High Quality Dynamic Reflectance and Surface Reconstruction from Video

Naveed Ahmed

Max-Planck-Institut für Informatik
Saarbrücken, Germany

Dissertation zur Erlangung des Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing)
der Naturwissenschaftlich-Technischen Fakultät I
der Universität des Saarlandes

delivered by
EUROGRAPHICS
DIGITAL LIBRARY
www.eg.org
diglib.eg.org

Datum des Kolloquiums — Date of Defense
10. Juli 2009 — July 10th, 2009

...

Dekan — Dean
Prof. Dr. Joachim Weickert
Universität des Saarlandes, Saarbrücken, Germany

Gutachter — Reviewers
Prof. Dr. Hans-Peter Seidel
MPI Informatik, Saarbrücken, Germany

Prof. Dr. Christian Theobalt
Stanford University, USA

Prof. Dr. Gabriel Brostow
University College London, United Kingdom

Naveed Ahmed
Max-Planck-Institut für Informatik
Campus E1 4
66123 Saarbrücken, Germany
`nahmed@mpi-inf.mpg.de`

Dedicated to the wonderful land and people of Deutschland.
Thank you for five amazing years!

Abstract

The creation of high quality animations of real-world human actors has long been a challenging problem in computer graphics. It involves the modeling of the shape of the virtual actors, creating their motion, and the reproduction of very fine dynamic details. In order to render the actor under arbitrary lighting, it is required that reflectance properties are modeled for each point on the surface. These steps, that are usually performed manually by professional modelers, are time consuming and cumbersome.

In this thesis, we show that algorithmic solutions for some of the problems that arise in the creation of high quality animation of real-world people are possible using multi-view video data. First, we present a novel spatio-temporal approach to create a personalized avatar from multi-view video data of a moving person. Thereafter, we propose two enhancements to a method that captures human shape, motion and reflectance properties of a moving human using eight multi-view video streams. Afterwards we extend this work, and in order to add very fine dynamic details to the geometric models, such as wrinkles and folds in the clothing, we make use of the multi-view video recordings and present a statistical method that can passively capture the fine-grain details of time-varying scene geometry. Finally, in order to reconstruct structured shape and animation of the subject from video, we present a dense 3D correspondence finding method that enables spatio-temporally coherent reconstruction of surface animations directly from multi-view video data.

These algorithmic solutions can be combined to constitute a complete animation pipeline for acquisition, reconstruction and rendering of high quality virtual actors from multi-view video data. They can also be used individually in a system that require the solution of a specific algorithmic sub-problem. The results demonstrate that using multi-view video data it is possible to find the model description that enables realistic appearance of animated virtual actors under different lighting conditions and exhibits high quality dynamic details in the geometry.

Kurzfassung

Die Entwicklung hochqualitativer Animationen von menschlichen Schauspielern ist seit langem ein schwieriges Problem in der Computergrafik. Es beinhaltet das Modellieren einer dreidimensionaler Abbildung des Akteurs, seiner Bewegung und die Wiedergabe sehr feinen dynamischen Details. Um den Schauspieler unter einer beliebigen Beleuchtung zu rendern, müssen auch die Reflektionseigenschaften jedes einzelnen Punktes modelliert werden. Diese Schritte, die gewöhnlich manuell von Berufsmodellierern durchgeführt werden, sind zeitaufwendig und beschwerlich.

In dieser These schlagen wir algorithmische Lösungen für einige der Probleme vor, die in der Entwicklung solch hochqualitativen Animationen entstehen. Erstens präsentieren wir einen neuartigen, räumlich-zeitlichen Ansatz um einen Avatar von Mehransicht-Videodaten einer bewegenden Person zu schaffen. Danach beschreiben wir einen videobasierten Modelierungsansatz mit Hilfe einer animierten Schablone eines menschlichen Körpers. Unter Zuhilfenahme einer handvoll synchronisierten Videoaufnahmen berechnen wir die dreidimensionale Abbildung, seine Bewegung und Reflektionseigenschaften der Oberfläche. Um sehr feine dynamische Details, wie Runzeln und Falten in der Kleidung zu den geometrischen Modellen hinzuzufügen, zeigen wir eine statistische Methode, die feinen Details der zeitlich variierenden Szenegeometrie passiv erfassen kann. Und schließlich zeigen wir eine Methode, die dichte 3D Korrespondenzen findet, um die strukturierte Abbildung und die zugehörige Bewegung aus einem Video zu extrahieren. Dies ermöglicht eine räumlich-zeitlich zusammenhängende Rekonstruktion von Oberflächenanimationen direkt aus Mehransicht-Videodaten.

Diese algorithmischen Lösungen können kombiniert eingesetzt werden, um eine Animationspipeline für die Erfassung, die Rekonstruktion und das Rendering von Animationen hoher Qualität aus Mehransicht-Videodaten zu ermöglichen. Sie können auch einzeln in einem System verwendet werden, das nach einer Lösung eines spezifischen algorithmischen Teilproblems verlangt. Das Ergebnis ist eine Modelbeschreibung, das realistisches Erscheinen von animierten virtuellen Schauspielern mit dynamischen Details von hoher Qualität unter verschiedenen Lichtverhältnissen ermöglicht.

Summary

Creating high quality animations of virtual human actors has long been a focus of research in computer graphics. In the past decade, a variety of methods have been proposed that could estimate the motion of a performer and animate a model accordingly. Nevertheless, it is still very taxing to estimate the surface material properties so that the virtual actor can be rendered under arbitrary lighting conditions. It is also very difficult to obtain a spatio-temporally coherent surface representation of an animated model directly from multi-view video. Finally, transferring dynamic geometry detail from a real world actor to a virtual avatar is a very challenging problem in itself.

Previous methods for material and surface detail reconstruction were primarily geared towards reconstruction of static scene geometry. All the methods start with the acquisition of images of the object using still cameras. In contrast to still cameras, resolution of video cameras is still extremely low, which hampers the development of algorithmic solutions for dynamic scenes. Moreover, algorithms for video need to consider the additional temporal domain, which makes the development of the solutions even more challenging. With the advent of high resolution video cameras, solving the above mentioned problems in the video domain has not only become feasible, but it also has opened the possibility to solve the reconstruction problems in a spatio-temporally coherent way.

In this thesis, we demonstrate that using multi-view video data we can extract all necessary information that is required for the reconstruction of high quality 3D human animation from video.

We start with a novel spatio-temporal approach to create a personalized avatar from multi-view video data of a moving person. The avatar's geometry is generated by shape adapting a template human body model. Its surface texture is assembled from multi-view video frames showing arbitrary different body poses. The generated static texture can be used to render the complete human animation with just a single texture. This model description, an animated template geometry and a surface texture, is ideal to use in multi-user virtual environments where real-world people interact via digital avatars. The resulting avatars of humans exhibit true shape and photo-realistic appearance.

Free-viewpoint or 3D video allows the photo-realistic rendering of the virtual human from novel viewpoints. Recently the concept is extended to relightable free-viewpoint video that can also be rendered under arbitrary lighting. The relightable free-viewpoint videos are reconstructed using synchronized multi-view video streams that are recorded under calibrated lighting conditions. We make use

of the earlier work in this area, and using the same multi-view video data, present two methods that result in higher quality of relightable free-viewpoint video. First, we propose a solution for improving spatio-temporal texture registration, which is necessary for the accurate measurement of the surface reflectance properties. Additionally, a method to reduce the bias in the estimated surface reflectance is proposed to get as good as possible realistic renditions under arbitrary lighting conditions. The resulting model description enables us to faithfully reproduce the appearance of animated virtual actors under different simulated lighting conditions.

Models used in the reconstructed human animations, either animated templates or reconstructed directly from the video, do not depict high quality dynamic details that are visible in the clothing of the actor. Adding these dynamic details manually is a very complex process. Full body laser scanners can capture very fine quality details of the model, but unfortunately they are also static and look baked on the surface when those models are used for animation. We propose a statistical method that can capture highly-detailed dynamic surface geometry of humans from multi-view video streams under calibrated lighting even in the presence of measurement uncertainties. The output is a complete moving model of the human actor that features subtle dynamic geometry detail, such as wrinkles and folds in clothing.

Using an animated template model has its own benefits and drawbacks. It guarantees spatio-temporal coherence, but as the model has to be deformed for each frame to match the shape and size of the actor in the input video frame, the accuracy of the model with respect to the original actor is compromised. A better option would be to reconstruct the model directly from the video data, thus optimizing the consistency between the model and the actor. It is possible to reconstruct a mesh from each frame of the video. The obvious problem with this solution is that the reconstruction from each frame results in meshes with different connectivity. Ideally, one would like to create a spatio-temporally coherent animation from the individual reconstructions. To bridge this gap, we present a spatio-temporal dense 3D correspondence finding method from multi-view video data that enables the reconstruction of spatio-temporally coherent dynamic 3D geometry from a sequence of unrelated meshes.

Each of the algorithmic solutions can be used independently, as per the requirement of some specific system. Moreover, they can also be used together and combined in a single system resulting in an animation pipeline that can reconstruct and render very high quality animation of virtual actors from multi-view video data.

Zusammenfassung

Die Erzeugung hochqualitativer Animationen von virtuellen menschlichen Darstellern ist seit langem ein Schwerpunkt in der Forschung im Bereich Computergrafik. Im vergangenen Jahrzehnt wurde eine Vielzahl von Methoden vorgestellt, welche die Bewegung eines Akteurs abschätzen und ein Modell entsprechend animieren können. Gleichwohl ist es immer noch anspruchsvoll die Materialeigenschaften der Oberfläche einzuschätzen, sodass ein virtueller Charakter unter beliebigen Beleuchtungverhältnissen dargestellt werden kann. Es ist ebenfalls sehr schwierig von einem Multi-View-Video eine räumlich und zeitlich zusammenhängende Darstellung eines animierten Modells zu erhalten. Schließlich, stellt die Übertragung der Details dynamischer Geometrie von einem echten Schauspieler auf einen virtuellen Avatar selbst eine große Herausforderung dar.

Bisherige Vorgehensweisen im Bereich der Rekonstruktion von Material und Oberflächendetails zielten hauptsächlich auf statische Geometrie ab. Normalerweise, beginnen alle Methoden mit dem Erfassen eines Bildes des Objekts, mithilfe einer Fotokamera. Im Gegensatz zu Fotokameras ist die Auflösung von Videokameras immer noch extrem niedrig, was die Entwicklung algorithmischer Lösungen für dynamische Szenen erschwert. Darüberhinaus müssen Videoalgorithmen die zusätzlichen zeitlichen Komponente berücksichtigen, was das Erarbeiten von Lösungen noch komplizierter macht. Das Aufkommen hochauflösender Videokameras hat nicht nur eine Lösung der oben genanten Probleme für Videos ermöglicht sondern hat auch das Lösen der Rekonstruktionsprobleme auf räumlich und zeitlich zusammenhängende Art möglich gemacht.

In dieser Arbeit werden algorithmische Lösungen für vier spezielle Probleme präsentiert:

Wir beginnen mit einem neuartigen räumlich-zeitlichen Ansatz um einen individuellen Avatar auf der Basis von Multiview Videodaten einer sich bewegenden Person zu erzeugen. Die Gestalt des Avatars wird durch die Anpassung der Form einer Vorlage für menschliche Körper erhalten. Seine Oberflächentextur wird zusammengesetzt aus mehreren Multi-View-Video-Frames die beliebige verschiedene Posen beinhalten. Die so erhaltene statische Textur kann dazu benutzt werden die gesamte Animation mit einer einzigen Textur darzustellen. Diese Modelbeschreibung gemeinsam mit einer animierten Geometrievorlage und einer Oberflächentextur sind ideal um in einer virtuellen Multi-User-Umgebung in der echte Menschen durch digitale Avatare miteinander interagieren eingesetzt zu werden. Die Resultate für menschliche Avatars zeichnen sich durch eine wahrheitsgetreue Form und einen fotorealistischen Gesamteindruck aus, was durch die Rekonstruktion von Fotos einzelner Posen nicht möglich gewesen wäre.

3D videos erlauben die fotorealistische Darstellung des virtuellen Menschen aus neuen Blickwinkeln. Um ihn korrekt unter verschiedenen Beleuchtungen darstellen zu können, müssen auch die Reflexionseigenschaften seiner Oberfläche bekannt sein. Wir beschreiben einen Ansatz um diese abschätzen zu können. Dieser benutzt eine animierte Vorlage für menschliche Körper die gleichzeitig Gestalt, Bewegung und sich räumlich verändernde Reflexionseigenschaften durch wenige synchronisierte Multi-View-Videoaufnahmen erfasst. Wir stellen auch eine Lösung vor um die Registrierung räumlich und zeitlich veränderlicher Texturen zu verbessern. Das ist notwendig um eine genaue Messung der Reflexionseigenschaften der Oberfläche zu gewährleisten. Darüberhinaus, zeigen wir eine Methode, die den systematischen Fehler in der Schätzung der Oberflechenreflexion reduziert um möglichst relaistische Darstellung unter beliebigen Beleuchtungsverhältnissen zu erzielen. Die daraus resultierende Modelbeschreibung ermöglicht die originalgetreue Erscheinung virtueller Akteure unter verschiedenen simulierten Beleuchtungen.

Modelle die zur Rekonstruktion menschlicher Bewegungen, seien es animierte Vorlagen oder solche direkt von Videos, beschreiben nicht die hochqualitativen dynamischen Details der Kleidung des Darstellers. Diese dynamischen Details von Hand hinzuzufügen ist ein sehr komplexer Vorgang. Ganzkörper Laserscanner sind in der Lage sehr feine Details des Modells zu erfassen, aber diese sind leider auch statisch und wirken künstlich auf der Oberfläche wenn solche Modelle für Animationen genutzt werden. wir stellen eine statistische Methode vor die detailreiche dynamischer menschliche Oberflächengeometrie von mehreren Videoaufnahmen unter kalibrierten Beleuchtungen erfassen kann, sogar bei eventuell vorhandenen Messungenauigkeiten. Das Ergebnis ist ein komplett bewegliches Modell eines Menschlichen Schauspielers das selbst kleinste dynamische Details der Geometrie, wie zum Beispiel Falten auf der Kleidung, aufweist.

Eine animierte Vorlage zu benutzen hat seine Vor- und Nachteile. Es garantiert räumliche und zeitliche Stimmigkeit aber da das Modell für jeden Frame verformt werden muss um sich an die Gestallt und Größe des Darstellers im Eingabevideo-Frame anzupassen, wird die Genauigkeit des Models in Bezug auf das Original beeinträchtigt. Es wäre besser das Modell direkt von den Videodaten zu rekonstruieren um die Übereinstimmung zwischen Modell und Akteur zu optimieren. Es ist möglich ein Gitternetz aus jedem Videoframe zu erzeugen. Das Problem hierbei ist offensichtlich, dass deren Konnektivität sich von Frame zu Frame unterscheidet. Im Idealfall möchte man ein räumlich und zeitlich kohärente Animation individueller Gitternetze erzeugen. Um diese Lücke zu überwinden, stellen wir eine Methode vor die räumlich und zeitlich nahe dreidimensionale Korrespondenzen finden kann, und es somit erlaubt räumlich und zeitlich kohärente dynamische Geometrie von einer Sequenz unabhängiger Gitternetze zu erzeugen.

Jede dieser algorithmischen Lösungen kann unabhängig benutzt werden, um die jeweiligen Anforderungen eines speziellen Systems zu erfüllen. Darüberhinaus können sie auch gemeinsam und kombiniert in einem einzigen System benutzt werden, was in einer Animations-Pipeline endet die aus Multi-Video-Daten hochqualitative Animationen virtueller Akteure erzeugen und darstellen kann.

Acknowledgements

This thesis would not have been possible without the help and support of many individuals. First of all, I would like to thank my supervisors Prof. Dr. Hans-Peter Seidel and Prof. Dr. Christian Theobalt. I am extremely thankful to Prof. Seidel who gave me the opportunity to work in the truly remarkable research environment in the Computer Graphics group at the MPI and supported me throughout my thesis work. It was a privilege to work in one of the best Computer Graphics groups in the world.

I am also indebted to Prof. Theobalt, who was not only my supervisor but also acted as my mentor throughout my research work. He has been working with me from the start of my PhD and we have worked together on all projects described in this thesis. His support and guidance was invaluable for all my research work. I am also thankful to him for being a reviewer of this dissertation.

Furthermore, I would like to thank Dr. Macrus Magnor for being my senior supervisor during the beginning of my PhD. I am also thankful to Prof. Dr. Sebastian Thrun for reviewing some of the later projects and providing extremely helpful advice, and Prof. Dr. Gabriel Brostow who have agreed to be part of my graduation committee.

I would especially like to thank all my former and present colleagues in the Computer Graphics group at the MPI. I would like to thank them for their cooperation, and for the time they devoted in the discussions for different research projects. I would especially like to thank Edilson de Aguiar, who not only worked with me on some projects but also supported me throughout in all of the research work and also reviewed this thesis. I am also very thankful to Zhao Dong for all his support in various ways during the years. Additionally, I owe thanks to Christian Rössl, Hendrik P. A. Lensch, Gernot Ziegler, who were co-authors on some of my papers. I would like to thank Rhaleb Zayer, Ivo Ihrke, and Hitoshi Yamauchi for discussions and technical advice for some of the projects. I am very thankful to Art Tevs for his support in some of the projects, especially the Relightable Free-viewpoint video project and also for this thesis. I am also thankful to Peter Dobrev for his excellent contribution in the time-varying geometry reconstruction project. I would also like to thank Jan Petersen for his work in the Relightable Free-viewpoint Video project.

I have received support from many people in different ways throughout the course of my work. I am thankful to Shahzad Ahmed for proofreading part of this thesis. I would like to thank Khawar Deen for his support. I am also thankful to Akiko Yoshida and Jens Kerber for their help. I owe thanks to Carsten Stoll, Eda Happ, Kuangyu Shi and Natascha Sauber for their multi-view video recordings. I am also extremely grateful to the group secretaries, Sabine Budde, Conny Liegl and Sonja Lienard, admins of the AG4 group, the helpdesk team and other non-scientific members of the MPI who made the work an extremely pleasant experience.

I would also like to thank the Max-Planck Society, Max-Planck Center for Visual Computing and Communication, and the International Max-Planck Research School for Computer Science for their partial financial support during my PhD studies.

Finally, I would like to thank Andrea and Arthur Hermann from Dudweiler, Saarbrücken for their support. I would also like to thank my parents, my wife, other family members and relatives in Pakistan and around the world for their support and encouragement.

Contents

1 **Introduction** 1
- 1.1 Main Contributions and Organization of the Thesis 3
  - 1.1.1 Part I - Background and Basic Definitions 3
  - 1.1.2 Part II - Automatic Generation of Personalized Human Avatars 3
  - 1.1.3 Part III - High Quality Relightable Free-Viewpoint Video . 4
  - 1.1.4 Part IV - Highly Detailed Dynamic Geometry via Simultaneous Reflectance and Normal Capture 4
  - 1.1.5 Part V - Spatio-Temporally Coherent Dynamic Scene Reconstruction Without A Prior Shape Model 5

I Background and Basic Definitions 7

2 **Preliminary Techniques** 9
- 2.1 The Camera Model 9
  - 2.1.1 The Pinhole Camera Model 10
  - 2.1.2 Camera Calibration 11
  - 2.1.3 Two-View Geometry 11
- 2.2 Modeling and Animating Humans 13
  - 2.2.1 Modeling the Appearance 13
  - 2.2.2 Modeling the Kinematics 16
  - 2.2.3 Animating a Human 16

3 **Multi-view Video Studio** 19
- 3.1 Related Multi-view Acquisition Facilities 20
- 3.2 Recording Studio 21
  - 3.2.1 Camera System 21
  - 3.2.2 Lighting Equipment 22
- 3.3 Acquisition 23
  - 3.3.1 Camera Calibration 23

3.3.2 Color Calibration 24
3.3.3 Lighting Calibration 25
3.3.4 Background Subtraction 26
3.3.5 Recording 27

II Automatic Generation of Personalized Human Avatars 29

4 Automatic Generation of Personalized Human Avatars 31
4.1 Related Work 32
4.2 Overview 33
4.3 Reconstructing a Personalized Surface Texture 34
4.3.1 Texture Parameterization 34
4.3.2 Spatio-temporal Texture Reconstruction 35
4.4 Results 38
4.5 Conclusion 39

III High Quality Relightable Free-Viewpoint Video 43

5 Problem Statement 45
5.1 Related Work 47

6 Reflectance Sharing and Spatio-Temporal Registration for Improved 3D Video Relighting 49
6.1 Overview 49
6.2 Acquisition 50
6.3 Reconstructing Dynamic Human Shape and Motion 51
6.4 Warp Correction 52
6.5 Cloth Shift Detection and Compensation 53
6.6 Dynamic Reflectometry 54
6.7 Spatio-Temporal Reflectance Sharing 55
6.8 Results and Validation 58
6.9 Conclusion 62

IV Highly Detailed Dynamic Geometry via Simultaneous Reflectance and Normal Capture 63

7 Problem Statement 65
7.1 Related Work 66

8 Reconstructing High Quality Time-Varying Geometry 69

8.1 Overview 69
8.2 Data Acquisition and Template Motion Estimation 70
8.3 Enhanced BRDF Estimation 70
8.4 Adding Spatio-Temporally Coherent Geometric Surface Detail . . 72
8.4.1 Measurement Potential 74
8.4.2 Prior Potentials 74
8.4.3 Practical Implementation 75
8.5 Results and Validation 77
8.6 Conclusion 80

V Spatio-Temporally Coherent Dynamic Scene Reconstruction Without A Prior Shape Model 83

9 Problem Statement 85
9.1 Related Work 87

10 Spatio-Temporally Coherent Dynamic Surface Reconstruction Using Dense Correspondence 91
10.1 Overview 91
10.2 Spatio-Temporal Correspondence Finding 92
10.3 Coarse Correspondences 93
10.4 Finding Dense Correspondences 95
10.5 Remarks on Practical Implementation 97
10.5.1 Computation of Coordinate Functions 97
10.5.2 Intersection with Iso-contours 98
10.5.3 Prefiltering of SIFT Features and Adaptive Refinement . . 98
10.6 Alignment by Deformation 99
10.7 Results 99
10.8 Validation 100
10.9 Conclusion 103

11 Conclusions and Future Work 107

Bibliography 111

A List of Publications 125

B Curriculum Vitae – Lebenslauf 127

Chapter 1

Introduction

High quality reconstruction of 3D human animation from real-world data has been an active focus of research in both computer graphics and computer vision. Traditionally, an animator would need to manually create the model, then hand-craft the animation and high quality details. Furthermore, if the animation is to be rendered under different lighting, which is a typical scenario for the animated models used in computer games, the surface material properties have to be crafted manually which can be a painstakingly complicated process. This typically takes hundreds of work hours for a single model and consequently the costs of these productions are very high.

