion, shading, silhouettes or face models. The most important selection criterion is the usability of the method in a forensic context. The reconstruction method should be bias-free, to guarantee a fair judgment in forensic cases. Although the Morphable Model approach is able to reconstruct 3D faces from single or multiple images with smooth visually pleasing reconstructions, the Morphable Model approach uses additional information from multiple faces which contaminates the data and is therefore unusable in a forensic context. Based on this criterion we selected a data driven reconstruction approach that needs a set of landmarks and images with different pose to reconstruct 3D faces. The main challenge using this approach is to reconstruct a dense 3D face model, based on a low number of landmarks.

What are the requirements for landmark based 3D reconstruction of random 3D objects?

Structure from motion methods use multiple landmarks in multiple images to reconstruct the shape of a face and estimate the rotation and translation of a face simultaneously. To determine the minimum requirements for the number of landmarks and frames, we performed several experiments on 2D projections of random point clouds. To increase the realism of the experiment we added Gaussian noise to the projections with varying strength. We evaluated the 2D reprojection error on the noise-free projections and the 3D error on the ground truth of the point clouds. We based the minimum requirements of landmarks and frames on the ability to cancel out the noise and on the stability of the 2D and 3D error. Although a larger number of landmarks increases the stability of the reconstruction method, we realized that in the forensic context only a limited number of landmarks is available. The experiments showed that a minimum of 20 landmarks in around 30 frames is required to obtain a 3D reconstruction of decent quality.

How can we perform landmark based reconstruction on rigid face data?

During the experiments with face like data, we discovered that a low number of landmarks necessitate calibration of the internal camera parameters to be able to distinguish variation in shape from variation in the field of view of the camera. We realized that reconstruction of uncalibrated face data is considerably complex and not feasible for low quality forensic data, therefore we assumed that it is possible to obtain the internal calibration parameters of the camera. Using the calibrated data we showed with experiments on a styrofoam head that the proposed reconstruction method is able to reconstruct a face-like structure with similar performance.

How can we obtain a coarse 3D face reconstruction using a realistic forensic case image sequence without introducing bias towards a model?

The proposed shape reconstruction algorithm is based on a sequence of images, which are manually landmarked. The initial shape estimate is based on a pair of frames with a suitable 2D reprojection error. In an iterative procedure multiple frames are added to improve the 3D shape estimate and the estimates of the rotation and translation parameters of the face in each frame. The low number of landmarks causes the optimization to be sensitive to outliers, therefore some

additional effort was made to prevent and correct outliers in the reconstruction process. After the reconstruction of the shape, we triangulate the reconstructed shape to obtain a surface for the texture of the face. Using the Lambertian illumination model we combine the texture information from multiple views by applying median filtering to the corrected images. The obtained 3D reconstruction is coarse and can only be used to render frontal reconstructions of a face. Experiments on a dataset of sequences of real face images showed that for low quality and non-frontal images, the face comparison scores increased considerably. With this complete reconstruction method the first step is made towards a working forensic face reconstruction method.

How can we improve the coarse reconstruction based on a realistic forensic case image sequence to obtain a dense 3D reconstruction?

The coarse shape reconstruction method of the forensic data is bias-free, but is unable to provide a dense reconstruction of the face. In the final proposed reconstruction method we revise the texture reconstruction method and incorporate a dense shape reconstruction method. The reconstruction method uses the proposed coarse shape reconstruction method as initialization. The dense reconstruction is optimized in an iterative procedure where the normals of the reconstructed face are used to estimate both the shape and the reflection coefficients of the face using the Lambertian illumination model. The reflection coefficients are used as texture for the reconstructed dense face shape. During the reconstruction process the resolution of the 3D reconstruction increases in each iteration. The final result is a dense 3D point cloud with an estimate of the reflection coefficient for each point on the face. The dense 3D reconstructions can be used to render frontal faces or faces under small pose. Since no statistical face models are used, the proposed reconstruction method is bias-free and is therefore suitable in a forensic face comparison process. The integrity of the forensic data is maintained during the entire reconstruction process.

7.2 FINAL CONCLUSION

In our research the aim was to make it feasible to use surveillance quality image sequences for face recognition in a forensic context. One of the major problems is that often in surveillance video recordings there are no frontal facial recordings of sufficient quality that can be used for facial comparison by state of the art facial recognition software. The solution we propose in this thesis is based on the extraction of 3D information of multiple images of the same subject from an image sequence and the reconstruction of a frontal facial image. A unique feature of our approach is that it is completely data driven. No statistical facial models are used in the reconstruction process. This is a great advantage in a forensic context, because the use of statistical models suggests that a bias may occur to subjects that are better represented by the model, something that should be avoided at all costs in forensic applications. We show that facial recognition is effectively improved using the reconstructed frontal facial images compared to the individual images from surveillance image sequences.

7.3 RECOMMENDATIONS

According to the experiments in Chapter 3 increasing the number of landmarks would increase the performance of the reconstruction process. To increase the number of landmarks in the proposed method, characteristic markings on the face like scars, freckles, wrinkles or birthmarks could be used as additional landmarks next to the current landmarks. Because these markings are person specific, determining the performance for unseen faces would become a challenging task. Since the proposed method is model free, new landmarks from markings can be added without any difficulties.

In Chapter 5 we introduced a coarse shape reconstruction method. The robustness of the proposed method depends on the choice of the initial pair of views and the order in which the frames are added. To add the frames one by one in an optimal way, each frame should be sufficient different from the other frames, but still have a large overlap with the previous frames. An improved automated approach to optimize the ordering would be beneficial for the stability of the reconstruction method. Another option would be to determine the order of the image sequence by a forensic examiner, which would help to maximize the chance of success during the reconstruction process.

One of the remaining open questions is how to handle uncalibrated image sequences. Taking the current quality of forensic case material into consideration, it is impossible with the proposed approach to distinguishing the difference between shape and variation in the field of view of the camera. However, if the quality of the images increases and therefore the number of landmarks increases, estimating the internal camera parameters might become feasible.

To limit the amount of work for the forensic researchers most of the manual work could benefit from semi-automated methods to determine the landmarks on the face and to find an optimal order for adding the frames. Visualization during the reconstruction process could help a forensic researcher to detect outliers in the data in an earlier stage. These outliers, either landmarks or frames, could be excluded from the input data, before restarting the reconstruction process. Although most of the techniques are currently available, developing the ultimate reconstruction tool is still a challenging task.

Appendix A

A.1 LANDMARKS

In this PhD thesis we use landmarks as facial features for the reconstruction process, see Figure A.1. Facial landmarks are predefined points on a face that describe a specific location on the face. Examples of landmarks are the tip of the nose and the corners of the eyes. These landmarks are highly distinctive and can be found using an automated approach. The position of other landmarks, such as the tip of the chin, are harder to define in images and automatic detection methods provide less accurate results here. Based on the forensic context, we assumed that the landmarks are manually annotated by a forensic researcher. Quantifying the accuracy of the manual landmarking process of the forensic researcher is beyond of the scope of this PhD thesis. We assume however that the accuracy of the manual labeling is higher than the accuracy of automatic approaches. Although someone could come up with many other features of a face, we choose landmarks as features because of their repeatability under scaling and change of resolution. Secondly, the number of landmarks can be adjusted according to the minimum resolution of an image sequence. Both of these aspects are important in a forensic context, where much variation of the image sequences is present.

Figure A.1: A face model consisting of landmarks only.

B

Appendix B

B.1 PROJECTION MODELS

A projection model is used to render a 3D model into a 2D image. The transformation from 3D to 2D is called projection. Each visible point in the 3D world is projected to a point on a 2D plane. Projection models are represented by projection matrices, which have a size of 3×4 . The projection matrix is used to convert homogenous 3D points into homogenous 2D points by matrix multiplication. The purpose of a projection matrix is to approximate the behavior of a real camera. Researchers have come forward with many projection models, where the most important factor is the number of variables used to determine a projection matrix. The easiest type of projection is the orthogonal projection, where the object is infinitely far from the image plane, leading to parallel projections of 3D points. There is no scaling involved, so the projections of the objects are as large as they appear in the 3D world, see Equation B.1.1.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (\text{B.1.1})$$

A slight variation of this model is the scaled orthogonal projection or weak projection model, where all points have the same distance to the camera, but can be scaled by a constant, see Equation B.1.2.

$$\begin{bmatrix} \alpha_1 & 0 & 0 & 0 \\ 0 & \alpha_2 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (\text{B.1.2})$$

A more general form is the affine projection model, involving for example rotation and translation, although this matrix has 9 variables, still no perspective effect can be modeled, see Equation

B.1.3.

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \\ \alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 \\ 0 & 0 & 0 & \alpha_9 \end{bmatrix} \quad (\text{B.1.3})$$

For objects far away from the camera with relatively small variation in depth, an affine projection model would suffice. However in the case of the ATM recordings the face is usually close to the camera and therefore the distance to the camera of each point on the face is of high importance for correctly estimating the position and 3D shape of the face. The pinhole camera model, where a 3D object is captured on a virtual image plane, is able to handle the perspective effect, see Figure B.1.

Figure B.1: The pinhole camera as can be found on wikipedia.

The pinhole camera uses a perspective projection model to project 3D points to the image plane, see equation B.1.4. Next to affine transformations also scaling, skewing and other effects are present.

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \\ \alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 \\ \alpha_9 & \alpha_{10} & \alpha_{11} & \alpha_{12} \end{bmatrix} \quad (\text{B.1.4})$$

Each 3D point converges in the camera center which leads to the perspective projection effect, see Figure B.2. Object closer to the camera appear to be bigger on the image plane. The size of the objects on the 2D plane is related to the depth of the 3D point. Even for faces close to a camera, the perspective projection model gives a realistic view on a 3D model.

More details about camera models or homogeneous coordinates can be found in for example^{31,90}. Recordings made with real cameras can severely suffer from lens distortion. This distortion can be corrected by calibrating the camera and determining the internal camera calibration

Figure B.2: A pinhole camera model using perspective projection. The 3D point Q is projected to the image plane in the 2D point q .

parameters. We used the Camera Calibration Toolbox for Matlab¹² to perform the calibration of the camera.