In both computer graphics and computer vision, the automatic reconstruction of the animation from multi-view video data has recently gained more attention. It involves the reconstruction of motion, shape, and appearance of humans. Optical motion capture using markers has been used to capture the human motion. Recently, the focus has been shifted from the marker-based to marker-less approaches. A pioneering work in the marker-less optical motion capture used an animated template human model and multi-view video data to capture the motion and photo-realistically render the virtual humans [Carranza03]. As an alternative to using an animated template model, the dynamic 3D geometry can be directly reconstructed from the video, thus resulting in high quality renditions [Starck07b]. Some of the methods do not use any 3D geometry, but create the novel views by interpolating the image data [Matusik04].

There are both benefits and drawbacks of the above mentioned methods. Nevertheless, for a true high quality reconstruction of human computer animations, there are still some very difficult problems that remain to be overcome. In this thesis we will show that many difficult problems that are encountered in the automatic reconstruction of human computer animation can be solved by means of algorithmic solutions using multi-view video data. People interact in the virtual environments by means of avatars which they choose based on their preferences. Many people prefer to use an avatar as close to their appearance as possible. Most of these virtual environments, be it the online chat rooms or massively multiplayer online games allow their users to create and customize their virtual appearance in many ways. However, it is very difficult to truly capture the correct appearance let alone the shape of the person using these rather simple tools. To create truly personalized human avatars, in Chapter 4, we propose a video-based approach that makes use of multi-view video data of the moving person and generates the life-like avatar of the person true to his/her shape and appearance. The method makes use of an animated template model to capture the motion and create a static texture that can be used to texture the geometry for the photo-realistic appearance.

The model description used for rendering the avatar is good enough as long as the lighting of the virtual environment is similar to the recording environment. In order to display him in a virtual world, which is different from the recording environment, his appearance must be adapted to the new illumination conditions. For this adaptation, the knowledge of surface reflectance properties of the human subject is necessary. Recently, Theobalt et al. [Theobalt05a], using an animated template human geometry, proposed a method to reconstruct these reflectance properties of moving actors using multi-view video data. We extend this method in Chapters 5 and 6, and propose two enhancements that can result in higher quality of relightable free-viewpoint video. Using the same multi-view video data we later extended this work even further and in Chapters 7 and 8, present a new passive approach to capture true time-varying scene geometry that can reconstruct even slightest of the dynamic details. Our method can reproduce dynamic surface details at millimeter-scale accuracy.

Instead of using a prior template, video data can be directly used to reconstruct the dynamic geometry. Most methods that utilize the video data to reconstruct geometric models for the purpose of animations provide very convincing shape and appearance for each frame. Unfortunately, they fall short of providing spatio-temporally coherent models, which is an extremely desirable property in the captured animations. Spatio-temporal coherence greatly facilitates or is even inevitable for many tasks such as editing, compression or spatio-temporal post processing. On the other hand, the methods that use an animated template model provide spatio-temporal coherence, but the tracking methods employed for animating and deforming the template model remain short of the accuracy provided by the reconstruction methods. In Chapters 9 and 10, we therefore propose a new 3D spatio-temporal dense correspondence finding method that enables us to reconstruct coherent scene geometry. Thus a template model is not needed and we obtain an accurate spatio-temporally coherent scene geometry directly from multi-view video data.

1.1 Main Contributions and Organization of the Thesis

This thesis is divided into 5 parts and contains 11 chapters. Apart from part I, which deals with the necessary theoretical and technical background and covers the preliminaries, each subsequent part presents algorithmic solutions based on multi-view video data that solve some of the problems that are encountered in automatic reconstruction of high quality 3D human animations. The algorithmic solutions described in part II, III, IV and V have been published before in a variety of peer-reviewed conference and journal articles. The main contributions of the thesis along with the references to the published work are briefly summarized in the following sections:

1.1.1 Part I - Background and Basic Definitions

This part covers the theoretical preliminaries required for the understanding of the rest of the thesis. In Chapter 2, we begin with the review of the camera model that is employed in computer graphics and computer vision. Thereafter, we discuss how to model the shape, appearance and kinematics of a human in a computer. We also review the techniques that are employed for character animation.

In Chapter 3 we describe our acquisition setup, which is a multi-view video studio that captures synchronized multi-view video streams. The recorded multi-view video data is used in all of the algorithmic solutions presented in this thesis. The details of obtaining multi-view video streams and their post-processing is described in this chapter.

1.1.2 Part II - Automatic Generation of Personalized Human Avatars

In multi-user virtual environments real-world people interact via digital avatars. In order to make the step from the real world onto the virtual stage convincing, the digital equivalent of the user has to be personalized. It should be possible to reflect the shape and proportions, the kinematic properties, as well as the textural appearance of its real-world equivalent. In Chapter 4, we present a novel fully-automatic method to build a customized digital human from easy-to-capture input data [Ahmed05]. The inputs to our method are multiple synchronized video streams that show only a handful of frames of a human performing arbitrary body motion. The avatar's geometry is generated by shape adapting a template human body model. Its surface texture is assembled from multi-view video frames showing arbitrary different body poses.

1.1.3 Part III - High Quality Relightable Free-Viewpoint Video

Free-View point video allows the user to view a dynamic scene from an arbitrary viewpoint. Theobalt et al. [Theobalt05a] presented a method for joint shape, motion and reflectance capture using multi-view video data that allows the reconstruction of relightable free-viewpoint video which can be viewed under arbitrary lighting. We improve their work and in Chapter 5 and Chapter 6, we introduce two methods that result in higher quality of relightable free-viewpoint video.

First, we present a novel spatio-temporal registration method that detects and compensates for the shifting of cloth across the body's surface of the actor [Ahmed07a]. Our second contribution was a spatio-temporal reflectance sharing method that reduces the bias in the estimated dynamic reflectance. This method assures that the estimated reflectance properties are not biased towards the recording environment [Ahmed07b].

1.1.4 Part IV - Highly Detailed Dynamic Geometry via Simultaneous Reflectance and Normal Capture

Models used for rendering the reconstructed animations lack high quality time-varying surface details that are normally visible in the moving apparel of a human actor, such as folds or wrinkles. Adding these dynamic details can dramatically increase the level of realism of the human animations. In Chapter 7, we start with the introduction of our passive method that can capture subtle time-varying surface details, e.g. folds and wrinkles, on a moving model. The starting point of the method is the enhancement of the solutions presented in Part III. Thereafter, we review the closely related work in the area of dynamic surface reconstruction, normal field integration, photometric stereo and reflectance estimation.

In Chapter 8, we present the crux of our statistical passive method that can add high quality dynamic details to the models [Ahmed08a]. First, an enhanced surface reflectance and normal estimation approach is described which employs robust statistics to handle sensor noise more faithfully. Next, a new spatio-temporal deformation framework is presented that enables us to transform the moving geometry and the time-varying normal field into true spatio-temporally varying scene geometry that reproduces geometric surface detail at high accuracy.

1.1.5 Part V - Spatio-Temporally Coherent Dynamic Scene Reconstruction Without A Prior Shape Model

A fast and versatile alternative template based methods for dynamic scene reconstruction is to reconstruct the geometric model from each frame of the video, e.g. by means of shape-from-silhouette methods. This reconstruction works fine for simpler animations but due to the lack of spatio-temporal coherence the usability of this data is very limited. In Chapter 9, we introduce and motivate our 3D dense correspondence finding method between a sequence of unrelated shapes that allows the reconstruction of a spatio-temporally coherent mesh sequence. The chapter ends with a review of the most important related work in the area of surface reconstruction, correspondence finding and mesh animation.

In Chapter 10, we present the main algorithmic solution for the spatio-temporally coherent reconstruction of a mesh sequence from unrelated shape-from-silhouette volumes [Ahmed08b]. This is achieved by employing a 3D dense correspondence finding method between two subsequent meshes, which is propagated over the whole sequence, resulting in a coherent animation.

Our work demonstrates that we can solve a variety of problems that are encountered in automatic reconstruction of 3D animation from video using multi-view video data. Our presented methods only require a small number (eight) of multi-view video streams, solve a wide range of problems, and can be used as the building blocks for high quality 3D animation reconstruction from video.

Part I

Background and Basic Definitions

Chapter 2

Preliminary Techniques

In this chapter, some general theoretical background is provided and some of the fundamental techniques which projects in this thesis employ are described.

All of the projects in this thesis rely on the synchronized multi-view video streams as input. These are captured by a multi-view camera system in our acquisition studio. In order to correctly use multi-view video streams, it is essential to simulate the real-world camera by means of a mathematical camera model. This mathematical camera model is presented in Sect. 2.1. We also discuss the process of camera calibration, and review the geometry from two-views.

In this thesis we focus on the reconstruction of human computer animations. Therefore we need a description of the human actor that can be used in the digital domain. In Sect. 2.2 we discuss how we model the shape, appearance and kinematics of the real-world human in a computer. We later describe a model for the kinematics and discuss how the model can be animated using the kinematic skeleton. We also discuss the animation of the model using deformation. Either of the two animation techniques has been used in all of the projects in this thesis.

2.1 The Camera Model

The camera captures a 2D image which is a projection of a 3D scene on a 2D plane. Function of the camera is very similar to the function of the human eye, where the 3D scene is the world around us and the 2D plane is the retina of the

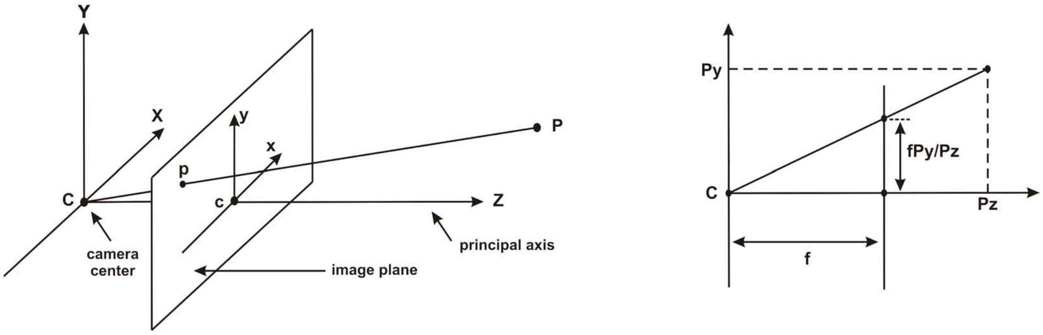


Figure 2.1: Pinhole camera geometry.

eye. Thus the role of the camera in computer graphics and computer vision is analogous to that of an eye in biological systems. Similar to the eye lens, the lens in the camera collects the incident illumination. The lens then converges the light rays towards a focal point, and the converged rays create an image of the observed scene over the image plane. In the following section, we will describe the pinhole camera model, which defines a mathematical relationship between the coordinates of a 3D point and its projection onto a 2D image plane. In later sections, we will describe the process of camera calibration and briefly review the concept of two-view geometry.

2.1.1 The Pinhole Camera Model

The pinhole camera is the simplest, and the ideal, model of camera function. It describes central projection of points in a space onto a plane [Hartley00]. Let a point in space with coordinates $\mathbf{P} = (P_x, P_y, P_z)^T$, the center of projection as the origin of the Euclidean coordinate system and the image plane $z = f$. The center of projection is also called the optical center or the camera center. The line from the camera center perpendicular to the image plane is called the principal axis, and it meets the image plane at the point called principal point.

The pinhole camera model maps $\mathbf{P}$ on the image plane where a line joining the point $\mathbf{P}$ to the center of projection meets the image plane, as shown in Fig. 2.1. It can be shown using the theory of similar triangles that the point $\mathbf{P}$ is mapped to the point $(fP_x/P_z, fP_y/P_z, f)^T$ on the image plane. Thus the 2D projection

$$(P_x, P_y, P_z)^T \mapsto (fP_x/P_z, fP_y/P_z)^T \tag{2.1}$$

describes the central projection mapping from world $\mathbb{R}^3$ to image coordinates $\mathbb{R}^2$.

2.1.2 Camera Calibration

To infer three-dimensional geometric information from an image, one must find the parameters that relate a point in the three-dimensional space to its two-dimensional position in the image. The parameters are classified as the *internal* and *external* parameters of the camera. There are four internal parameters: two for the position of the origin of the image coordinate frame, and two for the scale factors of the axes of this frame. As for the six external parameters: three are for the position of the center of projection, and three are for the orientation of the image plane coordinate frame.

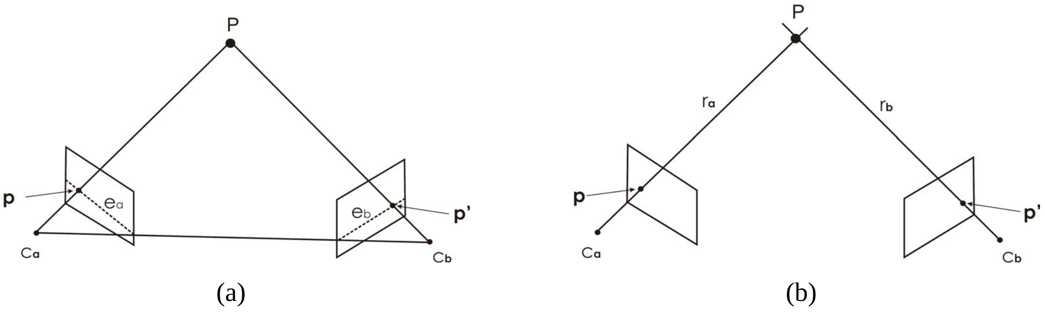
In addition, the physical properties of a real world camera lens differ from the properties of the ideal pinhole camera model. Due to these differences, the image formation process geometrically deviates from the pinhole camera. These deviations are typically caused by radial or tangential distortion artifacts. Radial distortion occurs, because unlike the ideal pinhole camera model, in the real lenses, the world point, image point and optical center are not collinear. Thus the world lines are not projected as lines. Radial distortion becomes more prominent as the focal length decreases. As a camera lens in itself is composed of many individual lenses, the misalignment of individual lenses with respect to the overall optical axis results in the tangential distortion [Weng90]. Most real world camera models take radial and tangential distortions into account, and include the parameters that compensate for the artefacts caused by them.

Majority of geometric camera calibration techniques [Tsai86, Jain95, Heikkila96] derive all of the above described parameters. Normally a calibration object with known physical dimensions is used to estimate the parameters. An optimization method is employed that modifies the model parameters until the predicted appearance of the calibration object optimally aligns with the captured images.

Color calibration refers to the correct reproduction of colors in the captured image under a given illumination condition. A simple color calibration technique is called white balancing, which involves the estimation of parameters that scale each color component with respect to a pure white or grey object. For our projects, we also perform color calibration that ensures color consistency across the cameras.

2.1.3 Two-View Geometry

Epipolar geometry refers to the geometry of stereo vision. It is the intrinsic projective geometry between two views, independent of scene structure, and only



(a) (b)

Figure 2.2: (a) Epipolar geometry: The point p in camera a corresponds to the point p' in camera b that lies on the epipolar line e_b. (b) Triangulation: the 3D position of a point P is calculated by the intersection of the two rays, r_a and r_b, through the respective cameras' centers of projection, c_a and c_b, and the respective projected image plane positions, p and p'.

depends upon the camera's internal parameters and relative pose [Hartley00]. It can be used to derive 3D structural information about the scene. Assuming a point $\mathbf{P}$ in 3-space is visible in both cameras, projected as $\mathbf{p}$ in the first camera, and as $\mathbf{p}'$ in the second camera. The epipolar geometry relates the two projected points by the so-called epipolar constraint, which describes that for the given $\mathbf{p}$, its correspondence $\mathbf{p}'$ should lie on the epipolar line e_b, Fig. 2.2a. Under the epipolar geometry the search for the correspondence for a given point is simpler as it only involves traversing a single line in the corresponding image plane instead of searching the complete two-dimensional image. The intrinsic epipolar geometry is encapsulated in the fundamental matrix F. It is a 3x3 matrix of rank 2, and for the two projected points satisfies the relation $\mathbf{p}'^{\mathbf{T}}F\mathbf{p} = 0$. The fundamental matrix can be inferred from 8 point correspondences between two uncalibrated cameras, and it is directly available for fully-calibrated camera pairs [Hartley00].

If both cameras are fully calibrated, with known correspondences $\mathbf{p}$ and $\mathbf{p}'$ in their image planes, then the 3D position of point $\mathbf{P}$ can be calculated via Triangulation, Fig. 2.2b. The position $\mathbf{P}$ is estimated by computing the intersection point of two rays, r_a and r_b. The ray r_a originates in the center of projection of camera $\mathbf{a}$, c_a, and passes the image plane in the position $\mathbf{p}$. The same construction is valid for ray r_b from camera $\mathbf{b}$, where the ray passes the image plane in the position $\mathbf{p}'$. However, due to measurement noise, the rays will not intersect exactly at a single point. In this case, a pseudo-intersection point that minimizes the sum of squared distance to each pointing ray is computed.

2.2 Modeling and Animating Humans

The human body is the entire physical structure of a human organism. It is a very complex system, in which an interplay of many physiological components result in its appearance, as well as physical and kinematics properties. General appearance of a human body is dependent on its skin, hair and in most of the cases when it comes to representing real-world humans, clothes. Appearance of the skin is dependent upon many underlying components, from the structure of the pigmentation to the deformation of the muscles. Given the fact that there are many different types of materials used in the clothes, the complexity of modeling the appearance increases even more. Physical properties of the human body model are influenced by its kinematics. The kinematic properties are determined from the body's skeleton. The skeleton is composed of bones which are connected with joints. In order to accurately capture a true human body in the computer, the model should represent the appearance, kinematics and physical properties as accurately as possible. In the following subsections we will review these representations.

Since the focus of this thesis is the reconstruction of human animations, accurate representation of the motion along with the appearance, shape and kinematics is equally important. We need to make sure that the model follows the motion of the human actor as accurately as possible, and for that we need techniques that can animate the model accordingly. In Sect. 2.2.3 we review two of the animation techniques that are employed in this thesis.

2.2.1 Modeling the Appearance

The realistic appearance of the virtual human model depends upon its geometry and its surface texture. The surface geometry of the virtual human is typically modelled by means of a triangle mesh. The triangle mesh is comprised of a set of triangles that are connected by their common edges. The triangles are also called the faces of the mesh, with each face made up of three vertices and three edges. The edge, which is formed by two vertices, is one side of the face. The vertex is the basic entity, and is typically shared between multiple triangles and edges.

There can be different ways to obtain the geometry for the human body model. It is possible to reconstruct the geometry from the input video data. Various methods are proposed to obtain geometry from multi-view images [Matusik00] [Kutulakos00] [Starck07b]. Fig. 2.3a shows a video frame from one of the camera, while the reconstructed visual hull can be seen in Fig. 2.3b. Another possibility is to use a generic template human body model as shown in

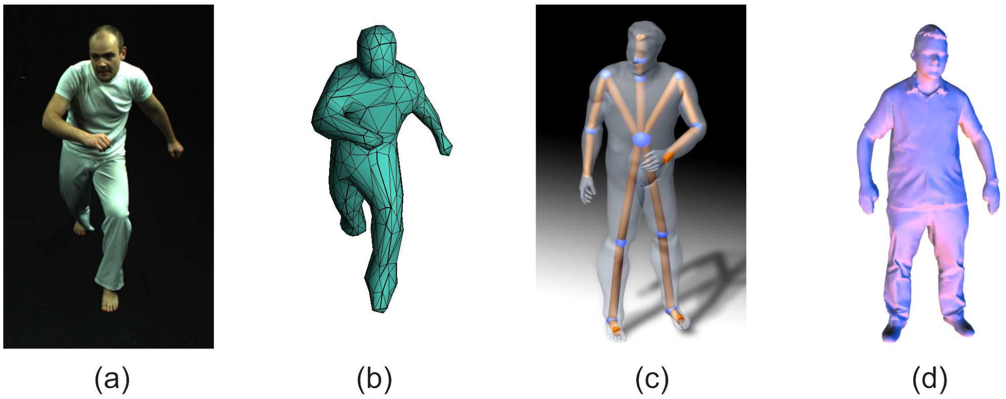


Figure 2.3: (a) Input video frame from one of the camera. (b) Reconstructed coarse geometry rendered from the same camera. (c) Template single skin human body model with superimposed kinematic skeleton. (d) A full-body laser scan of a human.

Fig. 2.3c, or make use of a full-body laser scanner and obtain the template geometry by measuring a real subject, Fig. 2.3d.

The second component for the realistic appearance of the virtual human model is its surface texture. A consistent surface texture for the model can be employed for photo-realistic renderings [Ahmed05]. Unfortunately a static texture cannot capture the true time-varying details, such as wrinkles and folds in the clothing, that evolve with the body pose.

If the model follows the poses of the human actor in the video, then it can be dynamically textured with multi-view video data, to reproduce the time-varying details [Carranza03]. This approach is feasible only when the virtual actor is reproduced under the illumination conditions that are very similar to the recording environment. Thus the illumination conditions should remain fixed during display of an animation.

If the model is to be rendered under arbitrary novel illumination conditions then however its surface reflectance properties must also be known. For the animated model, it requires the estimation of dynamic reflectance description (Chapter 5). The visual appearance of the surface is determined by the way incident light interacts with it and is sent back to the eye of the observer. In the most general case when light interacts with matter, there is one photon striking the surface at one particular point and one photon leaving the surface. In order to describe the general interaction case, a 12D function is necessary [Rusinkiewicz00].

This model can be significantly simplified if phosphorescence and fluorescence are ignored, wavelength changes are not considered, the wavelengths are discretized into bands, and the effects of subsurface scattering are not taken into account [Lensch04].

This results in a six-dimensional function, known as the spatially-varying *bidirectional reflectance distribution function* (BRDF) f_r. This representation is usually sufficient for realistic renditions of most of the materials. It is defined at all surface points $\vec{x}$ as the ratio of outgoing radiance l_o in hemispherical direction $\hat{v} = (\omega_o, \theta_o)$ to incoming irradiance $L_i \cos\theta_i \; d\omega_i$ arriving from direction $\hat{l} = (\omega_i, \theta_i)$:

$$f_r(\hat{v}, \vec{x}, \hat{l}) = \frac{dl_o(\vec{x}, \hat{v})}{L_i(\vec{x}, \hat{l}) \cos\theta_i \; d\omega_i} \tag{2.2}$$

In general BRDF can describe any surface reflectance characteristics and can be represented in many ways. Tabulated BRDFs store BRDF values in look-up tables and make use of the interpolation to represent novel incoming and outgoing directions. It provides good quality, but the storage cost is very high. Typically, in computer graphics, parametric models are used to evaluate reflectance for some specific illumination condition. The parameters differ for each material, and their variations result in a wide range of representable reflectance characteristics using the same mathematical expression. Most of the model are consist of a diffuse albedo component along with an analytic expression for evaluating the specular/glossy reflection. In our project on relightable free-viewpoint video (Chapter 6), we make use of two parametric BRDF models, the Phong model [Phong75] and the Lafortune model [Lafortune97b].