Bibliography

- [1] 3dMD Limited (2010). 3dMD Scanner. <http://www.3dmd.com/3dMDface/>.
- [2] Abate, A. F., Nappi, M., Riccio, D., & Sabatino, G. (2007). 2d and 3d face recognition: A survey. *Pattern Recognition Letters*, 28(14), 1885–1906.
- [3] Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '13* (pp. 3444–3451).: IEEE Computer Society.
- [4] Basso, C., Vetter, T., & Blanz, V. (2003). Regularized 3d morphable models. In *Higher-Level Knowledge in 3D Modeling and Motion Analysis, 2003*. (pp. 3–10).: Springer.
- [5] Bergen, J. & Hingorani, R. (1990). *Hierarchical motion-based frame rate conversion*. Technical Report NJ 08540, David Sarnoff Research Center Princeton.
- [6] Beumer, G., Tao, Q., Bazen, A., & Veldhuis, R. (2006). A landmark paper in face recognition. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0, 73–78.
- [7] Blanz, V. (2001). *Automatische Rekonstruktion der Dreidimensionalen Form van Gesichtern aus enem Einzelbild*. PhD thesis, Universität Tübingen, Germany.
- [8] Blanz, V., Grother, P., Phillips, P. J., & Vetter, T. (2005). Face recognition based on frontal views generated from non-frontal images. In *Computer Vision and Pattern Recognition* (pp. 454–461).
- [9] Blanz, V. & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques, SIGGRAPH '99* (pp. 187–194). New York, NY, USA: ACM Press/Addison-Wesley Publishing Co.
- [10] Blanz, V. & Vetter, T. (2003). Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9), 1063–1074.
- [11] Boom, B., Spreeuwens, L., & Veldhuis, R. (2009). Model-based illumination correction for face images in uncontrolled scenarios. In *CAIP '09: Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns* (pp. 33–40). Berlin, Heidelberg: Springer-Verlag.
- [12] Bouguet, J. Y. (2008). Camera calibration toolbox for Matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/.

- [13] Bowyer, K. W., Chang, K., & Flynn, P. (2004). A survey of approaches to three-dimensional face recognition. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 1* (pp. 358–361). Washington, DC, USA: IEEE Computer Society.
- [14] Bowyer, K. W., Chang, K., & Flynn, P. (2006). A survey of approaches and challenges in 3d and multi-modal 3d + 2d face recognition. *Computer Vision and Image Understanding*, 101(1), 1–15.
- [15] Bruce, V. & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, 77(3), 305–327.
- [16] Chowdhury, A. K. R. & Chellappa, R. (2003). Face reconstruction from monocular video using uncertainty analysis and a generic model. *Computer Vision and Image Understanding*, 91, 188–213. Special Issue on Face Recognition.
- [17] Cognitec Systems GmbH (2013). FaceVACS SDK 8.8.0. <http://www.cognitec-systems.de>.
- [18] Delaunoy, A. & Pollefeys, M. (2014). Photometric bundle adjustment for dense multi-view 3d modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1486–1493).
- [19] Dovgard, R. & Basri, R. (2004). Statistical symmetric shape from shading for 3d structure recovery of faces. In *European Conference on Computer Vision* (pp. 99–113).
- [20] Fidaleo, D. & Medioni, G. G. (2007). Model-assisted 3d face reconstruction from video. In S. K. Zhou, W. Zhao, X. Tang, & S. Gong (Eds.), *AMFG*, volume 4778 of *Lecture Notes in Computer Science* (pp. 124–138).: Springer.
- [21] Fooprateepsiri, R. & Kurutach, W. (2014). A general framework for face reconstruction using single still image based on 2d-to-3d transformation kernel. *Forensic science international*, 236, 117–126.
- [22] Fua, P. (2000). Regularized bundle-adjustment to model heads from image sequences without calibration data. *International Journal of Computer Vision*, 38(2), 153–171.
- [23] Garg, R., Roussos, A., & Agapito, L. (2013). Dense variational reconstruction of non-rigid surfaces from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1272–1279).
- [24] Gherardi, R. & Fusiello, A. (2010). Practical autocalibration. In *11th European Conference on Computer vision: Part I (ECCV '10)* (pp. 790–801).: Springer-Verlag.
- [25] Gouraud, H. (1971). Continuous shading of curved surfaces. *IEEE Transactions on Computers*, 20(6), 623–629.
- [26] Guo, H., Liu, C., & Zhang, L. (2005). An automatic method of building 3d morphable face model. In *Advanced in Biometric Person Authentication* (pp. 50–58).
- [27] Gupta, H., RoyChowdhury, A., & Chellappa, R. (2004). Contour-based 3d face modeling from a monocular video. In *Proc. BMVC* (pp. 1–10).
- [28] Hamsici, O. C., Gotardo, P. F. U., & Martinez, A. M. (2012). Learning spatially-smooth mappings in non-rigid structure from motion. In *Proceedings of the 12th European Conference on Computer Vision, ECCV '12* (pp. 260–273).: Springer-Verlag.