The Phong model is an empirical isotropic reflectance model that consists of diffuse object color and a specular lobe

$$f_r^{rgb}(\hat{l}, \hat{v}, \vec{x}, \rho) = k_d^{rgb} + \frac{k_s^{rgb}}{\hat{n} \cdot \hat{l}} (\vec{r}(\hat{l}) \cdot \hat{v})^{k_e} \tag{2.3}$$

Light source position $\vec{L}$ and viewing position $\vec{V}$ determine the light vector $\hat{l} = \vec{L} - \vec{x}$, viewing vector is $\hat{v} = \vec{V} - \vec{x}$, and given the surface normal $\hat{n}$, reflection direction is $\vec{r}(\hat{l}) = \hat{l} - 2(\hat{l} \cdot \hat{n})\hat{n}$. For evaluating both diffuse and specular color, we have to consider the red, green, and blue color channel separately. Seven model parameters $(k_d^{rgb}, k_s^{rgb}, k_e)$ then describe diffuse object color, specular color, and the Phong exponent which controls the size of the specular lobe.

A more advanced model based on the Phong model has been presented by Lafortune et al. [Lafortune97b]. It can additionally incorporate off-axis specular peaks, backscattering and even anisotropy:

$$\begin{aligned} f_r^{rgb}(\hat{l},\hat{v},\vec{x},\rho) &= k_d^{rgb} \\ &+ \sum_i [C_{x,i}^{rgb}(l_x v_x) + C_{y,i}^{rgb}(l_y v_y) + C_{z,i}^{rgb}(l_z v_z)]^{k_{e,i}} \end{aligned} \tag{2.4}$$

Besides diffuse color k_d^{rgb}, the model includes several specular lobes i whose individual direction, specular albedo and directedness are defined by $(C_{x,i}^{rgb}, C_{y,i}^{rgb}, C_{z,i}^{rgb}, k_{e,i})$. The vectors $\vec{l} = (l_x, l_y, l_z)$ and $\vec{v} = (v_x, v_y, v_z)$ are the normalized vectors corresponding to the hemispherical directions $\hat{l}$ and $\hat{v}$. For a more detailed discussion on reflectance models, we would like to refer the interested reader to [Lensch04].

2.2.2 Modeling the Kinematics

The computational model for the human skeleton is a kinematic skeleton. A Kinematic skeleton is a mathematical model which represents the human skeleton as a hierarchal arrangement of joints and interconnecting bones. The result in an articulated figure consisting of a set of rigid segments connected with joints. The set of rigid body segments form a kinematic chain, which is essentially an hierarchal assembly of rigid bodies. The relative orientation between one segment and the following rigid body segments in a kinematic sub-chain is controlled via a rigid body transformation. This rigid body transformation describes a joint rotational and translational transformation between two the local coordinate frames of two subsequent rigid bodies. As the kinematic skeleton is a hierarchal structure, the transformation on the top level influences all the connected rigid bodies. Consequently, the transformation on the lowest level rigid body only affects that specific body.

Fig. 2.3c shows a kinematic skeleton superimposed on a human body model. The skeleton models most important joints and segments that are necessary for the correct representation of the human. It consists of 16 segments and 17 joints, unlike the real human body skeleton which consists of 206 bones and more than 200 joints. The bone lengths in the skeleton implicitly encode the translational component of the transformation. Thus the joints of the model only represent the rotational component. Since the bone lengths are constant, we only need rotation information for each joint to define the pose of the skeleton. Varying angles of the joints yields an infinite number of configurations. A global translation for the root of the skeleton can be employed as the only required translational component.

2.2.3 Animating a Human

The geometry that we obtain from any of the method described in Sect. 2.2.1, should be somehow animated to reconstruct the motion of the human actor in the video. In this thesis we make use of two techniques, using the kinematic skeleton or deformation.

We make use of the animation based on the kinematic skeleton in the relightable free-viewpoint video project, Chapters 5 and 6. In this project, first a kinematic skeleton is implanted into the geometry of the single skin template human model, Fig. 2.3c. Thereafter, the skeleton is attached to the surface by assigning the weights to each vertex of the geometry in accordance with its relative position to each bone. A bone would exert more influence on its nearby vertices. This influence is represented by the weights, which control the deformation of the mesh as the joints are rotated. Each vertex can be influenced by multiple bones and the weights from each bone are blended. The technique of assigning the weights in this way is commonly called linear blend skinning [Baran07]. Finally, the motion description in terms of joint parameters is automatically estimated using a silhouette based analysis-through-synthesis method (Sect. 6.3).

Another approach for animating the model would be to use mesh-deformation methods [Botsch07]. These methods are employed to great effect in performance capture of humans [de Aguiar07a] [de Aguiar08]. In our work of parametrization-free animation reconstruction using dense 3D correspondences, we make use of a mesh deformation approach to animate the reconstructed visual hull, Chapters 9 and 10. Our solution is independent of any specific deformation approach, therefore we refer the reader to a recent survey in the area of surface deformation [Botsch07].

Chapter 3

Multi-view Video Studio

This chapter describes our recording studio. First, the studio room, the camera system and the lighting setup are described. Thereafter, the acquisition pipeline is presented, with all necessary steps to generate the input data for the projects described in this thesis.

All of the projects presented in this thesis require high quality multi-view video data as input. This data is recorded in our multi-view video studio, where we simultaneously capture video streams from eight synchronized video cameras.

In this chapter we will present our multi-view video studio in detail. The studio is an extension of [Theobalt03], which was a simpler multi-view acquisition setup. We present our new acquisition studio, which provides high quality data that are recorded not only using the calibrated cameras but also under completely calibrated illumination conditions. These data are the main requirement of our work on relightable free-viewpoint video (Chapters 5 and 6), and subsequently high quality reconstruction of time-varying geometry (Chapters 7 and 8). The acquisition setup of the studio is enhanced with the addition of hiqh frame rate and high resolution cameras along with the better lighting setup, which facilitate us greatly in the reconstruction of high quality surface models. High frame rate and high resolution data were also invaluable for our work on the parametrization-free animation reconstruction using dense 3D correspondences (Chapters 9 and 10).

We will start this chapter with a review of related multi-view acquisition systems. Thereafter, we will describe the recording studio, and discuss our camera and lighting system that is installed in the studio. Finally, we will present the acquisition process, which is comprised of camera, color and lighting calibration, background subtraction and finally the actual recording of the human actor.

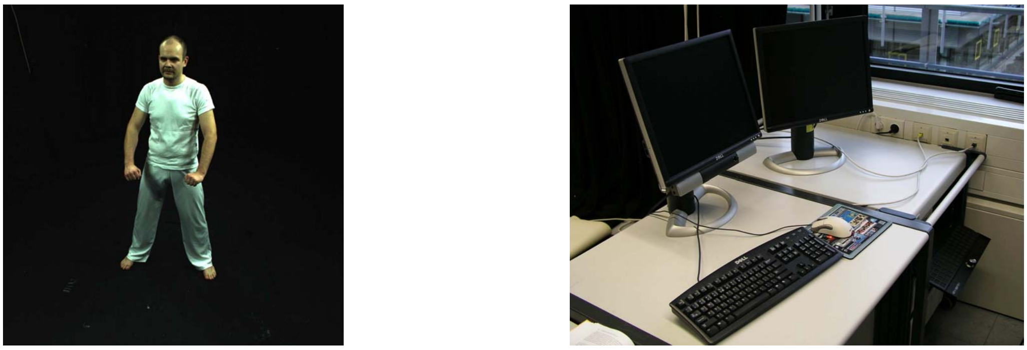
3.1 Related Multi-view Acquisition Facilities

Multi-view data is used in variety of research areas. Various setups for their acquisition exist, based on the specific needs of the research. The project presented in this thesis are versatile in the sense that they encompass many research areas that require these data. Therefore, our multi-view video studio is designed in such a way that the specific requirements for data are not compromised.

Image based reflectance estimation requires very high quality image data. For estimating the surface reflectance models of real-world object, a series of images obtained from different viewing directions and taken under different incident illumination conditions are required. For static scenes, acquisition setup using high quality photo cameras and a set of light sources have been proposed [Ward92, Goesele00]. [Debevec00] presented a light stage to capture the reflectance field of animatable face model. [Einarsson06] extended it further by using a large light stage, a tread-mill where the person walks, so that they can acquire simple motion and reflectance field of humans. Unfortunately, their setup can only process simple periodic motions, such as walking. In contrast our multi-view video studio allows the extension of the photo camera based reflectance estimation method into video based dynamic reflectometry, without any restriction on the type of motion.

Multi-view video streams are readily used in the area of video-based motion capture. In our work we focus on marker-less motion capture, because it allows recording of the human actor without any optical markers attached on the body. Video acquisition in a 3D room that allows recording with up to 48 cameras is presented by [Kanade98]. Systems for motion acquisition using reconstructed volumes are presented in [Cheung00, Borovikov00, Luck02, Brostow04]. Commercial solutions for marker-less motion capture are now also available [Motion]. For an extensive review of video-based motion acquisition systems, we would like to refer the interested reader to [Poppe07].

Another research area that makes use of multi-view video streams is 3D video. In addition to capturing the motion, multi-view video streams can be used to reconstruct the dynamic shape and appearance models of the human actor. This enables the user to change the viewpoint of the scene during the rendering. [Narayanan98] made use of 50 cameras and reconstructed 3D models of dynamic scenes using



(a) (b)

Figure 3.1: Our recording studio includes (a) the recording area and (b) the control room.

dense stereo. [Würmlin03] presented a method to record and edit 3D videos, and further extended it in [Waschbüsch05]. [Matusik04] presented a complete system for real-time acquisition, transmission and rendering of 3D Video. Recently [Starck07b] presented a 3D video system that captures appearance, shape and motion from multi-view video data.

3.2 Recording Studio

Our multi-view video studio is designed to be flexible and versatile such that it fulfils the requirements of all the research projects. It is built from off-the-shelf hardware. It is designed to acquire high quality video footage of humans that can be used in surface reflectance measurement, dynamic surface reconstruction, motion capture, dynamic shape deformation, and appearance modeling.

The studio is located in a room of approximately 9x4.8 meters in size. The ceiling has a height of approximately 4m. An area of 2.5x4.8 meters is separated, which serves as a control room of the studio. The remaining area of the studio, which can be optionally enclosed with black curtains and carpets to minimize the effects of indirect illumination, is the recording area. The recording area and the control room of the studio are shown in Fig. 3.1.

3.2.1 Camera System

The camera system in our studio is comprised of eight Imperx^{TM} MDC1004 single chip CCD cameras, Fig. 3.2a. The imaging sensor of the cameras has a resolution of 1004x1004 pixels with 12 bits per pixel color depth. The sensor uses a Bayer mosaic to record the red, green and blue color information. The CCD sensor is connected to two controller chips. It provides a sustained frame rate of 48 fps at full resolution when both controller chips are activated. In this mode, the photometric responses of the sensors is out of synch and an intra-frame color adjustment step is necessary. With only one chip activated, the CCD sensor provides a sustained frame rate of 25 fps at full resolution and there is no need for the color adjustment.

The cameras are linked to a control PC equipped with 8 high-speed frame grabber boards. Each frame grabber is connected to a camera through a Camera Link^{TM} interface. For maximal data rate, each capture card is equipped with an on board SCSI interface enabling direct streaming of image data to a RAID system. Eight RAID systems are employed in parallel to enable real-time storage of the video streams. The cameras are synchronized via a trigger pulse that is broadcasted to each capture card.

The cameras can be installed at any location in the studio. In general cameras are placed in an circular arrangement around the center of the scene. For the relightable free-viewpoint video project, we placed one camera on the top. A typical arrangement allows us to capture a volume of approximately 3.5x3.5x3 meters with all cameras.

3.2.2 Lighting Equipment

Along with the camera system, the lighting equipment in the studio is crucial for the image quality of multi-view video streams. In order to fulfill the need of appropriate illumination conditions for different applications, it is important to provide a flexible lighting system. For our research, it is important to have both an ambient scene lighting, as well as more specific spot light kind of set up.

For general lighting, we employ 8 NesyFlex 440 DI ^{TM} compact softlights [Nesys] that are optimized for universal use in TV and video studios, Fig. 3.2b. Each light component contains 8 fluorescent day light tubes that radiate even light at a wide angle. They illuminate objects in the center of the scene from the top of the recording area and spread the light homogeneously downwards. The system can be controlled as a single unit using the DMX ^{TM} controls. Additionally, each light

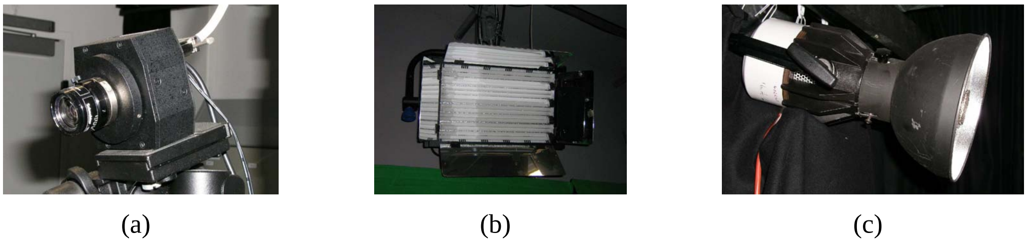

(a) (b) (c)

Figure 3.2: (a) Imperx™ MDC1004 camera, (b) NesyFlex 440 DI ™ softlight and (c) K5600™ Jokerbug spotlight.

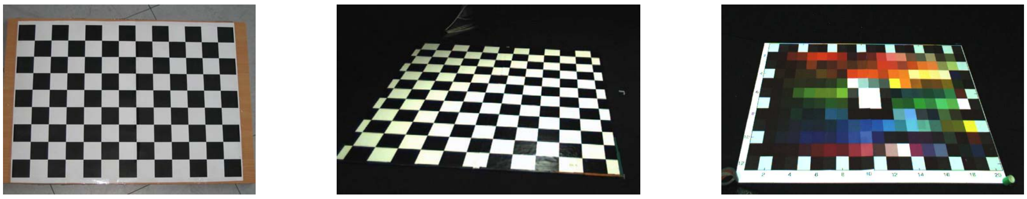
can be rotated to fulfil specific requirements. By this end, the lighting system prevents direct illumination of the camera lenses, avoiding glares, and produces a very uniform lighting in the scene, avoiding sharp shadows and unwanted highlights on the recorded subjects.

For our project on relightable free-viewpoint video, we employ two K5600™ Jokerbug 800 spot lights to illuminate our scenes, Fig. 3.2c. They are placed in opposite corners of our studio, and they are oriented towards the center of the recording area. The spot lights emit light with a daylight spectrum, and different lenses can be used to modify the shape of the beam according to our needs.

We have fully controllable lighting system in our studio. No exterior light can enter the recording area, and the influence of indirect illumination from the walls can be minimized by covering up all the walls by opaque black molleton. Optionally, the indirect illumination reflected off the floor and the visual appearance of cast shadows can be minimized by rolling out a black carpet.

3.3 Acquisition

With our multi-view video studio, we can efficiently acquire camera and lighting attributes along with multi-view video data that is used in all our research projects. Before commencing the actual recording of the human actor, we acquire all the necessary information that is needed for camera, color and lighting calibration. We also record the information required for the background subtraction. Finally, the actual recording of the human actor takes place.



(a) (b) (c)

Figure 3.3: (a) smaller checkerboard pattern used for determining intrinsic camera parameters, (b) large checkerboard pattern used for extrinsic camera parameters estimation and (c) color calibration pattern.

3.3.1 Camera Calibration

For our projects, we need to determine, both the internal and external parameters for each of the 8 cameras. For the camera calibration, we record two calibration objects of known dimension to be used by our calibration methods. For intrinsic calibration a small calibration pattern positioned in front of the cameras is recorded, Fig. 3.3a. A larger checkerboard visible from all the cameras is recorded to facilitate the extrinsic calibration, Fig. 3.3b.

For determining intrinsic camera parameters we employ Heikkila's method [Heikkila96]. The estimated parameters are used to undistort the calibration images and multi-view video streams. Extrinsic camera parameters are estimated by means of the Tsai algorithm [Tsai86]. Our calibration software automatically detects the corners of the checkerboard, with known world space positions. An optimization procedure estimates the extrinsic camera parameters by minimizing the reprojection error between the measured and predicted position of the checkerboard pattern.

3.3.2 Color Calibration

Accurate color reproduction among different cameras is very important not only for the correct renditions but also for the surface reflectance measurement. In the first step, to ensure the correct color reproduction, all the cameras are white balanced before the recording session. However, due to sensor noise, and slight physical differences in built-in camera components, there can be still discrepancies in the color response of each camera. To resolve these color discrepancies, we record a color calibration pattern which consists of an array of 237 uniformly colored squares with purely lambertian reflectance, Fig. 3.3c.

Using the recorded color calibration pattern, we perform relative photometric calibration. We define one camera as the reference and for each remaining camera, a color transformation is computed such that the color values of the pattern in the reference camera are reproduced. A trilinear transformation of the RGB color values is used for the color transformation. The determined color transformation for each camera is applied on each frame of the respective video stream to ensure the faithful color consistency.

3.3.3 Lighting Calibration

In order to measure the reflectance properties of an object, the light source should be carefully calibrated, i.e. their position, luminance intensity and color should be known. In order to find the photometric properties, we use the approach proposed in [Debevec97], to generate a High Dynamic Range (HDR) image from a set of images of a mirrored ball taken at different exposure times in the studio, Fig. 3.4a. Using the camera Olympus Camedia C5050$^{\text{TM}}$, first its response curve is calculated and then 13 different images are taken, with exposure times varying from $1/1000$s to 4s. These images and the response curve are used to generate the HDR mirrored ball image, which is converted to a cubic environment map representation, Fig. 3.4b. Using the HDR cube map, the algorithm described in [Agarwal03] finds the light position in the image domain. The luminance intensity and color of all light sources is found by integrating the respective values of all the pixels belonging to the light source. Both spot lights are approximated as point light sources.

In order to find the 3D position of the light sources, the following method is used: In addition to capturing HDR mirrored ball images, images of the mirrored ball are obtained from two calibrated cameras in the studio, Fig. 3.4c. In both images the center of the mirrored ball in the image plane C_{ip} is identified. Using extrinsic and intrinsic parameters of the cameras, the 3D position of the center of the mirrored ball C_{wp} is found by shooting the rays from both cameras towards their respective C_{ip} and calculating their intersection.

In order to find the correct orientation of the cube map with respect to the camera system in the studio, two calibrated cameras are marked with distinct colors. An image of the cube map is generated from the HDR cube map, such that the reflections of both marked cameras are visible, and it is projected onto a sphere. Four vectors are constructed: $\vec{v_1}$ and $\vec{v_2}$ from the marked studio cameras to C_{wp}, and $\vec{m_1}$ and $\vec{m_2}$ from the corresponding marked cameras in the cube map to C_{wp}. The correct orientation is found by rotating the sphere such that $\vec{m_1}$ and $\vec{m_2}$ overlap $\vec{v_1}$ and $\vec{v_2}$ respectively. Two rotations are enough to assure the correct orientation.

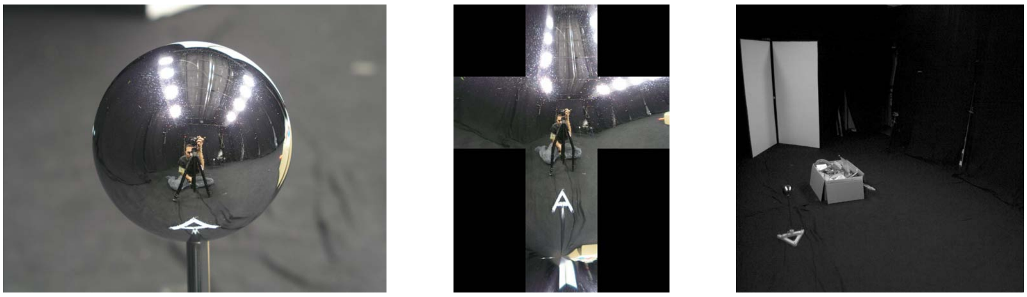

(a) (b) (c)

Figure 3.4: (a) captured light probe, (b) transformed cube map and (c) light probe captured from one of the multi-view video camera.

This correct orientation is used to find the accurate direction of the light sources using the cube map.

We apply the above calibration procedure for two mirrored balls placed in different positions in the room. This was done to find the accurate position of the light sources. For the luminance and intensity, the information from only one light probe is sufficient. From each mirrored ball we get a direction vector for each of the spot light. Therefore, for each spot light we have two direction vectors from the center of spheres towards each spot lights. The position of each light source is then computed trivially by intersecting the rays along these direction vectors.

3.3.4 Background Subtraction

All the projects in this thesis require the data in which the human actor is separated from the background. The lighting in our studio is completely controlled, and the effects of external light on the scene and cast shadows are minimized. This simplifies the process of background subtraction. For background subtraction we simply record the studio without the human actor from all cameras. This background image is used by the background subtraction algorithm to separate foreground from the background. This algorithm computes mean color and standard deviation for each background pixel. Foreground pixels are identified by a large deviation of their color from the background statistics. For details of this procedure, we would like to refer the reader to [Theobalt05a].

3.3.5 Recording

After all the necessary data required for camera calibration, color calibration, lighting calibration and background subtraction are recorded, the actual recording session commences. The human actor can perform any motion within the recording area. The performance is recorded by our eight synchronized video cameras. Every research project has a different set of requirements for the input data. The acquisition setup and the specific recording requirements are also briefly discussed in each of the projects separately later in the thesis.

Part II

Automatic Generation of Personalized Human Avatars

Chapter 4

Automatic Generation of Personalized Human Avatars

This part presents a method for generating personalized human avatars from multi-view video data. First, the related work in this area is reviewed, then a spatio-temporal method to adapt the shape and skeletal dimensions of the human model is presented. Finally, a method for reconstructing a consistent surface texture for the model using multi-view video frames from different camera views and different body poses is described.

In recent years, virtual environments in which real-world people can interact through controllable digital characters, so-called avatars, have become accessible even to the user at home. In order to make their appearance on the virtual stage convincing, many users want to give their digital equivalent a personal touch. Unfortunately, in most online games or 3D chat rooms, the degree to which a user can personalize his avatar is very limited. At best, he can manually modify a body shape taken from a database of template geometries, and texture the face of the virtual puppet with a digital photograph. It is obvious that, in order to make the personal touch fully convincing, the animatable human model should reflect the complete shape and textural appearance of the real-world human that it represents.

In order to serve this purpose, we have developed a novel fully-automatic method to build a customized digital human from easy-to-capture input data [Ahmed05]. The inputs to our method are multiple synchronized video streams that show only a handful of frames of a human performing arbitrary body motion. Our approach is based on a template human body model consisting of a triangle mesh surface representation and an underlying kinematic skeleton. This body representation is automatically deformed until it matches both the shape and the skeletal structure of its real-world counterpart captured in the video footage.

The main contribution of this work is the reconstruction of a consistent surface texture from multi-view video streams. This consistent surface texture is employed for the realistic rendition of the digital human. By simultaneously employing images from multiple camera views and multiple time steps of video, it is made sure that even temporarily invisible parts of the body surface are faithfully captured in the texture. With our novel method we quickly generate photo-realistic digital actors from real-world people using acquisition technology that may, in the near future, be available even to the user at home.

4.1 Related Work

Acquisition of visually realistic models of humans from images has been a long standing problem in computer graphics and virtual reality. In order to generate a realistic human avatar, the kinematics, shape and appearance have to be captured simultaneously.