- [29] Hartley, R. I. (1993). An investigation of the essential matrix.
- [30] Hartley, R. I. (1997). Kruppa's equations derived from the fundamental matrix. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 133–135.
- [31] Hartley, R. I. & Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition.
- [32] Hassner, T. (2013). Viewing real-world faces in 3d. In *IEEE International Conference on Computer Vision (ICCV)* (pp. 3607–3614).
- [33] Heisele, B., Ho, P., Poggio, T., Ho, . P., & Poggio, . T. (2001). Face recognition with support vector machines: Global versus component-based approach. In *In Proc. 8th International Conference on Computer Vision* (pp. 688–694).
- [34] Ho, H. T. & Chellappa, R. (2013). Pose-invariant face recognition using markov random fields. *IEEE Transactions on Image Processing*, 22(4), 1573–1584.
- [35] Hu, Y., Jiang, D., Yan, S., Zhang, L., & Zhang, H. (2004a). Automatic 3d reconstruction for face recognition. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on* (pp. 843–848).
- [36] Hu, Y.-L., Yin, B.-C., Cheng, S.-Q., & Gu, C.-L. (2004b). An improved morphable model for 3d face synthesis. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics 2004*, volume 7 (pp. 4362–4367).
- [37] Ishimoto, M. & Chen, Y.-W. (2009). Pose-robust face recognition based on 3d shape reconstruction. In *Fifth International Conference on Natural Computation*, volume 6 (pp. 40–43).
- [38] Jiang, D., Hu, Y., Yan, S., Zhang, L., Zhang, H., & Gao, W. (2005). Efficient 3d reconstruction for face recognition. *Pattern Recognition*, 38(6), 787–798.
- [39] Jo, J., Choi, H., Kim, I.-J., & Kim, J. (2015). Single-view-based 3d facial reconstruction method robust against pose variations. *Pattern Recognition*, 48(1), 73–85.
- [40] Kalinkina, D., Gagalowicz, A., & Roussel, R. (2007). 3d reconstruction of a human face from images using morphological adaptation. In *MIRAGE 2007 (Computer Vision / Computer Graphics Collaboration Techniques and Applications)* INRIA Rocquencourt, France: Springer Verlag.
- [41] Kang, S. B. & Jones, M. (2002). Appearance-based structure from motion using linear classes of 3-d models. *International Journal of Computer Vision*, 49(1), 5–22.
- [42] Keller, M., Knothe, R., & Vetter, T. (2007). 3d reconstruction of human faces from occluding contours. In *Mirage* (pp. 261–273).
- [43] Kemelmacher-Shlizerman, I. & Seitz, S. M. (2011). Face reconstruction in the wild. In *Proceedings of the 2011 International Conference on Computer Vision (ICCV '11)* (pp. 1746–1753). Washington, DC, USA: IEEE Computer Society.
- [44] Krishnamurthy, V. & MarcLevoy (1996). Fitting smooth surfaces to dense polygon meshes. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques* (pp. 313–324). New York, NY, USA: ACM.

- [45] Kuo, C., Huang, R.-S., & Lin, T.-G. (2002). 3d facial model estimation from single front-view facial image. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(3), 183–192.
- [46] Le Zou; Cheng, S.; Zixiang Xiong; Mi Lu; Castleman, K. (2007). Facial feature extraction from range images using a 3d morphable model. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007* (pp. 241–244).
- [47] Lee, J., Moghaddam, B., Pfister, H., Machiraju, R., Lee, J., Moghaddam, B., Pfister, H., & Machiraju, R. (2003). Silhouette-based 3d face shape recovery. In *In Graphics Interface* (pp. 21–30).
- [48] Levine, M. D. & (Chris) Yu, Y. (2009). 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.
- [49] Levy, M., Trosset, M., & Kincaid, R. (2003). Quasi-newton methods for stochastic optimization. In *Uncertainty Modeling and Analysis, 2003. ISUMA 2003. Fourth International Symposium on* (pp. 304–309). Washington, DC, USA: IEEE Computer Society.
- [50] Liao, C.-T., Wang, S.-F., Lu, Y.-J., & Lai, S.-H. (2008). Video-based face recognition based on view synthesis from 3d face model reconstructed from a single image. In *ICME* (pp. 1589–1592).
- [51] Marques, M. & Costeira, J. (2008). 3d face recognition from multiple images: A shape-from-motion approach. In *FG* (pp. 1–6).
- [52] Meuwly, D. & Veldhuis, D. R. N. (2012). Forensic biometrics: From two communities to one discipline. In *International Conference of the Biometrics Special Interest Group, BIOSIG 2012* (pp. 1–12). Darmstadt, Germany: IEEE Conference publications.
- [53] Meuwly, D. & Veldhuis, R. N. J. (2014). Biometrics - developments and potential. In *Wiley Encyclopedia of Forensic Science*, volume 1 (pp. 1–8). Chichester, UK: John Wiley & Sons Ltd.
- [54] Meyer, C. D., Ed. (2000). *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics.
- [55] Park, U. & Jain, A. K. (2007). 3d model-based face recognition in video. In S.-W. Lee & S. Li (Eds.), *Advances in Biometrics*, volume 4642 of *Lecture Notes in Computer Science* (pp. 1085–1094). Springer Berlin Heidelberg.
- [56] Patrick J. Grother, George W. Quinn, P. J. P. (2010). *Report on the Evaluation of 2D Still-Image Face Recognition Algorithms*. Technical report, NIST Interagency. Internal Report 7709 - MBE 2010.
- [57] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009a). A 3d face model for pose and illumination invariant face recognition. In *AVSS '09: Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance* (pp. 296–301). Washington, DC, USA: IEEE Computer Society.
- [58] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009b). Basel face model. <http://faces.cs.unibas.ch/>.

- [59] Pighin, F., Szeliski, R., & Salesin, D. H. (2002). Modeling and animating realistic faces from images. *International Journal of Computer Vision*, 50(2), 143–169.
- [60] Romdhani, S. (2005). *Face Image Analysis using a Multiple Features Fitting Strategy*. PhD thesis, Universität Basel.
- [61] Romdhani, S., Blanz, V., & Vetter, T. (2002). Face identification by fitting a 3d morphable model using linear shape and texture error functions. In *Computer Vision – ECCV 2002* (pp. 3–19).: Springer.
- [62] Romdhani, S., Ho, J., Vetter, T., & Kriegmann, D. (2006). Face recognition using 3d models: Pose and illumination. In *Proceedings of the IEEE*, volume 94 (pp. 1977–1999).
- [63] Romdhani, S. & Vetter, T. (2003). Efficient, robust and accurate fitting of a 3d morphable model. In *IEEE International conference on Computer Vision 2003*.
- [64] Romdhani, S. & Vetter, T. (2005). Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2* (pp. 986–993). Washington, DC, USA: IEEE Computer Society.
- [65] Roth, J., Tong, Y., & Liu, X. (2015). Unconstrained 3d face reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [66] Roy-Chowdhury, A. K. (2002). 3d face reconstruction from video using a generic model. In *In International Conference on Multimedia and Expo* (pp. 449–452).: Society Press.
- [67] Roy-Chowdhury, A. K., Chellappa, R., & Gupta, H. (2005). *3D Face Modeling From Monocular Video Sequences*, chapter 6, (pp. 185–218). Academic Press.
- [68] Roy-Chowdhury, A. K. & Xu, Y. (2007). *Pose and Illumination Invariant Face Recognition Using Video Sequences*, chapter 2, (pp. 9–25). Springer Berlin Heidelberg.
- [69] Scheenstra, A., Ruifrok, A., & Velkamp, R. C. (2005). A survey of 3d face recognition methods. In *In Lecture Notes in Computer Science* (pp. 891–899).: SpringerVerlag.
- [70] SCS (2015). PV3D Dataset. <http://scs.ewi.utwente.nl>.
- [71] Shan, Y., Liu, Z., & Zhang, Z. (2001). Model-based bundle adjustment with application to face modeling. *IEEE International Conference on Computer Vision*, 2, 644–651.
- [72] Sim, T. & Kanade, T. (2001). Combining models and exemplars for face recognition: An illuminating example. In *Proceedings of the CVPR 2001 Workshop on Models versus Exemplars in Computer Vision*.
- [73] Spreuwers, L. (2008). Multi-view passive 3d face acquisition device. In *Proceedings of the Special Interest Group on Biometrics and Electronic Signatures*, volume P-137 of *Lecture Notes in Informatics (LNI) - Proceedings* (pp. 13–24). Bonn: Gesellschaft fuer Informatik.
- [74] Strecha, C., Fransens, R., & Van Gool, L. (2006). Combined depth and outlier estimation in multi-view stereo. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2 (pp. 2394–2401).
- [75] Sturm, P. & Triggs, B. (1996). A factorization based algorithm for multi-image projective structure and motion. In *4th European Conference on Computer Vision (ECCV'96)* (pp. 709–720).: Springer-Verlag.