Full-body range scanning systems exist that can quickly acquire the full surface geometry of a human body. However, they are highly expensive and don't straight forwardly enable to estimate a skeleton of the human [Paquette96]. Alternatively, image- or video-based methods can be used to reconstruct body models. In one line of research, it is the primary goal to derive the kinematic structure and a simple surface geometry from image data [Kakadiaris95, Fua98, de Aguiar04]. A surface texture, however, is not reconstructed.

In 3D video, novel views of a real person are rendered from multiple input video streams [Moezzi97, Kanade97, Matusik00]. Unfortunately, these approaches do not reconstruct models that could be animated with arbitrary novel motion data.

We propose a novel model-based approach that creates a fully-animatable avatar comprising a customized geometry, a realistic surface texture, and an appropriately rescaled skeleton. Our work is similar to the methods proposed by Hilton et al. [Hilton99] and Lee et al. [Lee00], where a human template model is deformed until it aligns with multiple silhouette images. Surface textures are created by mapping photographs back onto the body representation.

In contrast, we present a novel approach that employs multiple time steps of multi-view video footage to capture shape and texture at higher accuracy. To achieve this

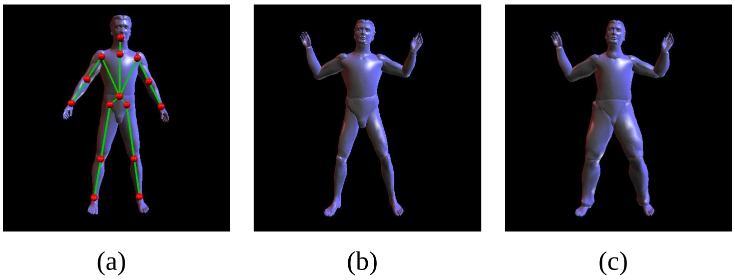


Figure 4.1: (a) Adaptable generic human body model; (b) initial model after skeleton rescaling and pose estimation; (c) model after spatio-temporal free-form deformation scheme.

goal, we build upon and extend the silhouette-based marker-free motion capture method detailed in [Carranza03]. As opposed to many previous approaches, our method is fully-automatic and even enables the extraction of multiple face textures depicting different facial expressions. This way, the appearance of the avatar can be changed on-the-fly such that it reflects its current mood.

4.2 Overview

Multi-view video (MVV) sequences used as input to our system are recorded in our multi-view studio (Chapter 3). In each MVV sequence that serves as input to our avatar creation algorithm, the person first strikes an initialization pose (Fig. 4.1b) for a short moment, and thereafter is free to move arbitrarily. In a post-processing step, the silhouette of the person in each frame is extracted via color-based background subtraction (Sect. 3.3.4). Typically, even short motion sequences of only 3-7 seconds are sufficient for our method.

We employ a template human body model whose shape and proportions can be customized in order to optimally reproduce the appearance of a person in the real world, Fig. 4.1a. The kinematics of the model are represented by means of a skeleton comprising 16 segments and 17 joints that provide 35 pose parameters in total. The surface geometry of each segment is represented via a closed triangle mesh.

We employ the method presented in [Carranza03] to capture the initial shape of the model and derive the correct body pose at each time step of the video, Fig. 4.1b. Additionally, we use the approach by de Aguiar et. al. [de Aguiar05] for spatio-temporal free-form deformation, in order to increase the quality of the captured shape of the model, Fig. 4.1c.

4.3 Reconstructing a Personalized Surface Texture

Once we have a shape adapted 3D model in the correct pose, the final component contributing to a realistic look of our avatar is a photo-realistic surface texture. Previous approaches to avatar creation reconstructed a static surface texture from multiple photographs showing the person in a single pose. Although the so-created virtual actors look authentic if they strike the same pose as the person in the images, very disturbing appearance artifacts may occur if their bodies are animated. One reason for such artifacts is texture undersampling due to insufficient visibility of certain body areas in a single pose. Also problematic are those parts of the body geometry that are temporally occluded by other body segments (e.g. in the shoulder or leg area) but which become suddenly visible as soon as the pose of the skeleton changes (Fig. 4.8c).

A third problem is that photographic textures “freeze” the local appearance of dynamic surface details as well as local illumination effects. We address all theses issues in conjunction by means of a spatio-temporal texture reconstruction scheme that samples from multiple time steps of the MVV sequence. It estimates color information also for temporally invisible areas of the body. In the following, we first describe our texture parameterization. Thereafter, we detail the spatio-temporal texture reconstruction algorithm.

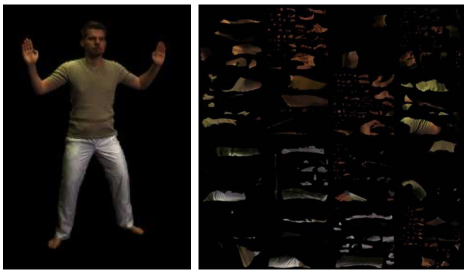


Figure 4.2: Input video frame and corresponding MVV texture for a male actor.

4.3.1 Texture Parameterization

Each body segment is parameterized separately over a planar rectangular domain using patches of minimal distortion [Ziegler]. The sixteen planar patch layouts are finally assembled into one texture atlas for the complete model. This way, we obtain a pose-independent bijective 3D-to-2D mapping between a surface element and a texel in the texture domain. Throughout our experiments, we use 1024x1024-texel texture maps. The graphics hardware is used to transform each video camera image into the texture domain. All data related to surface elements (view vectors, visibility etc.) can now be conveniently stored as textures. For each video time step, eight so-called multi-view video textures (MVV textures) are created (Fig. 4.2) by transforming the individual video frames into texture space.

4.3.2 Spatio-temporal Texture Reconstruction

Since we know the exact body pose of the model in each time step of multi-view video we can incorporate image data of multiple body poses into one consistent surface texture. This, in turn, enables us to fill-in color information for surface areas that are invisible in one body pose from images of the model in another body pose. There are two main reasons for why a surface point may not be visible from any input camera view.

- Mutual occlusion of directly adjacent body segments: Some areas of a segment can be occluded by the directly adjacent segment, e.g. parts of the upper arm segment that are inside the torso.
- Camera placement: For any possible arrangement of imaging sensors some parts of the model may be invisible, even though they are not occluded by any neighboring body segment.

In order to differentiate which of the two cases applies to a specific invisible surface point, we have developed the following two-step spatio-temporal texture reconstruction procedure which implicitly handles both cases:

Before texture reconstruction commences, U time steps of the input MVV sequence from which the color information for the final texture is assembled are automatically selected. In step 1, the *single-time-step texture assembly*, we create U individual consistent surface textures, $stex_i$, with $i \in \{1, \ldots, U\}$. Each $stex_i$ is only reconstructed from multi-view video images of time step i. The color of a texel is computed by weightedly blending the colors at its projected locations in each of the camera views. The blending weights are computed in such a way that a camera which sees a surface point more head-on is assigned a higher blending weight. To this end, we employ the view-independent weighting scheme described in [Carranza03].

In order to compute the visibility of each surface point in all of the camera views, we have developed a scheme which looks at each of the 16 body segments sep-

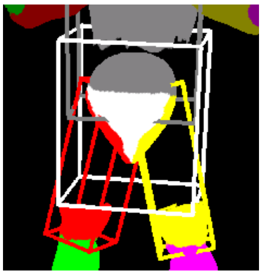


Figure 4.3: Trimming procedure for the pelvis segment: Vertices in the torso and the upper legs that lie inside the pelvis' bounding box are discarded. The white parts of the torso are also visible for camera 2 in the untrimmed model. All other colors indicate the adjacent body segments that, prior to trimming, occlude these vertices.

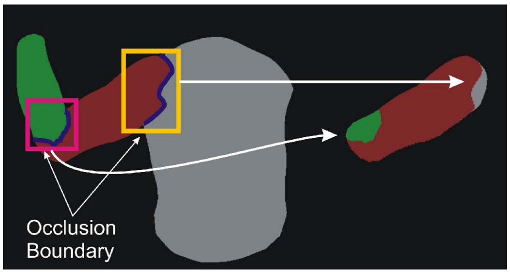
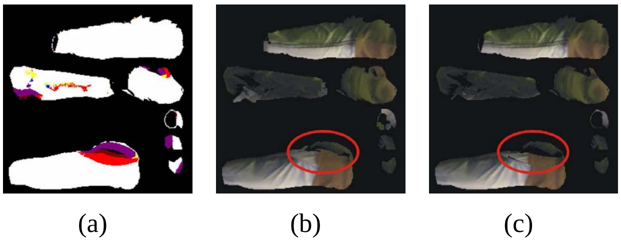


Figure 4.4: Texture information for surface segments that are occluded by adjacent triangle meshes is only taken from those parts of the occluding geometry that are close to the occlusion boundaries.

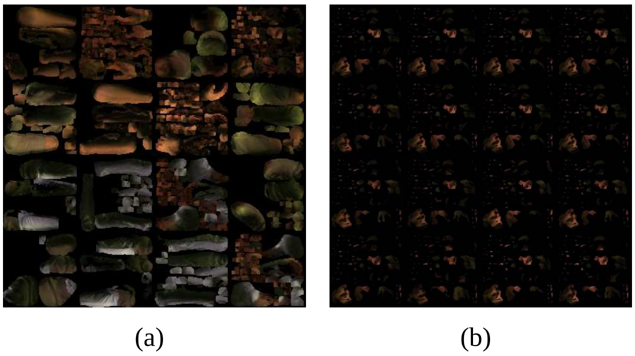
arately. Using the pelvis as an example, the scheme works as follows: First, a slightly enlarged bounding box of the pelvis is generated. All triangle vertices on directly adjacent segments (i.e. torso and upper legs) that are inside that bounding box are trimmed. For each input camera view, the visibility of each vertex in the pelvis is determined from the trimmed version of the model. In Fig. 4.3 the trimming procedure for the pelvis segment is visualized. The white regions on the pelvis illustrate those parts of the geometry that have been visible in input camera 2 even in the untrimmed model. All pelvis areas with another color were occluded by one of the directly adjacent segments. By this means, we implicitly create texture information for parts of the surface geometry that are invisible due to mutual occlusion between neighboring triangle meshes. Our visibility computation scheme makes sure that occluded texture parts are filled-in from those parts of the occluding geometry that are spatially close to the occlusion boundary on the 3D surface (Fig. 4.4). Texture parts of the occluder that are far from from the occlusion boundary do not contribute to the occluded texture area.



(a) (b) (c)

Figure 4.5: (a) Color-coded rendition showing from what time steps of multi-view video each texel in the final texture of the left upper leg was reconstructed. Consistent segment texture with (b) and without (c) mean filtering of areas that do not stem from the reference texture. Smoothed areas are encircled in red.

In step 2, the *texture combination* step, we merge all single-time-step textures, $stex_i$, into one final texture (Fig. 4.6a). The texture generated from the model in the initialization pose, $stex_1$, is considered as the reference texture. For every texel in $stex_1$ whose color is not known we make a look-up, in ascending order, into all remaining single-time-step textures $stex_i$,$i \in \{2, \ldots, U\}$. The color of the texel is copied from the first texture in which it is visible. Fig. 4.5a illustrates from what time steps of the multi-view video sequence the colors in the final texture of the left upper leg were taken. Color discontinuities in the final texture that may arise in those areas that have not been reconstructed from the reference time step are smoothed by locally applying a mean filter (Fig. 4.5b,c).



(a) (b)

Figure 4.6: An example of a complete body texture (a) and a packed face texture (b) for a male avatar.

In many virtual environments it is a nice feature to be able to express the mood of the avatar with a texture that shows a particular facial expression. One can store a complete body texture for each facial expression. Making use of our texture parameterization which maps each body segment to a 2d patch in the texture domain, we can optimize this storage. As an optional step, our approach thus allows the user to manually select a set of time steps from the input sequence that show interesting facial expressions. For each of these time steps, we create a separate texture of the face segment only. All face textures are efficiently stored in a packed format (Fig. 4.6b). The packed face texture can be loaded together with the full body texture and, depending on the actor's mood, the facial expression can be changed on-the-fly (Fig. 4.7).

4.4 Results

We have several multi-view video test sequences of a male and a female actor wearing different types of apparel at our disposition. Each of the input sequences is between 3 and 7 seconds long. We employed different number of frames for texture reconstruction. A comparison revealed that 5 frames are sufficient to create a complete texture without artifacts. On a PC featuring a Pentium^{TM} 4 CPU and an Nvidia GeForce 6800 GPU one iteration of the skeleton rescaling method on average takes around 1 minute. We employ 5 times steps for spatio-temporal free-form deformation, which takes around 15 minutes to find optimal scaling parameters. The spatio-temporal texture reconstruction method takes, on average, around 40 seconds if 5 time steps of the MVV sequence are considered. On the whole our method requires 17 minutes for processing a sequence.

Figs. 4.8a,b show a comparison between the actor as he appears in one of the video frames, and the rendered avatar in a novel body pose. Our approach faithfully captures the shape and the textural appearance of the actor in different types of apparel. Even in body poses that are significantly different from any of the captured ones, appearance artifacts due to texture undersampling are hardly visible.

Fig. 4.8c demonstrates that, if the surface texture is only reconstructed from a single time step, severe rendering artifacts may appear at segment boundaries. In contrast, our spatio-temporal texture reconstruction scheme generates a very consistent surface texture (Fig. 4.8d).

In addition, our approach enables to capture various face textures and store them in a compact format. Two renditions of the avatar with different facial expressions are illustrated in Fig. 4.7. The reconstructed realistic virtual humans can be used to realistically populate artificial virtual environments (Fig. 4.8e).

Our results show that the employment of image data of multiple body poses during texture reconstruction enables us to reconstruct human avatars that exhibit a very high visual quality. Although we employ specialized multi-view video hardware

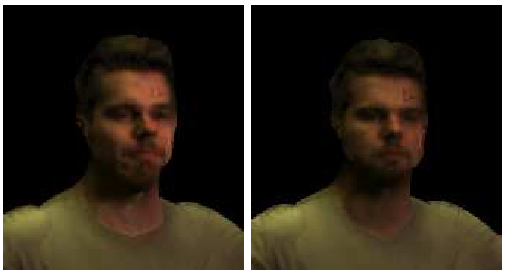


Figure 4.7: Different moods of the avatar can be expressed with different facial expressions.

for acquisition, images of a human in different body poses that were captured with several digital photo cameras could also be employed.

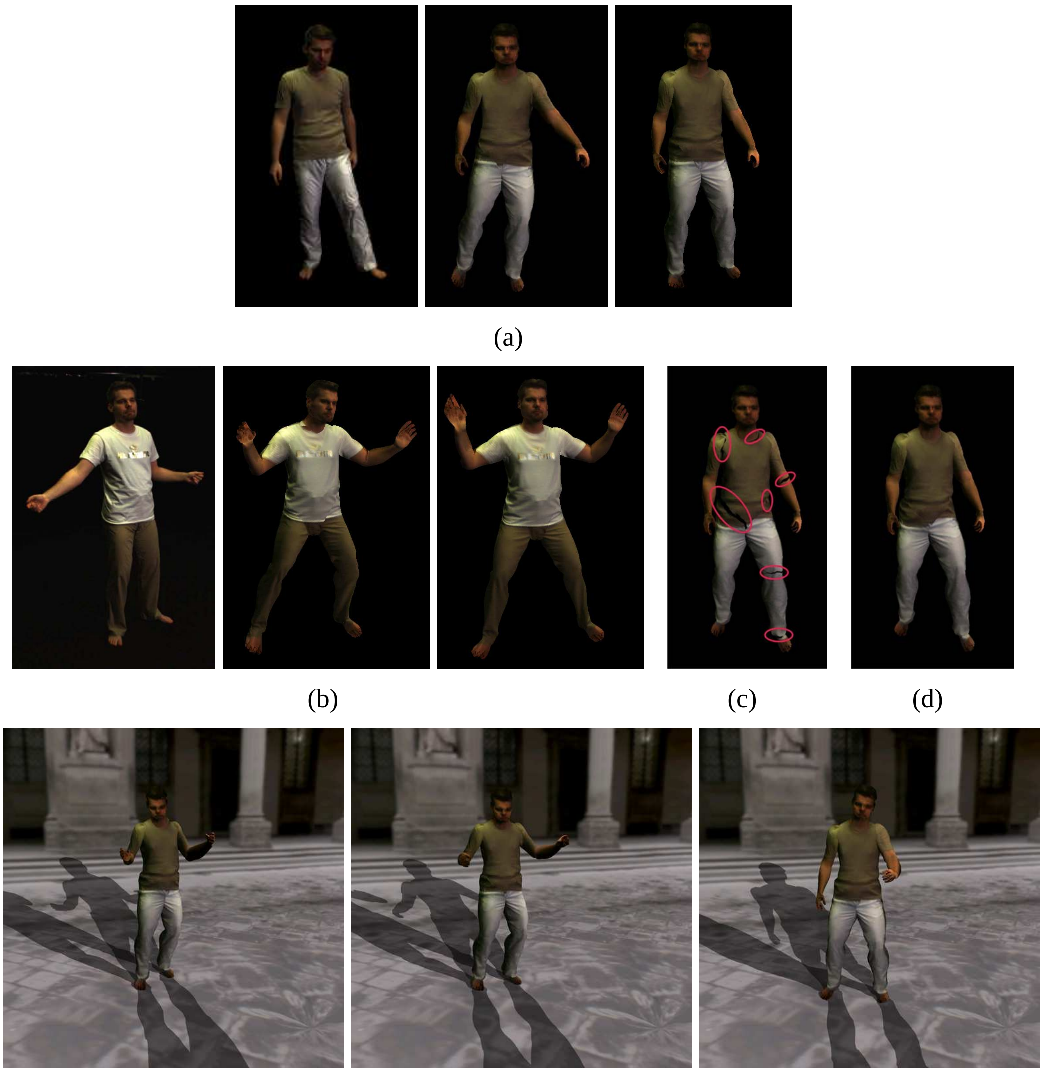
Despite the high achievable visual quality our approach is subject to a few limitations. If a surface point is never seen by any of the cameras, and if this non-visibility is not due to self-occlusion of adjacent body parts, no texture information can be reconstructed for that area. However, in practice this almost never happens. It is also not a principal limitation of our method but a problem that can not satisfactorily be solved by any image-based approach. Furthermore, the segmented geometry of our model may lead to discontinuities in the surface appearance if the model's stance is greatly different from the reference pose. This is a limitation of the segmented body model, and in the next chapters, we demonstrate that a single-skin model results in much higher quality of renditions.

Despite these limitations, we have demonstrated that we can robustly reconstruct highly realistic virtual humans based on simple-to-parameterize model from only a handful of images.

4.5 Conclusion

In this chapter we presented a fully-automatic approach to generate a personalized avatar from multi-view video data of a moving person. Our method is based on a generic human body model whose pose and geometry can be modified by optimizing only a handful of parameters. By employing dynamic multi-view image data for shape customization and texture reconstruction we obtain convincing virtual humans that exhibit a visual quality that would not have been achievable by reconstructing from single-pose photographs.

Some of the limitations in our approach arise from the use of the segmented model. In our work on relightable free-viewpoint video (Part III), we show that the use of single skin human body model can greatly enhance the reconstruction quality and results in even more realistic renditions. Similarly, unlike the static texture used to capture the appearance of the avatar, the later problems in this thesis focus on capturing dynamic surface details both in the geometry and surface texture. The work on relightable free-viewpoint video demonstrates that high quality time-varying details can not only be authentically reconstructed but can also be stored in a very compact manner in the form of time-varying surface textures. In Part IV of this thesis, we demonstrate that time-varying details can be directly captured in the geometry which results in a very high quality of reconstructed animations.



(a)

(b) (c) (d)

(e)

Figure 4.8: (a),(b) Comparisons between real and the virtual humans: In each triplet, the image on the left shows one of the multi-view video frames used to reconstruct an avatar. The two images to the right show renditions of the virtual human in novel body poses that have not been seen by any camera. (c) If the texture is only reconstructed from one time step of video black seams may appear at segment boundaries on the rendered model. (d) Our spatio-temporal texture reconstruction method eliminates these artifacts. (e) An avatar can be used to insert a real-world person into arbitrary virtual environments.

Part III

High Quality Relightable Free-Viewpoint Video

Chapter 5

Problem Statement

This part reviews two methods that enhance the reconstruction of relightable free-viewpoint videos from multi-view video data. First, the related work in this area is reviewed, then a method for improving spatio-temporal texture registration is presented. Finally, a method for reducing bias in the reflectance estimation approach is described.

In the previous chapter we presented a method for automatic reconstruction of personalized human avatars. The method captured the true shape of the human actor and reconstructed a static surface texture for the realistic renditions of the avatar. This technology is suitable for a wider audience as a part of a general setup but not very suitable for a specific application that requires higher quality renditions. A static surface texture for rendering 3D videos would result in very unlifelike animations. Additionally, recent advances in graphics hardware and rendering algorithms enable the creation of images of unprecedented realism in real-time. In order to capitalize on these novel rendering possibilities, however, ever more detailed and accurate scene descriptions must be created. The price to pay can be measured in working hours spent to create detailed geometry meshes, complex textures, convincing shaders, and authentic animations: Apparently, scene modeling is becoming a limiting factor in realistic rendering.

In order to avoid excessive modeling times, we can again look at capturing suitable models directly from the real world objects. Image- and video-based rendering (IBR/VBR) approaches pursue this notion, aiming at automatically generating visually authentic computer models from real world-recorded objects and events [Kanade97]. Many of these techniques show how to interactively render photo-realistic views from real world-captured, dynamic scenes (see also Sect. 5.1). While the ability to realistically display dynamic events from novel viewpoints has by itself already a number of intriguing applications, the next step is to use objects that have been captured in the real world for augmenting virtual scenes. To import a real-world object into surroundings different from the recording environment, however, its appearance must be adapted to the new illumination situation. To do so, the bi-directional reflectance distribution function (BRDF) must be known for all object surface points. Data-driven [Debevec00, Matusik03] as well as model-based [Marschner98, Lensch03] methods have been proposed to recover and represent the BRDF of real-world materials. Unfortunately, these methods cannot be directly applied to dynamic objects exhibiting time-varying surface geometry and constantly changing local illumination.

Theobalt et al. [Theobalt05b] [Theobalt05a] presented an approach that jointly captures shape, motion and time-varying surface reflectance of people. In their work, they used a silhouette based analysis through synthesis method to capture the shape and motion of the human actor [Carranza03]. They also presented an image-based warping method to enhance the multi-view photo consistency in the presence of inexact body geometry. Finally, they also presented methods to estimate surface reflectance properties and time-varying normal field of the moving actor.

In this part of the thesis, we present some methodical improvements to their original pipeline. We slightly modified acquisition setup, and now only employ a single lighting configuration using two spot lights (Chapter 3). Moreover, we discard the segmented human body model and employ a single skin template model [Theobalt07]. Extending their work on enhancing photo-consistency, using the same framework, we introduce a spatio-temporal registration method that compensates shifting of the apparel over the body [Ahmed07a] [Theobalt07]. Their original reflectance estimation considered each point surface individually. Without modifying individual component of their work flow pipeline, we introduce a new sampling method for the surface reflectance estimation that only modifies the reflectance samples for each surface point. Our novel spatio-temporal reflectance sharing method ensures that the surface reflectance properties are not biased towards the recording environment [Ahmed07b].

Contributions of this part are:

- An algorithm to detect and compensate lateral shifting of textiles,
- a spatio-temporal reflectance sharing method that reduces bias in the estimated BRDF parameters.

5.1 Related Work

We capitalize on previous research in many areas, but primarily pick up ideas from the fields of free-viewpoint video and image-based reflectance estimation.

Research in free-viewpoint video aims at developing methods for photo-realistic, real-time rendering of previously captured real-world scenes. The goal is to give the user the freedom to interactively navigate his or her viewpoint freely through the rendered scene. Early research that paved the way for free-viewpoint video was presented in the field of image-based rendering (IBR). Shape-from-silhouette methods reconstruct geometry models of a scene from multi-view silhouette images or video streams. Examples are image-based [Matusik00, Würmlin02] or polyhedral visual hull methods [Matsuyama02], as well as approaches performing point-based reconstruction [Gross03]. The combination of stereo reconstruction with visual hull rendering leads to a more faithful reconstruction of surface concavities [Li02]. Stereo methods have also been applied to reconstruct and render dynamic scenes [Zitnick04, Kanade97], some of them employing active illumination [Waschbüsch05]. Alternatively, a complete parameterized geometry model can be used to pursue a model-based approach towards free-viewpoint video [Carranza03]. On the other hand, light field rendering [Levoy96] is employed in the 3D TV system [Matusik04] to enable simultaneous scene acquisition and rendering in real-time.

IBR methods can visualize a recorded scene only for the same illumination conditions that it was captured in. For correct relighting, it is necessary to recover complete surface reflectance characteristics.

The estimation of reflection properties from still images has been addressed in many different ways. Typically, a single point light source is used to illuminate an object of known 3D geometry consisting of only one material. One common approach is to take HDR images of a curved object, yielding a different incident and outgoing directions per pixel and thus capturing a vast number of reflectance samples in parallel. Often, the parameters of an analytic BRDF model are fit to the measured data [Sato97, Lensch03] or a data-driven model is used [Matusik03]. Reflectance measurements of scenes with more complex incident illumination can be derived by either a full-blown inverse global illumination approach [Yu99, Gibson01, Boivin01] or by representing the incident light field as an environment map and solving for the direct illumination component only [Yu98, Ramamoorthi01, Nishino01]. Reflection properties together with measured photometric data can also be used to derive geometric information of the original object [Zhang99]. Rushmeier et al. estimate diffuse albedo and normal map from photographs with varied incident light directions [Rushmeier97, Bernardini01]. A linear light source is employed by Gardner et al. [Gardner03] to estimate BRDF properties and surface normal. In [Georghiades03, Goldman04], reflectance and shape of static scenes are simultaneously refined using a single light source in each photograph.

Instead of explicitly reconstructing a mathematical reflectance model, it has also been tried to take an image-based approach to relighting. In [Hawkins04] a method to generate animatable and relightable face models from images taken with a special light stage is described. Wenger et al. [Wenger05] extend the light stage device such that it enables capturing of dynamic reflectance fields. Their results are impressive, however it is not possible to change the viewpoint in the scene. Einarsson et al. [Einarsson06] extend it further by using a large light stage, a treadmill where the person walks on, and light field rendering for display. Human performances can be rendered from novel perspectives and relit.

Our work on spatio-temporal reflectance sharing has been inspired by the reflectance sharing method of Zickler et al. to reconstruct appearance of static scenes [Zickler05]. By regarding reflectance estimation as a scattered interpolation problem, they can exploit spatial coherence to obtain more reliable surface estimate. Our algorithm exploits both spatial and temporal coherence to reliably estimate dynamic reflectance. However, since a full-blown scattered data interpolation would be illusive with our huge sets of samples, we propose a faster heuristic approach to reflectance sharing.

Chapter 6

Reflectance Sharing and Spatio-Temporal Registration for Improved 3D Video Relighting

This chapter describes two extensions to the earlier work on reconstructing relightable 3D videos. First a method for improving spatio-temporal registration of the dynamic texture is described, which detects and compensates shifting of cloth over the body surface. Finally, a reflectance sharing approach for reducing spatio-temporal bias in the estimated surface reflectance properties is presented.

6.1 Overview

Fig. 6.1 illustrates the workflow between the components of the joint shape, motion and reflectance capture approach presented by Theobalt et al. [Theobalt05a] after our two enhancements. Our proposed methods (Fig. 6.1, highlighted by magenta rectangles), sit in between the pipeline and do not modify the original surface reflectance and normal field estimation procedures.

Although the details of Theobalt et al. [Theobalt05a] original framework, as a whole, are not the subject of this thesis, for better understanding we briefly elaborate on the acquisition setup in Sect. 6.2 and employed model-based marker-less

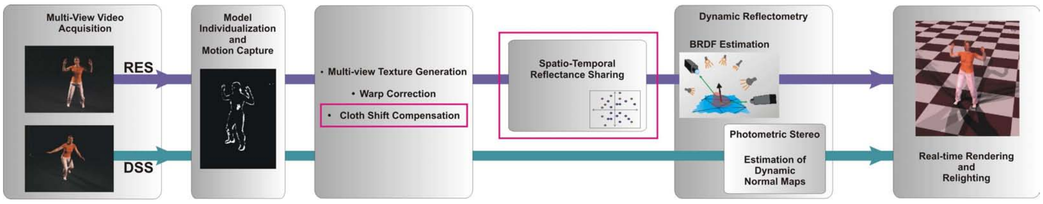

Multi-View Video Acquisition
RES
DSS
Model Individualization and Motion Capture
• Multi-view Texture Generation
• Warp Correction
• Cloth Shift Compensation
Spatio-Temporal Reflectance Sharing
Dynamic Reflectometry
BRDF Estimation
Photometric Stereo
Estimation of Dynamic Normal Maps
Real-time Rendering and Relighting

Figure 6.1: Algorithmic workflow of the original pipeline with our two enhancements (highlighted by magenta rectangles).

motion capture algorithm in Sect. 6.3. Their image-based warping method for the texture registration will be discussed in Sect. 6.4.

Our enhancement to the texture registration method, which addresses the issue of detecting and compensating the shifting of the apparel over the body surface by means of an automatic cloth shift detection procedure is presented in Sect. 6.5.

In order to motivate for our spatio-temporal reflectance sharing method we will describe the basic principals of dynamic reflectometry that is used in the original pipeline in Sect. 6.6.

The fixed arrangement of the camera and light sources in the acquisition system can lead to biased sampling of the reflectance space. To reduce this bias a novel spatio-temporal reflectance sharing method that combines dynamic reflectance samples from different surface points of similar material during BRDF estimation of each surface element in texture space (texel) is presented in Sect. 6.7.

6.2 Acquisition

Inputs to Theobalt et al. [Theobalt05a] method are synchronized multi-view video sequences captured with eight calibrated cameras that feature 1004x1004 pixel image sensors and record at 25 fps. The cameras are placed in an approximately circular arrangement around the center of the scene which is illuminated by two calibrated spot lights. Since the BRDF estimation from the estimation of the dynamic normal maps is conceptually separated, two types of multi-view video sequence for each person and each type of apparel are recorded. In the first type of sequence, the so-called reflectance estimation sequence (RES), the person performs a simple rotation if front of the acquisition setup. One RES is recorded for each actor and each type of apparel, and it is later used to reconstruct the per-texel BRDF models. In the second type of sequence, the so-called dynamic scene sequence (DSS), the actor performs arbitrary movements. Several DSS are

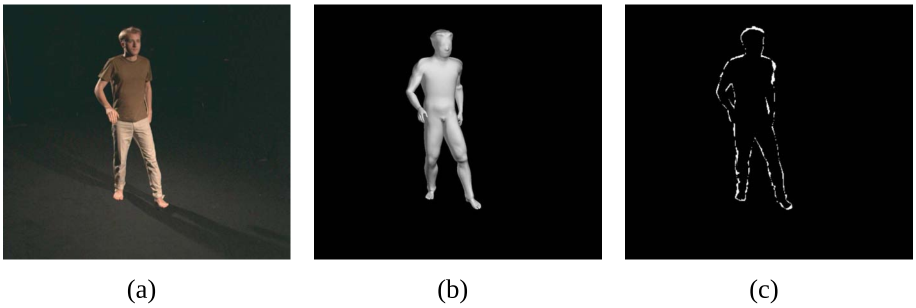

(a) (b) (c)

Figure 6.2: (a) Input frame, (b) body model in same pose, and (c) silhouette matching.

recorded, and from each of them, one relightable free-viewpoint video clip is reconstructed. Also the second component of their dynamic reflectance model, the dynamic normal maps, are reconstructed from each DSS.

6.3 Reconstructing Dynamic Human Shape and Motion

An analysis-through-synthesis approach is employed to capture both shape and motion of the actor from multi-view video footage without having to resort to optical markers in the scene. It employs a template human body model consisting of a kinematic skeleton and a single-skin triangle mesh surface geometry [Carranza03, Theobalt04]. In an initialization step, the shape and proportions of the template are matched to the recorded silhouettes of the actor. After shape initialization, the model is made to follow the motion of the actor over time by inferring optimal pose parameters at each time step of video using the same silhouette matching principle, Fig. 6.2. This dynamic shape reconstruction framework is applied to every time step of each captured sequence, i.e. both RES and DSS. This way for each time step of video the orientation of each surface point with respect to the acquisition setup is known, which is a precondition for the subsequent dynamic reflectometry procedure.

Given the moving geometry, all input video frames and all corresponding data required for reflectance estimation (e.g. image samples, normals, visibility information, light vectors) are transformed into sequences of textures. Throughout the work, 1024x1024-texel texture maps are used, where the texel is a surface element in the texture space. The model's surface is parameterized over a 2D square. For the BRDF and time-varying normal estimation a parameterization with minimal

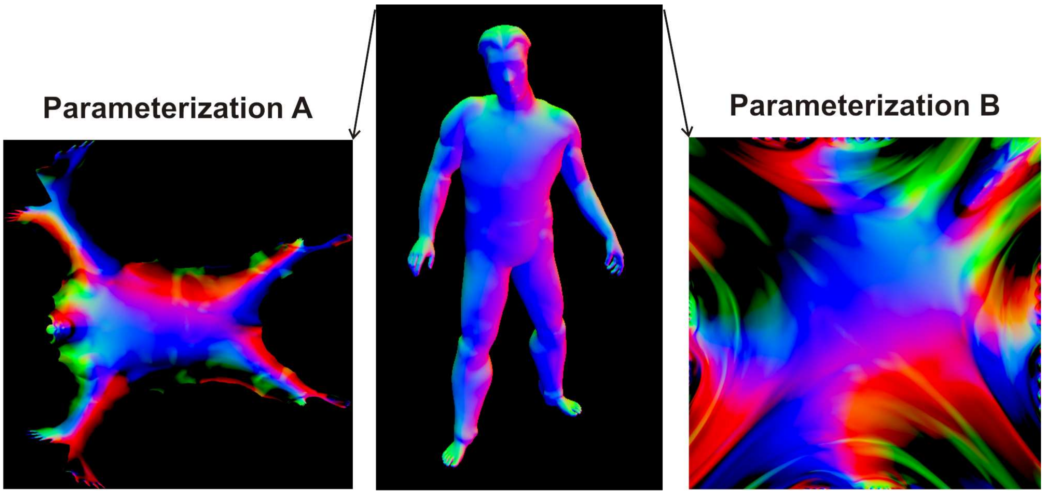


Figure 6.3: Human body model and the corresponding texture parameterizations (colors=normals encoded in RGB).

surface distortion is required. To achieve this, a parameterization (Parameterization A) that leaves the mesh boundary free and results in fairly uniform distribution of samples [Zayer05b] is employed, Fig. 6.3. For the purpose of cloth shift detection, on the other hand, a parameterization (Parameterization B) with a fixed square boundary is preferred, Fig. 6.3.

6.4 Warp Correction

Although the body model initialization procedure yields a faithful representation of the person's true geometry, small inaccuracies between the real human and its digital counterpart are inevitable. Due to these geometry inaccuracies, pixels from different input views may get mapped to the same texel position in different MVV textures, even though they do not correspond to the same surface element of the true body geometry.

One common strategy to enhance model-to-texture consistency is to deform the geometry until an overall photo-consistency measure is maximized. For instance, [de Aguiar05, Kück04] deform model geometry from input images by jointly optimizing multi-view silhouette- and photo-consistency. In a similar line of thinking, [Hernández04] jointly employs silhouette and stereo constraints to deform scene geometry from images. Geometry deformation-based optimization, however, tends to give unstable results, in particular due to nonlinear optimizations that are normally required.

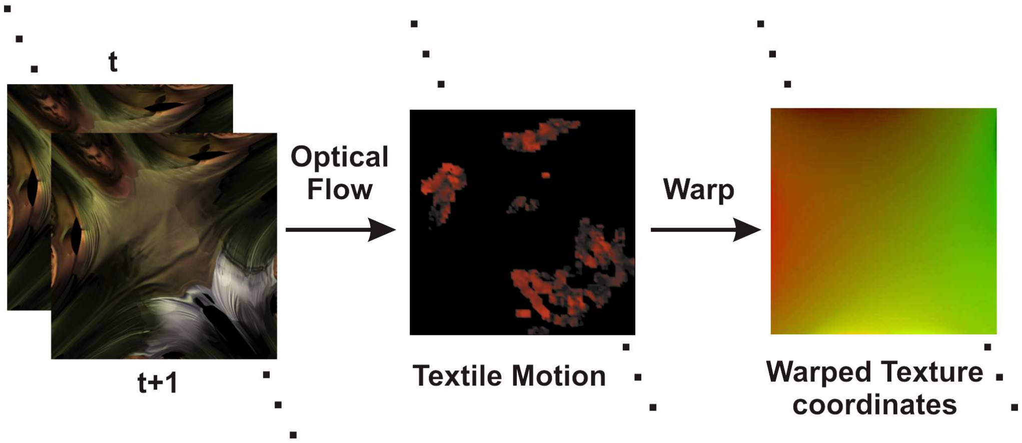


Figure 6.4: Cloth shift between two subsequent combined textures t and $t+1$ (in parameterization B) is found via optical flow. In the middle, detected shifted areas are shown in red. Finally, the shift is encoded in the warped texture-coordinates.

Theobalt et al. [Theobalt05a] presented an optical flow based image-warping approach that instead of moving surface elements to their correct locations in 3D, move the image pixels within the 2D input image planes until they all become photo-consistent given the available geometry. To establish per-pixel correspondences, the warping operation itself is based on the optical flow [Lucas81] between the reference image and the target image. A regular 2D triangle mesh is superimposed on the reprojected model image, per-vertex displacements are derived from the optical flow values, and the mesh is deformed accordingly via thin-plate spline interpolation [Farin99]. Finally, the warped reprojected image is created on the GPU.

In the next section we will present our cloth shift detection and compensation approach making use of this image warping technique.

6.5 Cloth Shift Detection and Compensation

BRDF estimation procedure presented in [Theobalt05a] assumes that a static set of material parameters can be assigned to each point on the model's surface. In reality, however, this assumption does not hold since the apparel of the person shifts across the body while she is moving. Prior to surface reflectance, we thus estimate the motion of the apparel over time and register all surface textures against a reference texture. Please note that we can still reproduce the true shifting of the apparel during rendering by making the cloth motion information accessible to the renderer. During display, the renderer warps the estimated static BRDF textures back into their true position. We employ the following method to detect the shifting of cloth in the texture domain, Fig. 6.4:

Our reference time step is the last frame of the RES because after this frame the actor goes on performing for the DSS. MVV textures for this frame and all the frames of the DSS are resampled into a weightedly blended single texture in parameterization B. Cloth shift is detected by computing an optical flow field between subsequent blended textures. This flow field describes for each texel how it shifts across the body surface. This texel motion information is made accessible to the reflectance estimation process as well as the renderer in the form of warped texture coordinates.

Please remember that we use texture parameterization A for sampling, but texture parameterization B for cloth shift computation. We make use of this parameterization because it has well defined boundaries in the texture space, unlike the free boundary representation in parameterization A where the there is no well defined correspondence between the boundary pixels of each side of the cut. Therefore, we project the parameterization A texture coordinates of the reference frame into parameterization B to obtain the texture coordinate image $I_{CoordAB}(0)$. Given the accumulated displacements from the pairwise flow fields we can deform $I_{CoordAB}(0)$ such that it matches the texture at each time of input video using the method from Sect. 6.4. Note that it is essential to compute the cloth motion relative to the previous frame and accumulate the displacement over time. Only this way, appearance differences due to lighting changes can be robustly handled.

The sequence of deformed texture coordinates enables us to account for cloth shifting during estimation and rendering, although only a static set of BRDF parameters are estimated.

6.6 Dynamic Reflectometry

Dynamic reflectance model presented in presented in [Theobalt05a] consists of two components, a static parametric isotropic BRDF for each surface point [Phong75, Lafortune97a], as well as a description of the time-varying direction of the normal at each surface location. The first component of the reflectance model is reconstructed from the video frames of the reflectance estimation sequence, the second component is reconstructed from each dynamic scene

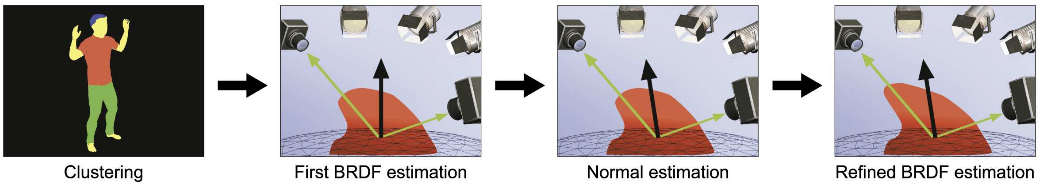


Figure 6.5: Steps to estimate per-texel BRDFs.

sequence. BRDF reconstruction is formulated as an energy minimization problem in the BRDF parameters [Theobalt05a]. This minimization problem has to be solved for each surface point separately.

In the first step the reflectance samples are clustered to determine what material a surface element, i.e., each texel in the texture maps, belongs. The number of materials is preset a priori. A straightforward color based clustering approach that considers raw color values is employed. The clustering step is important because unlike the diffuse BRDF which is measured for each texel separately, the specular component of the BRDF is estimated for each cluster. The energy functional measures the error between the recorded reflectance samples of the point under consideration and the predicted surface appearance according to the current BRDF parameters. Estimates of the BRDF parameters are used to refine the surface geometry by keeping the reflectance parameters fixed and minimizing the same functional in the normal direction, Fig 6.5. Once the BRDF parameters have been recovered from the RES, a similar minimization procedure is used to reconstruct the time-varying normal field from each DSS.

In the original pipeline, as it was summarized above, BRDF parameters were estimated for each surface point by taking only reflectance samples of this particular point itself into account. In the following, we present a novel spatio-temporal sampling scheme that reduces the risks of a bias in the BRDF estimates by also taking into account dynamic reflectance samples from other surface points with similar material properties.

6.7 Spatio-Temporal Reflectance Sharing

Although Theobalt et al. [Theobalt05a] showed that it is feasible to reconstruct dynamic surface reflectance properties using only eight cameras and a static set of light sources, this type of sensor arrangement leads to a biased sampling of the reflectance space. By looking at its appearance from each camera view over time, we can generate for each surface point, or equivalently, for each texel $\vec{x}$ a set of N

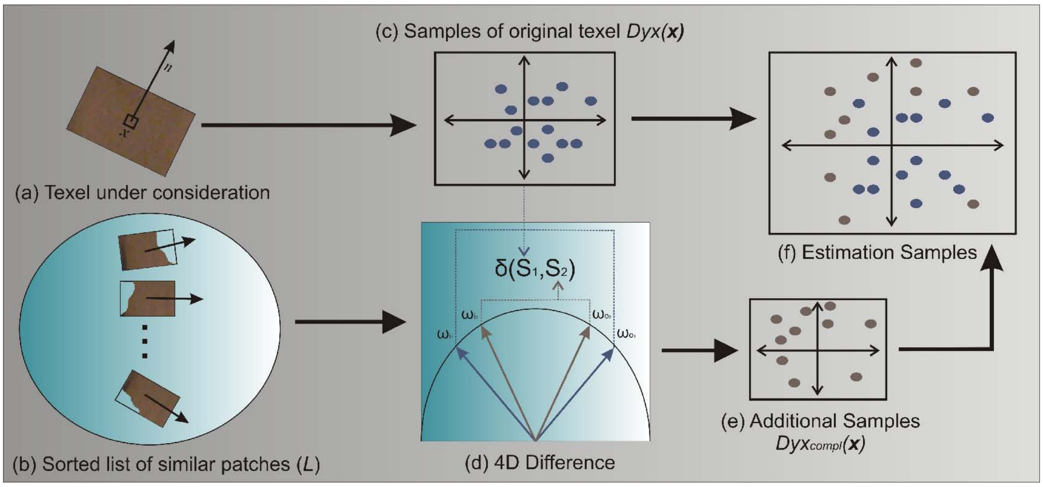

(c) Samples of original texel Dyx(x)
(a) Texel under consideration
δ(S1,S2)
ωi
ωo
(f) Estimation Samples
(e) Additional Samples Dyxcompl(x)
(b) Sorted list of similar patches (L)
(d) 4D Difference

Figure 6.6: Weighted selection of samples. Samples from the similar patches are added to the samples from the original texel. Additional samples are selected according to a weighting criteria that is based on their maximum angular difference from the samples of original texel.

appearance samples

$$\text{Dyx}(\vec{x}) = \{S_i \mid S_i = (I_i, \hat{l}_i, \hat{v}_i), i \in \{1, \ldots, N\}\} \tag{6.1}$$

Each sample S_i stores a tuple of data comprising of the captured image intensity I_i (from one of the cameras), the direction to the light source $\hat{l}_i$, and the viewing direction $\hat{v}_i$. Please note that only if a point has been illuminated by exactly one light source, a sample is generated. If a point is totally in shadow, illuminated by two light sources, or not seen from the camera, no sample is created. Our acquisition setup comprising of only 8 cameras and 2 light sources is comparably simple and inexpensive. However, the fixed relative arrangement of cameras and light sources may induce a bias in $\text{Dyx}(\vec{x})$. There are two primary reasons for this:

- Due to the fixed relative arrangement of cameras and light sources, each surface point is only seen under a fixed number of half vector directions $\hat{h} = \hat{l} + \hat{v}$.
- Even if the person performs a very expressive motion in the RES, samples lie on "slices" of the hemispherical space of possible incoming light and outgoing viewing directions.

Both of these factors possibly lead to BRDF estimates that may not generalize well to lighting conditions that are very different to the acquisition setup.