- [76] Sun, Y., Wang, X., & Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13 (pp. 3476–3483).: IEEE Computer Society.
- [77] Suwajanakorn, S., Kemelmacher-Shlizerman, I., & Seitz, S. (2014). Total moving face reconstruction. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision - ECCV 2014*, volume 8692 of *Lecture Notes in Computer Science* (pp. 796–812). Springer International Publishing.
- [78] Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [79] Torr, P. H. S. & Zisserman, A. (2000). MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78, 138–156.
- [80] Torresani, L., Hertzmann, A., & Bregler, C. (2008). Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(5), 878–892.
- [81] Triggs, B. (1997). Autocalibration and the absolute quadric. In *Conference on Computer Vision and Pattern Recognition (CVPR '97)*(pp. 609–614). Washington, DC, USA: IEEE Computer Society.
- [82] van Dam, C., Spreeuwers, L., & Veldhuis, R. (2012). Towards 3d facial reconstruction from uncalibrated cctv footage. In *33rd WIC Symposium on Information Theory in the Benelux* (pp. 228–234). Enschede, the Netherlands: Werkgemeenschap voor Informatie- en Communicatietheorie, WIC.
- [83] van Dam, C., Spreeuwers, L., & Veldhuis, R. (2013a). Landmark-based model-free 3d face shape reconstruction from video sequences. In A. Brömme & C. Busch (Eds.), *2013 BIOSIG*, Proceedings of the International Conference of Biometrics Special Interest Group (pp. 265–272).: Gesellschaft für Informatik e.V.
- [84] van Dam, C., Spreeuwers, L., & Veldhuis, R. (2013b). Model-free 3d face shape reconstruction from video sequences. In *34th WIC Symposium on Information Theory in the Benelux*: WIC.
- [85] van Dam, C., Spreeuwers, L., & Veldhuis, R. (2015). Face reconstruction from image sequences for forensic face comparison. *IET biometrics*, 5(2), 140–146.
- [86] van Dam, C., Spreeuwers, L., & Veldhuis, R. (2017). Model-free 3d face reconstruction from video for forensic facial comparison. Submitted.
- [87] van Dam, C., van Rootseler, R., Spreeuwers, L., & Veldhuis, R. (2010). Reconstruction of 3d face models from video and image sequences: An overview. Internal report, SAS-05-10.
- [88] van Rootseler, R., Spreeuwers, L., & Veldhuis, R. (2011). Application of 3d morphable models to faces in video images. In *Proceedings of the 32nd Symposium on Information Theory in the Benelux (WIC11)*(pp. 34–41). Delft, the Netherlands: Werkgemeenschap voor Informatie- en Communicatietheorie.

- [89] Veldhuis, R. (2015). The relation between the secrecy rate of biometric template protection and biometric recognition performance. In *Biometrics (ICB), 2015 International Conference on* (pp. 311–318).
- [90] Wang, G. & Wu, J. Q. (2010). *Guide to Three Dimensional Structure and Motion Factorization*. Springer Publishing Company, Incorporated, 1st edition.
- [91] Wang, S., Zhang, L., & Samaras, D. (2005). Face reconstruction across different poses and arbitrary illumination conditions. In *AVBPA* (pp. 91–101).
- [92] Wang, S.-F. & Lai, S.-H. (2008). Reconstructing 3d shape, albedo and illumination from a single face image. *Comput. Graph. Forum*, 27(7), 1729–1736.
- [93] Widanagamaachchi, W. & Dharmaratne, A. (2008). 3d face reconstruction from 2d images. In *DICTA '08: Digital Image Computing: Techniques and Applications* (pp. 365–371).
- [94] Xiong, X. & Torre, F. D. L. (2013). Supervised descent method and its applications to face alignment. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '13* (pp. 532–539).
- [95] Xu, Y. & Roy-Chowdhury, A. K. (2005). Integrating the effects of motion, illumination and structure in video sequences. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision* (pp. 1675–1682). Washington, DC, USA: IEEE Computer Society.
- [96] Yang, C., Chen, J., Xia, C., Liu, J., & Su, G. (2013). A sfm-based sparse to dense 3d face reconstruction method robust to feature tracking errors. In *IEEE International Conference on Image Processing (ICIP)* (pp. 3617–3621).
- [97] Zhang, R., Tsai, P.-S., Cryer, J., & Shah, M. (1999). Shape-from-shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8), 690–706.
- [98] Zhang, X. & Gao, Y. (2009). Generalised ambient reflection models for lambertian and phong surfaces. In *International Conference on Image Processing* (pp. 3993–3996).
- [99] Zhao, W. & Chellappa, R. (2000). Sfs based view synthesis for robust face recognition. In *FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000* (pp. 285–292). Washington, DC, USA: IEEE Computer Society.
- [100] Zhou, S. K. & Chellappa, R. (2005). Beyond a single still image: Face recognition from multiple still images and videos.

Summary

In this thesis we explore how multiple images from a sequence which individually are considered not usable for a forensic procedure, can be combined to reconstruct a face model that is usable in a forensic face comparison procedure. We started by examining the current procedure of forensic face recognition. In the current forensic procedure usually only one image from an entire image sequence is used for face comparison, whereas the remaining images are discarded. We started with the observation that all the images originate from the same 3D face and therefore contain useful information of the face. The best way to improve the current forensic face comparison procedure is to incorporate multiple images in the reconstruction process to reconstruct a high quality frontal view of the face. We reviewed the existing literature on 3D face reconstruction and found two different ways of reconstructing 3D faces. The first approach is a model driven approach, involving a Morphable Model, which is able to construct 3D faces from single or multiple images. Such an approach would be a considerable advance in forensic face recognition. Unfortunately the model driven approach uses information from external face models which biases the reconstruction towards the average face and is therefore unsuitable in a forensic context. The second approach is a data driven approach, which needs a considerable number of landmarks and images with different pose to reconstruct 3D faces. However, the quality of the images in many forensic cases is low and only a few landmarks can be detected, which makes the reconstruction a challenging task. We concluded that none of the existing face reconstruction methods currently support an unbiased reconstruction method suitable for usage in a forensic context. Based on the available literature and the forensic setting, we chose a structure from motion approach based on landmarks, which seemed the most suitable method for forensic face recognition. The structure from motion method uses landmarks in multiple images to reconstruct the shape of the face and estimates the rotation and translation of the face simultaneously. We performed experiments on random point clouds to determine the minimum number of views and landmarks needed to obtain a proper 3D reconstruction. Experiments on random point clouds show that a minimum of 20 landmarks in around 30 frames is required to obtain a coarse 3D reconstruction. Secondly, we discovered that a low number of landmarks necessitates calibration of the internal camera parameters to be able to distinguish shape variation from variation in the field of view of the camera. In follow-up experiments with rigid face data we showed that it is possible to reconstruct face-like shapes with similar performance. In the proposed shape reconstruction algorithm, the initial reconstruction is based on a pair of frames with a suitable 2D reprojection error. In an iterative procedure multiple frames were added to improve the 3D reconstruction and the estimates of the rotation and translation of the face in each frame. The low number of landmarks caused the optimization to be sensitive to outliers. After the reconstruction of the shape, we triangulated the landmark model to obtain a surface for the texture of the face. Based on the Lambertian illumination model, we corrected