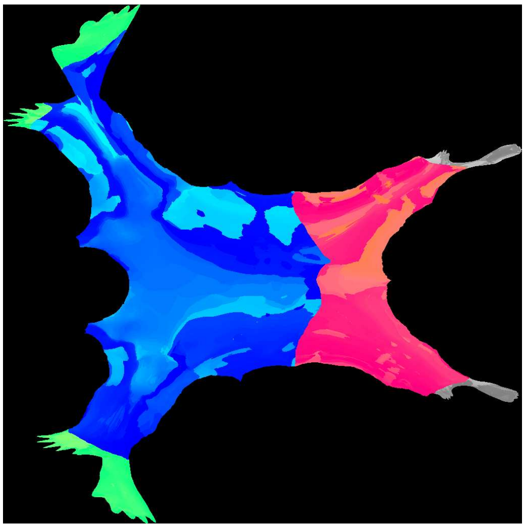


Figure 6.7: Texture-space layout of surface patches. Patches of same material are clustered according to the average normal direction. For this illustration, patches of the same material are colored in the same overall tone (e.g. blue for the shirt) but different intensities.

By means of a novel spatio-temporal sampling strategy, called spatio-temporal reflectance sharing, we can reduce the bias, Fig. 6.6. The guiding idea behind this novel scheme is to use more than the samples $\mathrm{Dyx}(\vec{x})$ that have been measured for the point $\vec{x}$ itself while the BRDF parameters for the point $\vec{x}$ are estimated. The additional samples, combined in a set $\mathrm{Dyx}_{\mathrm{compl}}(\vec{x})$, stem from other locations on the surface that are made of similar material. These additional samples have potentially been seen under different lighting and viewing directions than the samples from $\mathrm{Dyx}(\vec{x})$ and can thus expand the sampling range. It is the main challenge to incorporate these samples into the reflectance estimation at $\vec{x}$ in a way that augments the generality of the measured BRDFs but does not compromise the ability to capture spatial variation in surface appearance.

By explaining each step that is taken to draw samples for a particular surface point $\vec{x}$, we illustrate how we attack this challenge:

In a first step, the surface is clustered into patches of similar average normal directions and same material, Fig. 6.7. Materials are clustered by means of a simple k-means clustering using average diffuse colors. The normal direction $\hat{n}$ of $\vec{x}$ defines the reference normal direction, Fig. 6.6a. Now, a list L of patches consisting of the same material as $\vec{x}$ is generated. L is sorted according to increasing angular deviation of average patch normal direction and reference normal direction, Fig. 6.6b. Now, n_p many patches $P_0, ..., P_{n_p}$ are drawn from L by choosing every

lth list element. From each patch, a texel is selected at random, resulting in a set of texels, $T = \vec{x}_{P_0}, ..., \vec{x}_{P_{n_p}}$. The set of texels T has been selected in a way that maximizes the number of different surface orientations. From the reflectance samples associated with texels in T, we now select a subset $\text{Dyx}_{\text{compl}}(\vec{x})$ that maximizes the coverage of the 4D hemispherical space of light and view directions. In order to decide which samples from T are potential candidates for this set, we employ the following selection mechanism.

A weighting function $\delta(S_1, S_2)$ is applied that measures the difference of two samples $S_1 = (\hat{l}_1, \hat{v}_1)$ and $S_2 = (\hat{l}_2, \hat{v}_2)$ in the 4D sample space as follows:

$$\delta(S_1, S_2) = \Delta(\hat{l}_1, \hat{l}_2) + \Delta(\hat{v}_1, \hat{v}_2) \tag{6.2}$$

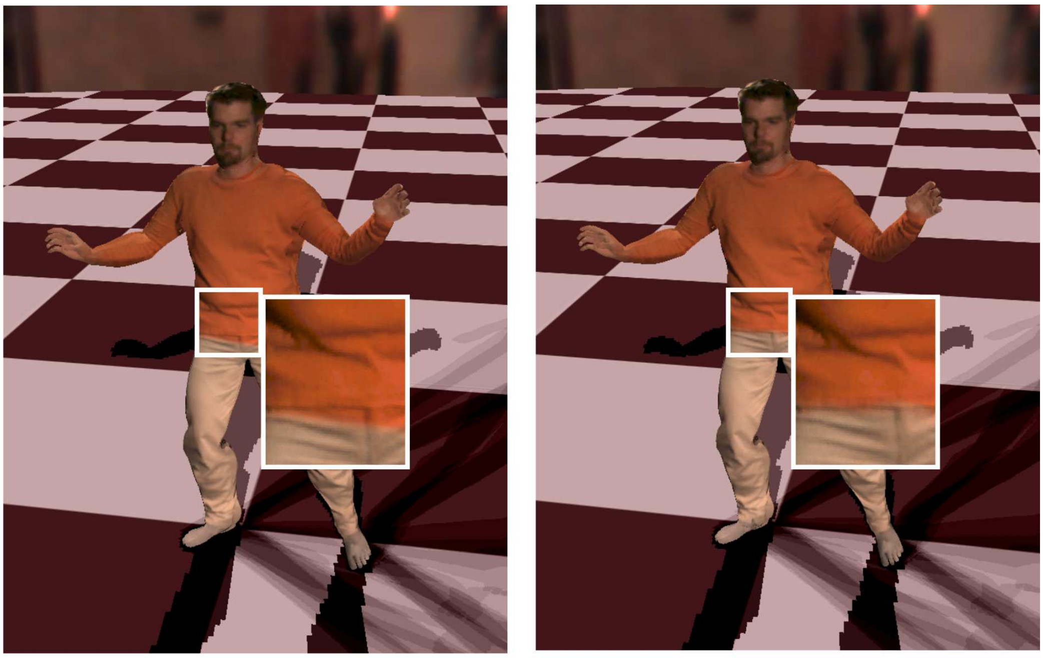
where Δ denotes the angular difference between two vectors. We employ δ to select for each sample S_r in T its closest sample S_{closest} in $\text{Dyx}(\vec{x})$, i.e. the sample for which $\omega_{S_r} = \delta(S_r, S_{\text{closest}})$ is minimal, Fig. 6.6d. Each sample S_r is now weighted by ω_{S_r}. Only the $\lceil \alpha N \rceil$ samples from T with the highest weights eventually find their way into $\text{Dyx}_{\text{compl}}(\vec{x})$, Fig. 6.6e. In Sect. 6.8, we show that at around 34% we get maximum improvement from additional samples, therefore we set the value of $\alpha = 0.66$. The BRDF parameters for $\vec{x}$ are estimated by taking all of the samples from $\text{Dyx}(\vec{x}) \bigcup \text{Dyx}_{\text{compl}}(\vec{x})$ into account, Fig. 6.6f. For estimation, we make use of the original dynamic reflectometry method detailed in Sect. 6.6.

6.8 Results and Validation

In the previous sections, we presented two enhancements to the original work on joint motion and reflectance estimation scheme of Theobalt et al. [Theobalt05a]. We presented a spatio-temporal registration technique and a novel spatio-temporal reflectance sharing method for enhancing the quality of relightable free-viewpoint videos.

We have validated our approach by visual inspection and quantitative evaluation. We have processed 2 different input sequences using Phong and Lafortune BRDFs. They cover 2 different human subjects,2 different types of apparel, and comprise 150 to 350 frames each. For numerical verifications, we restrict ourselves to Phong sequences.

For texture registration, we assess the multi-view warping quality by comparing the image differences between reference views and reprojected model views before and after the warp. The local registration improvements in single image pairs lead to a global improvement in multi-view texture-to-model consistency. With



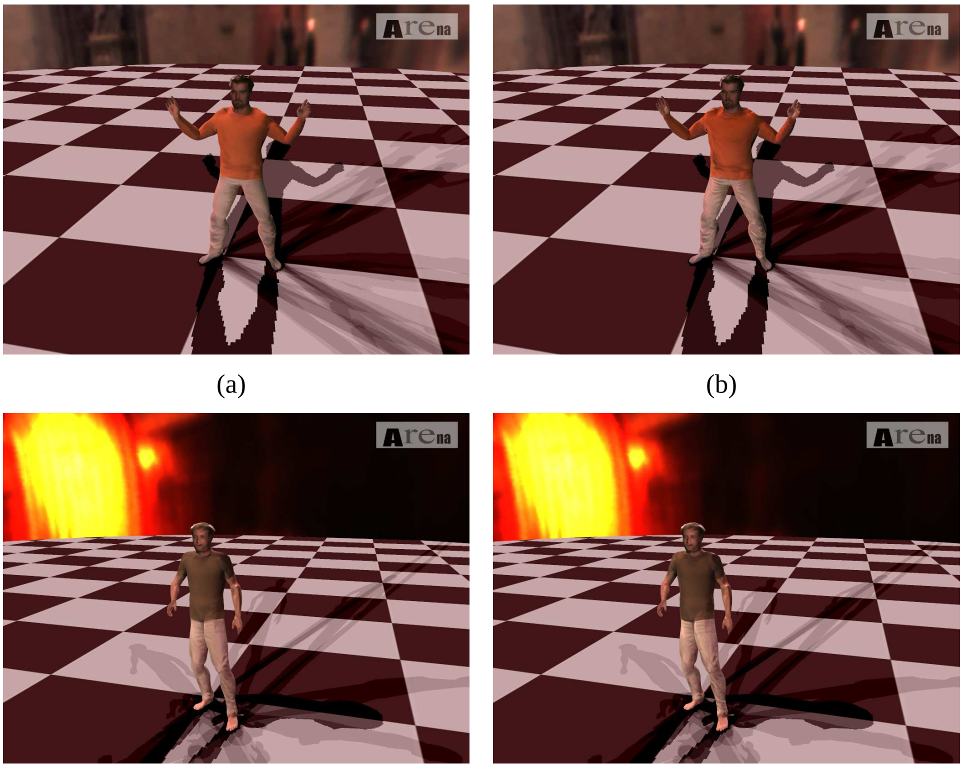
(a) Without Cloth Shift Compensation

(b) With Cloth Shift Compensation

Figure 6.8: Screen-shots of relightable 3D videos rendered under captured real-world illumination. (a) Without cloth-shift detection, the seam of the t-shirt is rendered incorrectly. (b) With cloth shift detection, it is reproduced accurately.

respect to one input stream not used for reconstruction we have obtained a peak-signal-to-noise-ration improvement of 0.2 dB using cloth shift compensation. On a Pentium IV 3.0 GHz, cloth shift compensation takes around 35s for each time step of the video. Although these quantitative improvements may appear small, their influence on the overall visual is quality is well-pronounced. Fig. 6.8 shows how it corrects the movement of seams of the shirt over the surface.

Although cloth shift compensation and warp correction lead to visual improvements in the majority of cases, isolated local deteriorations are still possible. Cloth shift detection, for example, sometimes erroneously classifies evolving wrinkles as shifting of apparel. Also, in case of strongly incorrect body geometry, warp correction may induce noticeable discontinuities on the surface, e.g. due to changing reference cameras or visibility boundaries. Luckily, for the types of scene we intend to handle, body shape is already so close to the true geometry that these discontinuities play no significant role. We nonetheless leave the decision if ei-



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

Figure 6.9: Comparison of renditions under captured real-world illumination such as the St Peter's Basilica environment map (a),(b) and the Grace Cathedral environment (c),(d) courtesy of Paul Debevec. One can see that compared to renditions obtained without spatio-temporal reflectance sharing ((a) (c)), subtle surface details are much better reproduced in the renditions obtained with spatio-temporal reflectance sharing ((b) (d)). A high quality video comparison can be seen in http://www.mpi-inf.mpg.de/˜nahmed/Mirage07.avi.

ther of the two methods are used to the user.

We verified our spatio-temporal reflectance sharing method both visually and quantitatively, and show that the novel reflectance sampling method leads to BRDF estimation that generalizes better to lighting conditions different from the acquisition setup. Fig. 6.9 shows a side-by-side comparison between the results obtained with and without spatio-temporal reflectance sharing. Both human subjects are rendered under real world illumination using HDR environment maps. One can see that with the exploitation of spatial coherence, more surface detail is preserved under those lighting conditions which are strongly different from acqui-

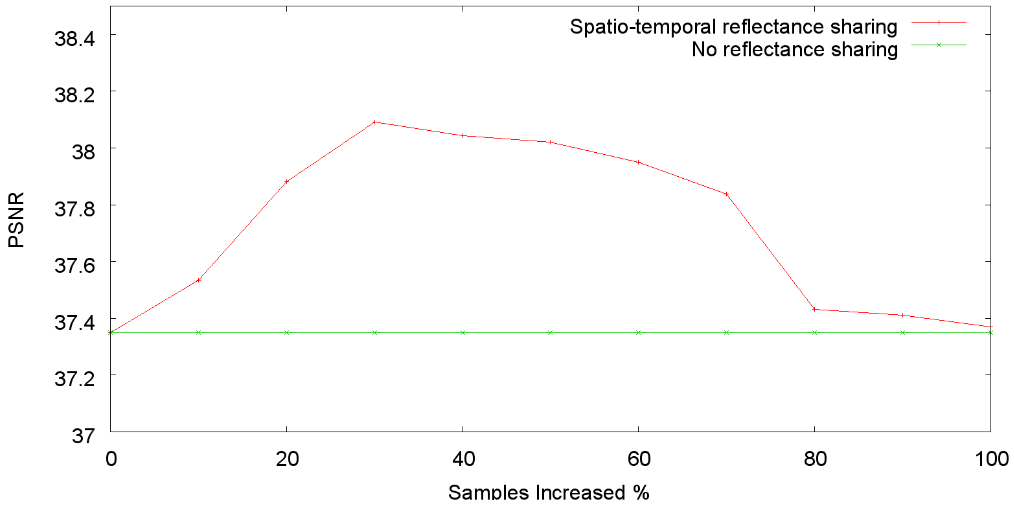


Figure 6.10: PSNR values with respect to ground truth for different numbers of additional samples $\mathrm{Dyx}_{\mathrm{compl}}(\vec{x})$**.**

sition setup. The difference is more pronounced in the accompanying video which can be downloaded from http://www.mpi-inf.mpg.de/~nahmed/Mirage07.avi.

In addition to visual comparison, we also validated the method by comparing the average peak-signal-to-noise-ratio with respect to input video stream obtained under two calibrated lighting conditions as described above. We reconstructed the BRDF of the test subject under lighting setup LC B with and without our new reflectance sampling. Subsequently, we calculated the PSNR with the ground truth images of the person illuminated under setup LC A. Using our novel sampling method, we have estimated surface reflectance using different percentages of additional samples. For each case, we computed the PSNR with respect to the ground truth. Fig. 6.10 shows the results that we obtained. Note that the graph of the original method (green line) is constant over the increasing number of samples just for the illustration purpose because it only considers the samples from a single texel. With spatio-temporal reflectance sharing (red line) both results are exactly the same in the beginning as no additional samples are considered, but it can be seen that the PSNR improves as additional samples are taken into account. We get a peak at around 30%-40% of additional samples. With the inclusion of more samples the PSNR gradually decreases as the ever increasing number of additional samples compromises the estimation of the reflectance's spatial variance. At maximum, we obtain a PSNR improvement of 0.75 dB. Although we have performed the PSNR evaluation only for one sequence, we are confident that for others it will exhibit similar results. This assumption is further supported by the more compelling visual appearance obtained for all the other test data that we have used.

6.9 Conclusion

In this chapter we presented two improvements to the original work of Theobalt et al. [Theobalt05a] on joint shape, motion and reflectance capture. We presented an image-based spatio-temporal registration technique that compensates for the shifting of cloth across the body's surface enables high-quality reconstruction of model-based relightable 3D videos. Quality improvements in the real-time renderings were shown both quantitatively and visually.

Our spatio-temporal reflectance sharing method reduces the bias in BRDF estimation for dynamic scenes. Our algorithm exploits spatial coherence by pooling samples of different surface location to robustify reflectance estimation. In addition, it exploits temporal coherence by taking into consideration samples from different steps of video. Despite the spatial-temporal resampling, our algorithm is capable of reliably capturing spatially-varying reflectance properties. By means of spatio-temporal reflectance sharing, we obtain convincing 3D video renditions in real-time even under lighting conditions which differ strongly from the acquisition setup.

Our methods are independent of this specific framework and can be employed independently. In the overall pipeline of the earlier work they do not change the estimation procedure in dynamic reflectometry. In the next part of this thesis we will continue with the improvements and make use of estimated dynamic normal field in the original method and transfer it into highly detailed time-varying geometry deformations.

Part IV

Highly Detailed Dynamic Geometry via Simultaneous Reflectance and Normal Capture

Chapter 7

Problem Statement

This part proposes a new data fusion framework for adding high quality details to dynamic geometry. The animated template mesh used in relightable free-viewpoint video does not exhibit subtle dynamic surface details, e.g. wrinkles in clothing. These changes were captured in the time-varying normal field. Using the estimated reflectance field and the dynamic normal field, high quality time-varying details are added to the template geometry. First the related work in this area is discussed and later the solution to get the detailed dynamic geometry is proposed.

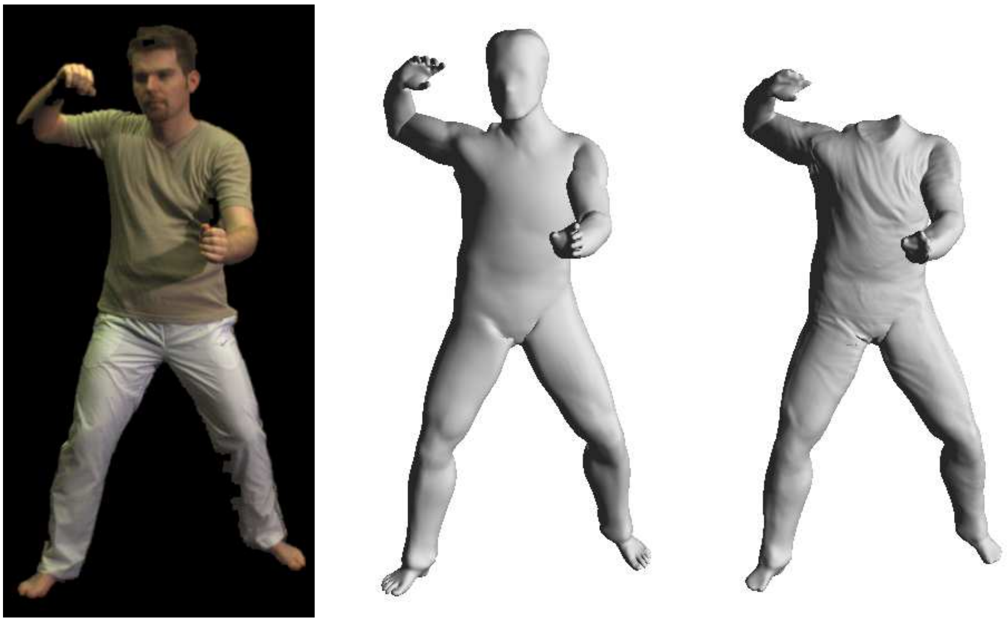
In the previous part of this thesis we presented two improvements to the work of Theobalt et al. [Theobalt05a] for reconstructing high quality relightable free-viewpoint video. In their work they first performed marker-less motion capture on the input data in order to make a coarse kinematic template (shown in Fig. 7.1b) follow the motion of the actor and also captured the true shape of the actor. Subsequently, a reflectance model for each point on the surface was reconstructed, and was exploited to measure a dynamic surface normal field parameterized over the smooth template mesh. The dynamic normal field was applied as a bump map over the smooth template geometry. While the animated template along with the BRDF parameters and dynamic normal field was sufficient for rendering relightable free-viewpoint video, one of the limitation of the original template geometry that it did not incorporate true time-varying geometry, remained. For realistic renderings of 3D videos from novel viewpoints, having the true detailed geometry would result in the accurate appearance as opposed to only using the bump maps. Hand-crafting detailed moving scene geometry is a cumbersome process, as it requires tedious manual work or computationally expensive numerical simulations (e.g. for clothing). The development of scanning devices that deliver fine-grained shape models of at least static scenes has therefore greatly facilitated animation production. Unfortunately, capturing high-quality time-varying shape of dynamic scenes at the same level of fidelity is still a big challenge. First approaches to reach this goal were based on active video-based measurement, such as structured light, or employed a combination of visual hull and stereo. While the former approaches are merely usable for small-scale scenes (e.g. faces) and interference makes multi-view recording difficult, stereo approaches often fall short in delivering the high level of accuracy that computer animation requires (Sect. 7.1).

In this part of the thesis, we propose a new method to passively capture highly-detailed dynamic surface geometry of humans from multiple video recordings under calibrated lighting [Ahmed08a]. We make use of the previous work on relightable free-viewpoint video ([Theobalt05a] and Chapter 6) and present a solution of the difficult problem of converting a potentially noise-contaminated normal field parametrized over an arbitrarily shaped smooth surface into highly-detailed time-varying scene geometry. The first contribution of this method is an improvement over our original surface reflectance and normal estimation approach which now employs robust statistics to handle sensor noise more faithfully, Sect. 8.3. The second and most important contribution is a new spatio-temporal deformation framework that enables us to transform the moving template geometry and the time-varying normal field into true spatio-temporally varying scene geometry that reproduces geometric surface detail at millimeter-scale accuracy, Sect. 8.4. Standard normal field integration schemes are not feasible in this setting as they often perform poorly in the presence of noise and as they do not easily generalize to the case of arbitrarily oriented base surfaces in 3D. In contrast, we formulate the problem as a spatio-temporal Markov Random field such that we can reconstruct fine-grained geometry that is spatially accurate, as well as temporally smooth, even if the input was affected by noise.

We demonstrate and validate the accuracy of our method based on several real-world sequences, Sect. 8.5.

7.1 Related Work

Most systems that can capture dynamic scene geometry at millimeter scale accuracy are restricted to confined spatial volumes, e.g. structured light systems for facial performance capture [Zhang04]. Mainly due to interference and spa-



(a) (b) (c)

Figure 7.1: Input video frame (a), smooth 3D template model in same pose (b), our detailed 3D surface model with true geometric detail such as wrinkles on the shirt (c).

tial resolution issues, it is hard to apply these methods for capturing humans from multiple views. While a combination of shape-from-silhouette and stereo is one way to approach the latter scenario, the inherent difficulty and lack of robustness in stereo make it hard to achieve very high accuracy and resolution [Hernández04, Starck06].

An alternative to multi-view stereo reconstruction the potential to capture fine-grained surface detail is photometric stereo, which is a variant of shape-from-shading. In photometric stereo one makes assumptions about surface reflectance properties to recover normal orientation from images taken under varying lighting [Woodham89, Zhang99]. It has also been tried to simultaneously estimate reflectance (e.g. BRDF information) and normal data from a variety of 2D images which were taken under calibrated lighting [Georghiades03, Goldman04]. In this single 2D view case, normal field integration schemes can be applied to transform orientation data into true highly-detailed height values [Frankot88, Agrawal06]. Chang et al [Chang07] used level set methods to integrate multi-view normal fields.

While it is feasible to estimate BRDF and normal orientation also for more general static 3D objects that were photographed under a variety of viewpoints and light directions [Lensch03], the deformation of geometry based on normals parametrized over a general 3D shape is non-trivial. Standard integration schemes (assuming orthographic projection and height fields that are parametrized over a plane) are not applicable anymore since absolute 3D position has to be recovered and coherence of the displacements over non-planar geometry needs to be assured.

One way to attack this problem is to measure 3D position approximately, e.g. by stereo or structured light scanning, and use normal information obtained via shape from shading to improve the initial position estimates and the degree of surface detail [Hernández07]. While early work in this direction produced comparably coarse 3D geometry [Fua94, Lange99], the work by Nehab et al. [Nehab05] produces detailed models of static objects by refining scanned 3D point positions until photometrically measured normals are well approximated. Jones et al. [Jones06] applied the latter technique to improve captured dynamic face geometry, but they did not formulate it as a spatio-temporal problem nor does their setup scale easily to larger scenes. Hernandez et al. [Hernández07] use structured light scanning to produce very high quality dynamic geometry, whereas in comparison we propose a spatio-temporal coherent passive method.

We capitalize on this idea as well but develop a more advanced reconstruction approach suitable for large-scale dynamic scenes. In contrast to previous work, our approach generates geometry that is accurate and detailed at each time step, *and* that is coherently deforming over time. We also incorporate characteristics of measurement noise into the reconstruction process by posing our problem as a spatio-temporal Markov Random Field (MRF).

The starting point is the work by Theobalt et al. [Theobalt05a] on the relightable free-viewpoint video, and our improvement to their work that were presented in the previous part of this thesis. In that work they captured shape, motion, reflectance and time-varying normals of human actors from only eight of synchronized video recordings under calibrated lighting. The method parametrizes shape, motion, and reflectance based on a smooth template body model that lacks any geometric detail. In this work, we improve the previous reflectance and normal field estimation approach by using robust statistics. We then propose a new spatio-temporal MRF framework which transforms smooth geometry and normals into highly detailed dynamic scene geometry even in the presence of notable measurement noise. As we can process normal fields over arbitrarily shaped time-varying base surfaces in 3D, we can capture time-varying geometry at detail levels comparably higher by other related approach, such as purely stereo-based reconstruction methods mentioned earlier.

Chapter 8

Reconstructing High Quality Time-Varying Geometry

This chapter describes a passive approach to capture true time-varying scene geometry in large acquisition volumes from multi-view video. First, an improved method for estimating surface reflectance properties and a time-varying normal field using a coarse template shape is described. Later, a statistical method to transform the captured normal field into true 3D displacements is presented. Output is a spatio-temporally coherent geometry that models even the slightest dynamic shape detail as true 3D geometry displacements.

8.1 Overview

Our goal is to passively reconstruct accurate and highly detailed dynamic surface geometry of humans from only eight synchronized video recordings, Sec. 8.2 and Fig. 8.1. Starting point for the our methods is the earlier work by Theobalt et al. [Theobalt05a], which gives us the tracked motion of the actor in input video recordings. They also parametrizes dynamic scene geometry in the form of an adaptable kinematic body template with smooth surface geometry that lacks fine surface details. Using the video sequences recorded under calibrated lighting, they also estimated surface reflectance properties, i.e. per-surface-point BRDfs, as well as dynamic normal maps. We extended their work in the previous part of this thesis, and we pick up and extend those ideas, such that we can use the same acquisition setup, starting with the coarse template model, and add deformations to the smooth geometry to acquire highly detailed dynamic geometry. We demonstrate our method on a variety of real world sequences.

To achieve our goal, we first modify the original BRDF estimation pipeline by including robust statistics into the reconstruction framework, Sect. 8.3. This allows us to model the non-Gaussian measurement noise more faithfully. Thereafter, we estimate dynamic normal (bump maps) from the input video sequences that are defined over the smooth template geometry. Finally, we develop a spatio-temporal Markov-Random-Field-based surface refinement procedure which is one of the first to enable integration of normal fields on arbitrarily shaped time-varying template geometry. Our new spatio-temporal framework captures at the same time spatially accurate and temporally smooth geometry and handles sensor noise robustly, Sect. 8.4.

8.2 Data Acquisition and Template Motion Estimation

The acquisition procedure, the employed template model and the marker-less motion estimation approach have been described in detail in [Theobalt05a] and Chapter 6. For details we would like to refer the reader to Chapter 6.

8.3 Enhanced BRDF Estimation

After performing marker-less motion capture for each frame of multi-view video, the position and orientation of each $u_{i,j}$ with respect to the calibrated acquisition apparatus is known. In other words, due to the scene motion it becomes possible to collect for each point on the surface a variety of reflectance samples, each representing the appearance of the point from known outgoing viewing and incoming lighting directions. The original method described in the previous part exploits this fact in order to estimate for each $u_{i,j}$ a static parametric BRDF model from the RES. An energy minimization framework was used to compute parameters of an isotropic Lafortune BRDF f_r at each surface point such that the measured data are best approximated [Lafortune97b]. For solving the dynamic geometry reconstruction problem addressed here, we replace the original least-squares approach by a regression framework based on robust Huber statistics [Huber04] as

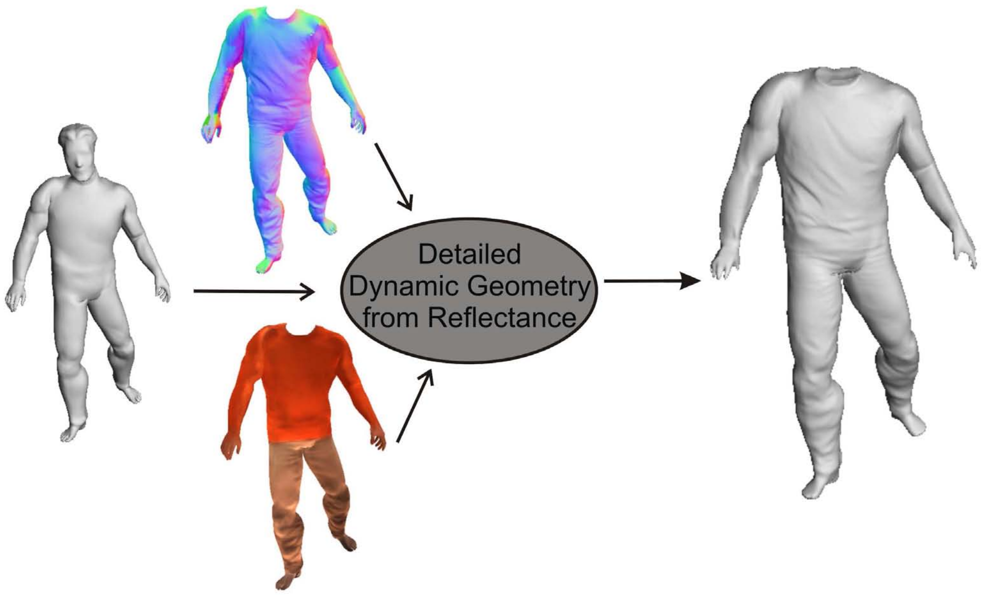


Figure 8.1: Overview: The tracked smooth template model (left), along with per-texel refined normal field (top) and per-texel BRDF parameters (bottom) are used to estimate detailed time-varying surface geometry (right).

this enables us to obtain more faithful estimates in the presence of non-Gaussian measurement noise. For each surface point $u_{i,j}$ on the template, we minimize the following energy functional to find an isotropic BRDF that reproduces the data in the RES:

$$\begin{aligned} E_{\text{BRDF}}(\rho(u_{i,j})) = \\ &\sum_{t}^{T}\sum_{c}^{8}\kappa_c(u_{i,j},t)\mathcal{H}\Bigg(S_c(u_{i,j},t) - \\ &[\sum_{e}^{2}\lambda_e(u_{i,j},t)(t)(f_r(\mathbf{l}(u_{i,j},t),\mathbf{v}_c(u_{i,j},t),\rho(u_{i,j})) \\ &\cdot I_e(\mathbf{n}_o(u_{i,j},t)\cdot\mathbf{l}(u_{i,j},t)))]\Bigg)^2 . \end{aligned} \tag{8.1}$$

E_{BRDF} is evaluated separately in the red, green and blue color channel. $S_c(u_{i,j},t)$ denotes the color of $u_{i,j}$ measured from camera c, and I_e denotes the intensity of light source e. The viewing directions $\mathbf{v}_c(u_{i,j},t)$ and light source directions $\mathbf{l}_e(u_{i,j},t)$ are expressed in $u_{i,j}$'s local coordinate frame based on the (template) surface normal $\mathbf{n}_o(u_{i,j},t)$. $\kappa_c(u_{i,j},t)$ and $\lambda_e(u_{i,j},t)$ encode the visibility of point

$u_{i,j}$ with respect to cameras and light sources, respectively. As opposed to the original least-squares minimization framework which assumes Gaussian noise in reflectance samples and thus may over-weight outliers, we employ robust Huber statistics $\mathcal{H}$ as penalizer. The Huber function $\mathcal{H}$ is defined as

$$\mathcal{H}(R) = \begin{cases} \frac{1}{2}R^2 & \text{, if } |R| \leq k \\ k|R| - \frac{1}{2}k^2 & \text{, if } |R| > k \end{cases} \tag{8.2}$$

where k is the clip threshold [Huber04]. $\frac{d\mathcal{H}}{d\mathcal{R}}$ is continuous and often referred to in the literature as the clip function. $\mathcal{H}$ preserves the advantageous convergence properties of an L_2 function for inliers, but resorts to an L_1 norm for samples that are likely to be outliers. By this means we implicitly model our noise characteristics more faithfully as a heavy-tail Gaussian. In order to find the clip threshold k for $\mathcal{H}$ we analyze the variance in captured reflectance samples in a series of consecutive video frames in which the person remains in a static pose relative to the cameras. For each color channel and each material we compute the average variance and use the squared values as material- and color-specific clip thresholds.

In practice BRDF parameters are estimated in a multi-step procedure. First, materials on the surface are clustered based on average diffuse color and a specular BRDF component is estimated for each material separately. Thereafter, a per-texel diffuse model is fit to each surface point after subtracting the previously estimated specular component from each sample. As our acquisition setup is the same as described in Sect. 6.2, comprised of eight cameras and two spot lights. Please note that we only use samples seen by exactly one light source for estimation, which, due to the positioning of lamps in the studio, in reality is true for over 90% of samples. For numerical minimization, we employ the L-BFGS-B minimizer [Byrd95].

Given estimates of the BRDF parameters, we can also refine our knowledge about surface geometry by keeping the reflectance parameters fixed and minimizing the same functional in the normal direction. Once the BRDF parameters have been recovered from the RES, a similar minimization procedure is used to reconstruct the time-varying normal field from each DSS.

8.4 Adding Spatio-Temporally Coherent Geometric Surface Detail

Dynamic normal fields encode information on high-frequency surface detail without physically deforming the smooth template surface over which they are parametrized. This information is sufficient to render relightable 3D videos of humans from many angles apart from grazing ones. However, true 3D time-varying geometric detail is essential in many production quality animation settings where realistic renditions from novel viewpoints and under arbitrary illumination are expected. Only true deformed surface geometry will enable correct appearance of the shape under the final lighting simulation.

In the following, we therefore present a new data fusion framework that transforms the original setup for relightable 3D video capture into a system for high-quality capture of detailed dynamic surface geometry. Our method is grounded on the assumption that our smooth template, essentially capturing low frequency geometry, is already well-aligned with the input.

Our algorithm estimates for each surface point $\mathbf{u}_{i,j}$ on the smooth template at each time step t a 3D displacement vector $\mathbf{d}(u_{i,j}, t)$ that yields the true 3D position of the point u at t as $\mathbf{x}_d(u_{i,j}, t) = \mathbf{x}(u_{i,j}, t) + \mathbf{d}(u_{i,j}, t)$. Since the true displacements are expected to be small, it is safe to assume that the displacement direction is always along the direction of template normals.

As our measurements are potentially contaminated by noise, we employ a statistical framework to robustly find the most likely field of surface displacements given the data. To achieve this purpose we model the joint posterior distribution of the field of displacements at each time step as a Markov Random Field (MRF) which takes the form

$$\begin{aligned}\mathbf{p}(\mathbf{d}(u_{i,j}t,) \, |\mathbf{n}_m(u_{i,j}, t), \mathcal{M}(t)) = \\ \frac{1}{Z} e^{-(\alpha\Phi(t)+\beta\Psi(t)+\gamma\Omega(t)+\delta\Xi(t))} \ , \end{aligned} \tag{8.3}$$

where Z is a normalization constant, $\Phi(t)$ models our measurement process, and $\Psi(t)$, $\Omega(t)$ and $\Xi(t)$ are prior potentials. α,β, γ and δ are weighting factors summing to 1. Empirically we found that values of $\alpha = 0.6$, $\beta = 0.1$, $\gamma = 0.2$ and $\delta = 0.1$ produce most decent results (see also Sect. 8.5 for a discussion). The spatio-temporal neighborhood structure of our MRF connects each surface location to the four immediate spatially adjacent ones at the same time step (easily found from our surface parametrization), as well as to its instantiations at the two previous time steps.

As we are interested in the most likely solution given the current data only and not in the full posterior, we find the most likely surface as the maximum a posteriori (MAP) hypothesis by minimizing the negative log-likelihood of (8.3) as

$$\hat{\mathbf{d}}(u_{i,j}, t) = \underset{\mathbf{d}(u_{i,j},t)}{\text{argmin}} \ \alpha\Phi(t) + \beta\Psi(t) + \gamma\Omega(t) + \delta\Xi(t) \ . \tag{8.4}$$

In the following subsections, we first describe and motivate how assumptions about noise characteristics are encoded in measurement potentials, Sect. 8.4.1, and illustrate what prior potentials are appropriate to properly condition our solution space, Sect. 8.4.2. Finally, we describe how to practically solve for a maximum a posteriori (MAP) surface even in our large scenes with on average 350,000 surface points, Sect. 8.4.3.

8.4.1 Measurement Potential

The information that captures the true shape of the fine-grained surface details is encoded in our measured surface normal field $\mathbf{n}_m(u,t)$. Our measurement potential therefore aims at minimizing the angular difference $\Delta(\mathbf{n}_m(u_i,t),\mathbf{n}_r(u_i,t))$ between the measured normals and the normals of the displaced surface.

To properly constrain our problem, we don't formulate the error in normal field approximation based on individual locations $u_{i,j}$ (i.e. individual texels in the texture domain), but rather based on triangles obtained by regularly triangulating all texels in the parametrization. Normals for the obtained triangles are computed by simply averaging the normals at their three respective vertices (i.e. texels). Again, we capitalize on the Huber function $\mathcal{H}$ to obtain more reliable estimates in the presence of noise. Our measurement potential thus takes the form

$$\Phi(t) = \sum_{D=(u_a,u_b,u_c)\in\mathbf{D}} \mathcal{H}(\Delta(\mathbf{n}_m(D,t),\mathbf{n}_r(D,t)))\ , \tag{8.5}$$

where $D = (u_a, u_b, u_c)$ is a triangle formed by adjacent texels (surface points) u_a, u_b, and u_c, and $\mathbf{D}$ is the set of all such triangles. $\mathbf{n}_r(D,t)$ is the normal field according to the current deformed surface evaluated at D, and $\mathbf{n}_m(D,t)$ is the corresponding measured normal field. The clip threshold k was chosen conservatively in such that deviations of new and measured normals by more than 90° are considered outliers.

8.4.2 Prior Potentials

We make the general assumption that dynamic surfaces in the real world are smooth in both space and time. In other words, spatially adjacent surface locations should exhibit similar displacements and the change in displacement for the same surface location over time should be in reasonable bounds as well. The spatial smoothness constraint penalizes local deviation from an oriented plane in a 4-neighborhood around each point and is encoded in the potential

$$\begin{aligned}\Psi(t) = \sum_i \sum_j &\mathcal{H}(\mathbf{x}_d(u_{i-1,j},t) - 2\mathbf{x}_d(u_{i,j},t) + \\ &\mathbf{x}_d(u_{i+1,j},t)) + \\ &\mathcal{H}(\mathbf{x}_d(u_{i-1,j},t) - 2\mathbf{x}_d(u_{i,j},t) + \\ &\mathbf{x}_d(u_{i+1,j},t))\ ,\end{aligned} \tag{8.6}$$

where $\mathbf{x}_d(u_{i-1,j},t)$, $\mathbf{x}_d(u_{i+1,j},t)$, $\mathbf{x}_d(u_{i,j-1},t)$, and $\mathbf{x}_d(u_{i,j+1},t)$ are displaced 3D positions of surface locations adjacent to $u_{i,j}$. The clip threshold k of $\mathcal{H}$ in this case is chosen such that differences in local surface normal orientation of more than 30° are considered outliers.

Temporal smoothness is enforced by the potential

$$\begin{aligned}\Xi(t) = \sum_i \sum_j &(\mathbf{d}(u_{i,j},t) - 2\mathbf{d}(u_{i,j},t-1) - \\ &\mathbf{d}(u_{i,j},t-2))^2\ .\end{aligned} \tag{8.7}$$

This term favors a smooth rate of change of displacements over time, or putting it differently, favors small ”acceleration” in displacement change over time.

Lastly, we make the a priori assumption that displaced surface locations should remain close to the original smooth template shape. The latter constraint is essential as it prevents our surface from drifting arbitrarily far away from the original template. Our second prior therefore takes the form

$$\Omega(t) = \sum_i \sum_j \mathbf{d}(u_{i,j},t)^2 \tag{8.8}$$

8.4.3 Practical Implementation

The test sequences employed by us feature parametrizations of the smooth template of size 1024×1024 pixels. On average this corresponds to $350,000$ surface locations for which a displacement needs to be found at each time step. Please note that we compute displacements at a much higher level of granularity than the vertex density of the original template which is typically only 40,000.

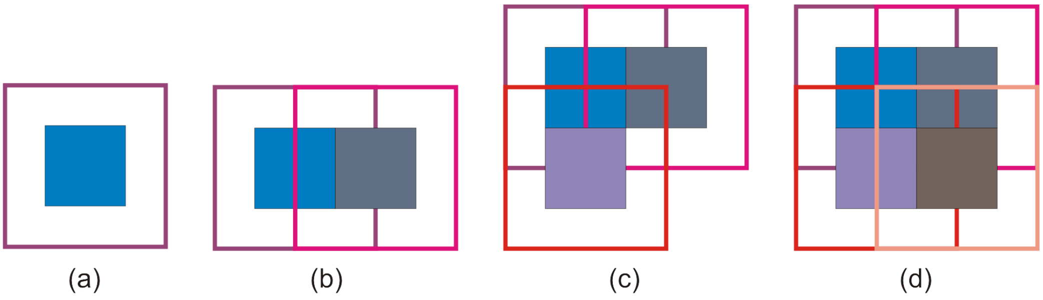


Figure 8.2: Patch-based optimization. A single patch, its boundary area, and its (blue) internal area (a). While the deformed surface is computed, the overlapping patches are processed in a sequence as shown in (b), (c) and (d) respectively. Only the interior patch area is preserved after displacement computation for one patch.

Parametrizations were obtained by manually cutting the template open and unfolding it over a 2D square by means of the conformal mapping technique described in [Zayer05b].

As we are only interested in a MAP solution to the final surface, we can conveniently resort to a standard off-the-shelf L-BFGS-B technique [Byrd95] to minimize Eq. 8.4.

To keep optimization tractable in the light of our very dense surface sampling, we also subdivide the overall surface reconstruction problem into a series of smaller ones. In practice, we subsequently compute displacements for individual surface patches and successively merge information from different patches to create the final result.

Each patch on our model corresponds to a square region of surface locations in our parametrization domain. Furthermore, each such square region is composed of an interior region and an exterior boundary area, Fig. 8.2. If we would simply deform individual adjacent patches we would with very high likelihood obtain discontinuities at patch boundaries since the mutual MRF dependencies across the rim are not properly considered. To prevent this source of error, we arrange subsequent patches in an interleaving, half-overlapping pattern, see Fig. 8.2b,c,d for the temporal sequence in which the patches are processed. Furthermore, after the displacements for one complete patch were estimated, we only preserve the displacement at the center of the patch. The boundary regions are thus only employed to initialize the optimization of any subsequent patch whose center region overlaps with the boundary. All patches are considered equal, thus the choice of the starting patch for our optimization is arbitrary. Overall, this interleaved optimization pattern produces a high quality surface estimate that preserves detail while preventing erroneous discontinuities along boundaries, see Sect. 8.5 for further discussion.

8.5 Results and Validation

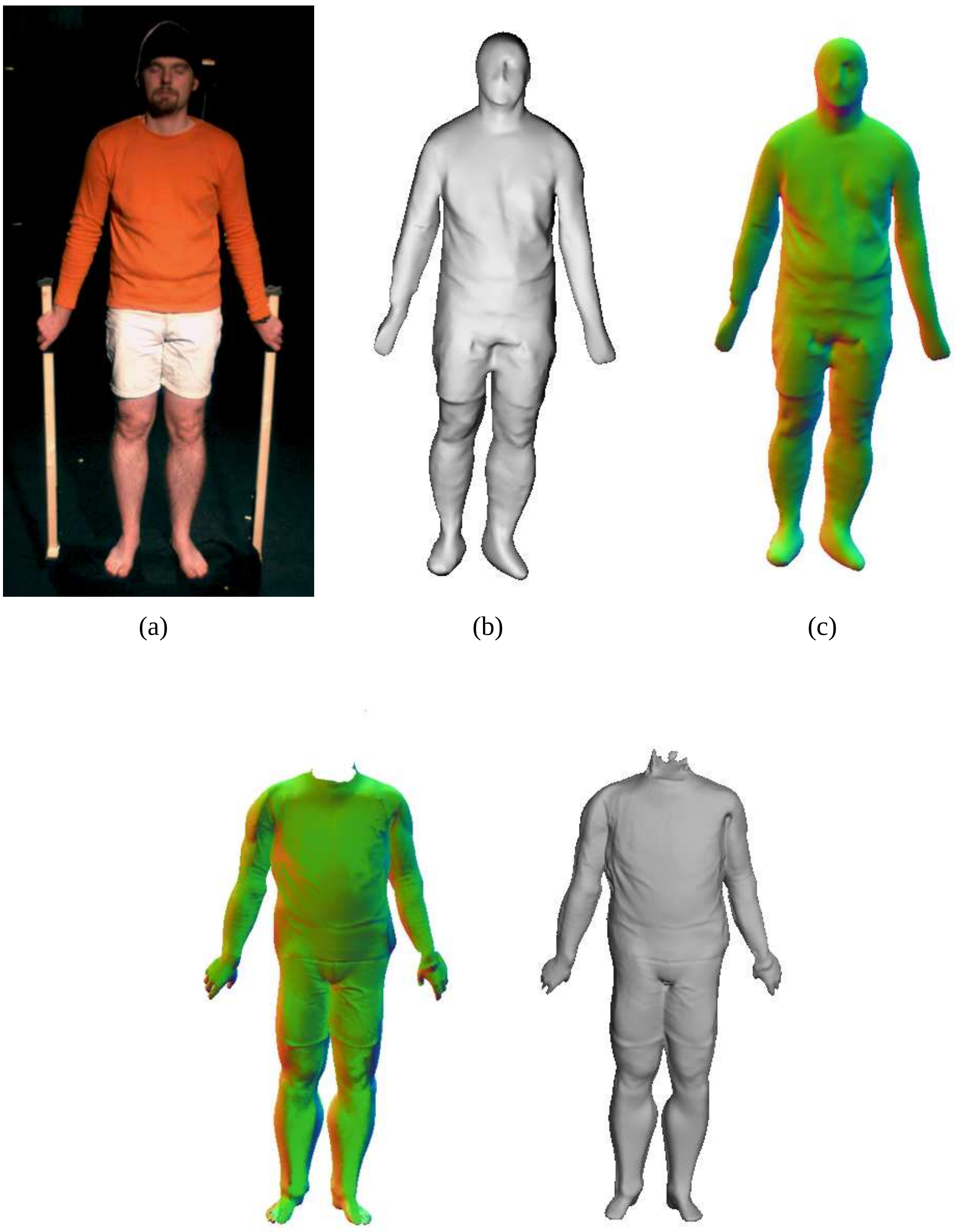
To demonstrate the results of our method, we have used two captured real-world motion sequences. The data for each sequence comprises of the moving low-detail template, all input image data (also in texture format already), full calibration data (cameras and lights), parametrization and warp-corrected texture coordinates. The latter is a set of data which encodes information on cloth shifting over the body's surface which was detected by the method detailed in Chapter 6.