and combined the texture from multiple views. The obtained 3D reconstruction is coarse and can be used to render frontal reconstructions of a face. Face recognition experiments showed that the reconstructed frontal faces already outperformed the original non-frontal images. In the final proposed method we revised the reconstruction of the texture and incorporated a dense shape reconstruction method into the proposed reconstruction method. The dense reconstruction method used the coarse shape reconstruction method as initialization, and creates a dense 3D face reconstruction. The dense reconstruction method is based on an iterative procedure where the normals of the reconstructed face are used to optimize both the shape and the texture of the face. The reconstructed dense 3D models can be used to render frontal faces or faces under small pose. Additional recognition experiments showed that the reconstructed frontal faces outperformed the original non-frontal images in most of the cases. The proposed reconstruction method is unbiased by design and is therefore suitable in a forensic face comparison process.

Samenvatting

In dit proefschrift onderzoeken we hoe meerdere plaatjes die op zichzelf niet bruikbaar zijn voor forensische gezichtsvergelijking, samengevoegd kunnen worden zodat er een gezichtsmodel ontstaat dat wel bruikbaar is binnen een forensische gezichtsvergelijkingsprocedure. We beginnen met het beschrijven van de huidige procedure voor forensische gezichtsvergelijking. In de huidige forensische procedure wordt vaak een enkel plaatje geselecteerd uit een hele reeks plaatjes. De overige plaatjes worden doorgaans niet gebruikt. We vervolgen met de observatie dat alle plaatjes van dezelfde 3D bron afstammen en daarom dus ook bruikbare informatie van het gezicht bevatten. De beste manier om de huidige gezichtsvergelijkingsprocedure te verbeteren is dan ook door in een reconstructieproces meerdere plaatjes te gebruiken om een frontale kwaliteitsafbeelding van een gezicht te reconstrueren. We onderzoeken de bestaande literatuur op het gebied van 3D reconstructie en vinden twee verschillende manieren om gezichten te reconstrueren. De eerste methode is gebaseerd op een model van het gezicht en maakt gebruik van een Morphable Model om gezichten uit een of meerdere afbeeldingen te kunnen reconstrueren. Zo'n methode zou een vooruitgang betekenen voor de forensische gezichtsvergelijking. Helaas gebruikt deze methode allerlei informatie uit externe bronnen, die ervoor zorgen dat de reconstructie meer dan gebruikelijk naar het gemiddelde gezicht getrokken wordt. Daarom is een dergelijke methode niet geschikt in forensische toepassingen. De tweede methode is een methode die enkel gebaseerd is op de data. Deze methode heeft een behoorlijk aantal herkenningpunten op het gezicht nodig en een behoorlijk aantal afbeeldingen met verschillende poses van het gezicht om een 3D-reconstructie te kunnen maken. Helaas is de kwaliteit van de afbeeldingen in forensische zaken vaak niet erg hoog en kunnen slechts enkele herkenningpunten op het gezicht worden gevonden, wat het reconstructieproces uitdagend maakt. We concluderen dat geen van de bestaande reconstructiemethodes een reconstructie oplevert die ofwel geen voorkeur heeft voor bepaalde gezichten danwel bruikbaar is in een forensische procedure. Op basis van de beschikbare literatuur kiezen we een structuur-uit-beweging-methode, die gebruikt maakt van herkenningpunten op het gezicht. Deze methode lijkt het meest geschikt binnen een forensische omgeving. De structuur-uit-beweging-methode gebruikt herkenningpunten in meerdere afbeeldingen om tegelijkertijd de vorm van een gezicht en de rotatie en verschuiving van dat gezicht te kunnen reconstrueren. We voeren enkele experimenten uit op willekeurige puntenwolken om het minimum aantal afbeeldingen en herkenningpunten te bepalen dat een geschikte reconstructie oplevert. De experimenten op de willekeurige puntenwolken tonen aan dat er een minimum van 20 herkenningpunten in 30 afbeeldingen nodig is om een ruwe reconstructie te maken. Ten tweede ontdekken we dat het lage aantal herkenningpunten ons ertoe dwingt om de interne parameters van de camera te bepalen om zo de variatie in vorm te kunnen onderscheiden van de variatie van het gezichtsveld van de camera. In vervolgsperimenten met statische gezichtsdata laten we zien dat het mogelijk is om op een gezicht

lijkende structuren te kunnen reconstrueren met vergelijkbare uitkomsten. In het voorgestelde reconstructiealgorithme van de vorm van het gezicht, is de initiële reconstructie gebaseerd op één paar van de afbeeldingen dat een geschikte 2D-herprojectiefout oplevert. In een iteratieve procedure worden daarna meerdere afbeeldingen toegevoegd om de reconstructie van de vorm en de schattingen van de rotatie en verplaatsing in elke afbeelding te verbeteren. Door het lage aantal herkenningspunten is deze optimalisatiemethode gevoelig voor uitschieters van de parameters. Nadat de vorm van het gezicht gereconstrueerd is, voegen we een raster van driehoeken toe tussen de herkenningspunten om een oppervlak te creëren voor de textuur van het gezicht. Met behulp van het Lambertiaanse belichtingsmodel corrigeren en combineren we de textuur uit meerdere afbeeldingen. De zojuist verkregen 3D-reconstructie is een ruwe reconstructie van het gezicht en kan daarom alleen gebruikt worden om een frontale afbeelding van het gezicht te genereren. In gezichtsherkenningsexperimenten tonen we aan dat de gereconstrueerde frontale gezichten vaak beter presteren dan de oorspronkelijke afbeeldingen. In de uiteindelijke voorgestelde methode herzien we de methode om de textuur te combineren en introduceren we een reconstructiemethode die een nauwkeurigere reconstructie maakt van de vorm van een gezicht. Deze reconstructiemethode maakt gebruik van de ruwe reconstructie om een nauwkeurigere reconstructie te maken. De nauwkeurigere reconstructiemethode maakt gebruik van een iteratieve methode waarin de normalen van het gereconstrueerde gezicht gebruikt worden om zowel de vorm als de textuur van het gezicht te optimaliseren. De gereconstrueerde gezichtsmodellen kunnen nu gebruikt worden om zowel frontale gezichten als gezichten onder een kleine hoek te genereren. Extra experimenten tonen aan dat de gereconstrueerde gezichten bijna in alle gevallen een verbetering opleveren ten opzichte van de niet frontale afbeeldingen. De voorgestelde reconstructiemethode heeft door het ontwerp geen voorkeur voor bepaalde gezichten en kan daarom in een forensische procedure gebruikt worden.

Acknowledgments

CHRIS VAN DAM

In 2009 I started my PhD at the University of Twente in the Signals & Systems group. The SAS group was situated in the Hogekamp building. I still remember the nice view over the surroundings especially during winter time and the extreme hot summer time. After the move of the SAS group to the Carré building - the building with four corners but definitely not squared - the view was less, but the climate control was much better. During this time many colleagues obtained their doctorate. Almar, Bas, Dirk-Jan, René, Rita, Sanja, Julia, Gerbert, Anne, Xiaoying, thank you all for the good time, the funny PhD movies we produced and for showing how to finish your PhD. I would like to thank all the roommates I had over the years: Robin, Pinar, Jen-Hsuan, Tauseef, Abhishek, Chanjuan, Meiru, Yuxi, Chris 2, sorry I was first. Thank you for the nice time I had at the University and for all the inspirational discussions and fun we had. I also thank the foreign PhD students for introducing us to many different cultures. As one of the few Dutch, I tried to teach you something about The Netherlands. Sometimes a bit of advise, a recipe for a nice Dutch meal, like *kroket* and *frikandel* or some nice candies, like *drop*. Some parts of our culture are a success, other aspects still need to be appreciated more in the future. I would like to thank Berk, Pinar, Jen-Hsuan and Wei-Shu for the tasteful dinners we had and the fun times playing Dixit, and yes I won! Also a special thanks to the technical support I received from Henny and Geert Jan for the recording using Ultra Violet light. Geert Jan created the hardware trigger device v2.0, for the set-up in Figure 5.1. Henny took care of the safety regulations in the experiment, so that nobody would get hurt. My next thankfulness goes to the secretaries that took care of most of the paperwork: Anneke, Sandra, Bertine & Suze. I would like to thank Professor Slump for giving me the opportunity to teach other students the art of image processing. Thanks to all the contributors of the PV3D project: Robin, Bas, Ferdi, Arnout (NFI) for their input and help during this project. A special thanks goes to my supervisors over the years: Raymond and Luuk. Always there to help, advise and guide me through the research. Finally I would like to thank my family, friends who supported me, and helped me finishing my PhD. A special thanks and love for my wife Ingeborg, who invested her time, effort and love in supporting me. The second last greetings fly out to my children, Noa and Ruben, who enlightened me and gave me joy and a bit of distraction during the course of my PhD. And last but not least I would like to thank God for His inspiration and love which helped me finishing this PhD.