The first sequence shows a scene in which the actor wears mostly diffuse clothing and walks back and forth in front of the cameras, Fig 8.5a,b. The RES (used for BRDF estimation) is 30 frames long and the DSS (used for geometry capture) comprises 184 frames. In the second sequence, the test subject wears a diffuse t-shirt and slightly specular trousers, and performs a basic taichi motion, Fig. 7.1 and 8.5c. While the RES contains again 30 frames, the actual motion in the DSS is 110 frames long.

As can be seen in Fig. 8.5 and Fig. 7.1, and also in the accompanying video [C08a], our reconstructed actor model faithfully captures even subtle detail, in particular wrinkles in clothing and folds, as *true* geometry. Fig. 8.4 zooms in on certain areas of the body model to illustrate that our MRF-based fusion method allows for reconstruction of subtle folds whose width is in the range of a few millimeters. This is a major improvement in shape quality over the original smooth template which was lacking any such detail, Fig. 7.1b and Fig. 8.4a,e. We would also like to point out that our final result is not only very detailed and almost free of artifacts at individual time steps, but due to the spatio-temporal MRF framework also faithful and smooth over time, see video [C08a]. The latter shows the unprecedented ability of our method to generate spatio-temporally smooth and detailed results even in the presence of measurement noise.

Although our visual results show qualitatively that we can measure highly-accurate scene geometry at sub-triangulation resolution, we also want to provide a more elaborate validation. Unfortunately, there exists no other scanning technology that would provide us with ground truth dynamic geometry at the same level of detail.

We therefore resort to data that was employed by Theobalt et. al [Theobalt07] for


(a) (b) (c)

(d) (e)

Figure 8.3: In this test an RES was recorded with a person standing in a static pose on a rotating turntable (a). Also, a scan was performed with a structured light laser scanner in order to obtain an as good as possible ground truth shape (b). (c) shows the normal field of the scan, where (global) normal direction is encoded in RGB color. (d) and (e) show the result that is obtained if we start from the smooth template fitted to the pose of the test subject, perform photometric stereo and run our MRF-based method to obtain the final detailed geometry. As one can see, our result captured some of the very fine wrinkles on the body much more faithfully than even the laser scanner.

the validation of the original surface reflectance estimation method. This data set contains an RES in which the actor strikes a static pose on a rotating turntable. In addition to the recording of the RES, a laser scan of the person was taken during preprocessing. Since we were also given the pose of the template at each frame, we were able to reconstruct the BRDF and normal map based on our method, and could use our MRF framework to generate detailed surface shape. Since the scan and template possess different triangulations direct vertex comparison is infeasible. However, visual comparison of our result Fig. 8.3e and the scanned ground truth Fig. 8.3b shows that all detail present in the original scan is also present in the deformed template, and that the resolution at which geometry was recovered is even higher in our result.

Typically, we reconstruct as many as 350,000 displacement values over the template surface. Even at this detail level and when using a small patch size of 16 pixels, it takes moderate 5 to 6 minutes per time step of video to find the final detailed surface. This time is in addition to the timings of acquisition, BRDF and normal estimation. Optimal values for the parameters α, β, γ and δ were found experimentally. To this end, we used a sequence of 3 of the reconstructed detailed meshes of the sequence 1 as a ground truth and used their normal fields as measured normal fields. reconstruction errors could now be measured for a reasonable sample of combinations of the coefficients. Optimal results are obtained for $\alpha = 0.6$, $\beta = 0.1$, $\gamma = 0.2$ and $\delta = 0.1$ which were used in all our experiments.

Our method is subject to a couple of limitations. An important assumption enabling us to properly localize our final geometric solution in space is the one that the template is close to the true geometry. Unfortunately, this assumption is not entirely true for the head of the template as there may be quite some differences to true hair style and face geometry. Simple free-form deformation performed for the shape capture cannot compensate for this. Therefore, we exclude the head from our reconstructions and note that this is a problem attributed to the provided input data.

Secondly, the currently employed template limits the types of scenes that we can handle to people wearing not too wide apparel. However, this is not a general limitation of our own contribution as we can easily apply our method to coarse geometry reconstructed with any other approach as well, as long as the geometry (triangulation) is coherent over time.

The original taichi input sequence also shows some jitter in the pose of the smooth template (slightly noticeable in the video result [C08a]), possibly due to tracking inaccuracies. We did not take any measures to compensate for this.

Finally, in any frame where a surface point is in shadow from the light source, no normal direction can be reconstructed and the template normal is used instead. In the video [C08a], this effect is sometimes noticeable when the arm casts a shadow on the torso. However, our method handles this situation gracefully and produces the best possible result given this hard-to-prevent occasional lack of data in general moving scenes.

Despite these limitations we have presented one of the first approaches to reconstruct high-quality and high detail geometry of large dynamic scenes in a purely passive way.

8.6 Conclusion

In this chapter we presented one of the first passive methods to reconstruct geometry of large dynamic scenes showing moving actors at very high detail and accuracy from video only. One of the strength of our work is that we only need multi-view video data recorded under calibrated cameras and lighting. This allows not only the highly detailed reconstruction but also the video data can be directly used for video-based rendering or relighting. In contrast, active methods solely concentrate on the reconstruction, and normally project a pattern over the scene, which renders the video data useless for most of the other video-based modeling tasks.

In this work, as a first step we built on our earlier work that allows for capturing of coarse geometry, surface reflectance and dynamic normal maps. We then applied a new MRF-based spatio-temporal surface deformation approach that converts the geometric details encoded in the normals into true 3D displacements over the smooth template. Our method faithfully handles typical heavy-tail measurement noise, and is one of the first to allow for spatially accurate and temporally consistent height reconstruction over curved dynamic base geometry.

Our work is not limited to any particular representation of the model. In the next part of this thesis, we will present a passive method to reconstruct spatio-temporally coherent arbitrary scene geometry from multi-view video data. The reconstructed spatio-temporal mesh animation of any subject can be used for surface reflectance, dynamic normal field estimation, and subsequently for reconstructing high quality time-varying details.

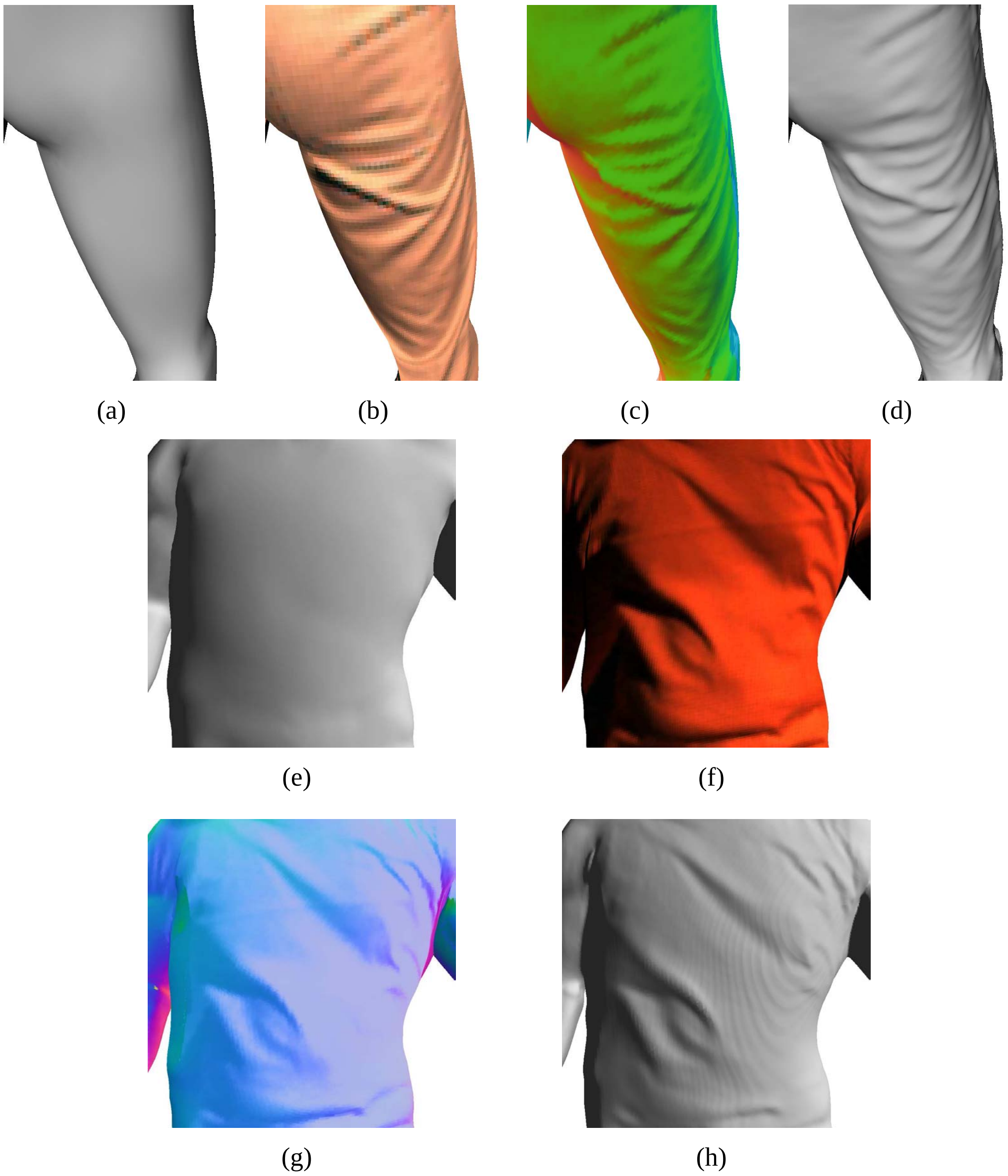
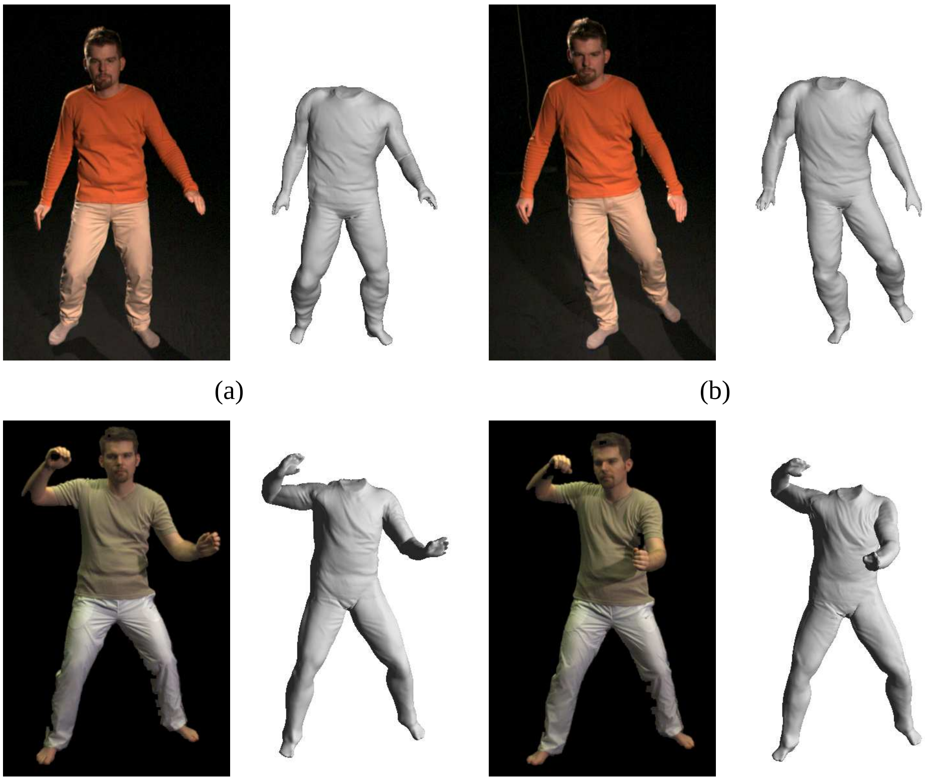


Figure 8.4: Our method can capture even subtle folds and wrinkles whose size is in the range of a few millimeters only. Zoom-in on leg: (a) smooth template, (b) template with texture, (c) color-coded normal field, (d) our final result rendered in OpenGL using Gouraud shading. (e)-(h) show a similar zoom onto the torso of the subject in the walking sequence. Also here, surface details were faithfully recovered in geometry.



(a) (b)

(c) (d)

Figure 8.5: Each pair of images shows, side-by-side, one original input video frame and the full 3D surface model with all geometric detail rendered in OpenGL from the same perspective. Sample input video frames of the motion sequences along with the corresponding detailed geometry. The direct comparison shows that our method captures even subtle dynamic geometric details in the actor's clothing very accurately.

Part V

Spatio-Temporally Coherent Dynamic Scene Reconstruction Without A Prior Shape Model

Chapter 9

Problem Statement

This part proposes a solution to the problem of reconstructing a structured mesh sequence from synchronized multi-view video streams. After reviewing the related work, a method for establishing dense correspondences between unrelated meshes is presented, which is used to obtain a spatio-temporal coherent mesh sequence.

In the previous chapters of this thesis, we focused on reconstructing realistic human animations from multi-view video data. All of the presented methods relied on a template human body model that was deformed and animated to capture the true shape and motion of the human actor. Although we have demonstrated that our animation reconstruction methods result in high quality animations, they were nevertheless limited by the model description of the template. It can only handle the specific scene for which the template model is available and can not handle any arbitrary scene. Ideally, instead of using a template model, one would like to reconstruct the time-varying shape and appearance of the arbitrary real-world scene performers from multi-view video data directly without having to craft a scene model beforehand. To some extent there are already quite a few methods that can achieve this goal, Sect. 9.1.

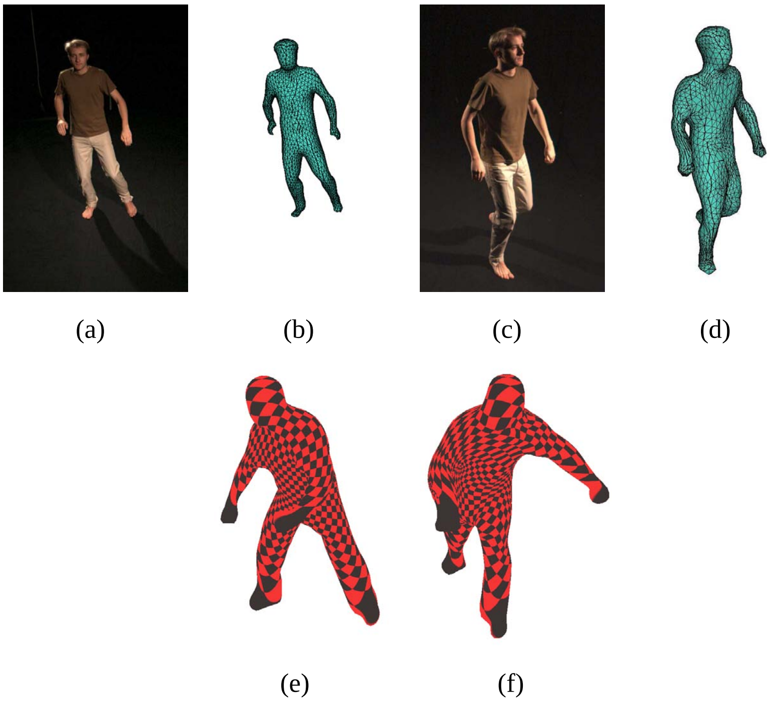
Most of these methods provide convincing shape and appearance for each time step of an input animation individually. However, they fall short of reconstructing spatio-temporally coherent scene geometry for arbitrary subjects since the challenging 3D correspondence problem is not addressed. Spatio-temporal coherence is an important and highly-desirable property in captured animations, as it greatly facilitates or even is inevitable for many tasks such as editing, compression or spatio-temporal postprocessing.

We therefore propose a new spatio-temporal dense 3D correspondence finding method that enables us to capture coherent dynamic scene geometry using standard shape-from-silhouette methods [Ahmed08b]. Our algorithm is tailored to the characteristics of video-based reconstruction methods which often capture high spatial detail in the input video frames, but provide relatively sparsely sampled 3D geometry with a much lower level of shape detail and with a considerable level of noise.

In a first step, shape-from-silhouette surfaces are reconstructed for each time step of video yielding a sequence of shapes made of triangle meshes with varying connectivity. Thereafter, sparse 3D correspondences between subsequent pairs of surfaces are computed by matching 3D positions of optical features that can be accurately extracted from high-resolution input video frames, Sect. 10.3. These sparse correspondences represent control points for anchoring appropriate bivariate scalar functions on each reconstructed surface mesh, Sect. 10.4. The choice of these functions enables us to establish dense correspondence essentially by matching function values. The dense correspondences can be used to straightforwardly align one mesh to all other reconstructions by performing a sequence of pairwise registrations, Sect. 10.6. The output of our approach is a spatio-temporally coherent animation, i.e. a sequence of meshes with constant graph structure and low tangential distortion. Main contributions, advantages and novelties of our algorithm are the following

- As an object space method it does not suffer from parametrization-induced limitations.
- It establishes dense correspondence fields independently of the level and structure of surface discretization which makes surface alignment straightforward.
- It explicitly addresses the characteristics of shape-from-silhouette-based animation reconstruction. By combining both accurate image feature and function matching, we are able to robustly match even coarsely reconstructed surface geometry lacking coherent and dense surface details.
- In practice, robustness to topology changes.

In the following section we will review the related work in domains of surface reconstruction, mesh animation and correspondence estimation between the meshes.


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

(e) (f)

Figure 9.1: Input video frames (a), (c) and corresponding spatio-temporally coherent meshes rendered back into same camera view (b), (d). The checkerboard texture shows the consistently small tangential surface distortion in our reconstruction even between temporally far apart frames (e), (f).

9.1 Related Work

Technological progress in recent years has made it feasible to reconstruct shape and appearance of dynamic scenes using video [Matsuyama04] or video plus active sensing [Waschbüsch07]. Multi-view video methods based on the shape-from-silhouette [Matusik01] or stereo principle [Zitnick04] bear the intriguing advantage that they enable reconstruction of arbitrary moving subjects. Unfortunately, none of these methods is designed to reconstruct scene geometry with coherent connectivity over time since the 3D correspondence problem is not addressed. Model-based approaches employ shape priors [de Aguiar07b, Cheung03, Hernández07] which limits them to certain types of scenes. The algorithm proposed in this part enables coherent dynamic shape reconstruction while maintaining the flexibility of shape-from-silhouette methods.

In geometry processing, the 3D correspondence problem is addressed in parametrization and its application in (compatible) remeshing see, e.g., the surveys [Hormann07, Alliez07] where the goal is to match the connectivity of one *single* shape model to the connectivity of another one. Generally, the required robust parametrization techniques are limited to fixed topology and are computationally involved, especially in the presence of additional constraints from given correspondences.

The key to spatio-temporally coherent reconstruction is a robust solution to the 3D correspondence problem. Conceptually similar to this problem, albeit in a reduced problem domain, is the shape matching problem [Rusinkiewicz05]. One way to solve this problem is to localize and match salient geometric features between two shapes [Gal06]. By combining feature matching with pose transformation, two shapes can be aligned [Huber03]. Some probabilistic alignment methods register laser scans by finding the most probable embedding of one shape into the other [Anguelov04]. Iterative closest point (ICP) procedures use a much simpler correspondence criterion that iteratively pairs locations closest to each other [Hähnel03]. ICP methods may easily get stuck in local minima if no decent initial registration is provided. None of the aforementioned algorithms explicitly addresses the problem of multi-frame animation reconstruction.

Only few methods so far explicitly address the problem of reconstructing coherent animated surfaces from real-time scanner data, such as real-time structured light scanners [Wand07, Stoll06]. Unfortunately, in a video-based setting like ours, the applicability of these methods is either limited by high computational complexity, or by the requirement of high spatial and temporal sampling density which is typically not fulfilled.

Similar to our approach is the algorithm proposed by Shinya et al. [Shinya04] who deform a 3D model into sequences of visual hull meshes by minimizing a deformation energy. In contrast to our algorithm, accurate optical feature information is not exploited, and the ICP-like correspondence criterion is vulnerable to erroneous local convergence.

Matsuyama et al. [Matsuyama04] suggest a method to deform a mesh based on multi-view silhouettes and multi-view photo-consistencies. By optical means only, the required dense matches are difficult to find, and therefore the strongly constrained non-linear minimization takes several minutes computation time per frame. In contrast, our algorithm is computationally more efficient and creates dense correspondences despite only sparse optical matches.

Starck et al. [Starck05] also aim at establishing coherence in sequences of shape-from-silhouette meshes. Their method establishes correspondences in a spherical parametrization domain which may fail in extreme poses and may introduce distortion-dependent matching inaccuracies close to singular points. In a recent follow-up, Starck et al. [Starck07a] apply a Markov random field to match isometry-invariant surface descriptors based on local parametrization. This enables establishing correspondence over wide time-frames, which is in fact a different problem. For both, [Starck05, Starck07a], numerical problems are more involved and computational costs are orders of magnitude higher than for our method.

Chapter 10

Spatio-Temporally Coherent Dynamic Surface Reconstruction Using Dense Correspondence

This chapter describes a dense correspondence finding method that enables spatio-temporally coherent reconstruction of surface animations from multi-view video. First, a method to establish sparse correspondences between the two surfaces is presented. Using sparse correspondences as the anchor points, dense correspondence is established between the surfaces. This dense correspondence is propagated from the start to the end of the sequence, in order to obtain a spatio-temporally coherent sequence.

10.1 Overview

The input to our method is a sequence of calibrated synchronized video streams that were recorded from multiple viewpoints around the scene and that show a subject performing in the scene's foreground. Our test acquisition system features eight synchronized video cameras arranged in a circular setup and delivering 25fps at 1004x1004 pixel frame resolution (Chapter 3).

Background subtraction yields a foreground silhouette for each of the N captured video frames. In a pre-processing step a polyhedral visual hull method [Franco03] is applied to each time-step of video. In order to cure triangle degeneracies in the input data and to produce a more uniform surface discretization, the visual hull surfaces are resampled and the resulting point clouds are fed into a Poisson surface reconstruction approach [Kazhdan06] (we use their implementation). This way, a sequence of triangle meshes with varying vertex connectivity is produced that captures the shape of the subject at each time step.

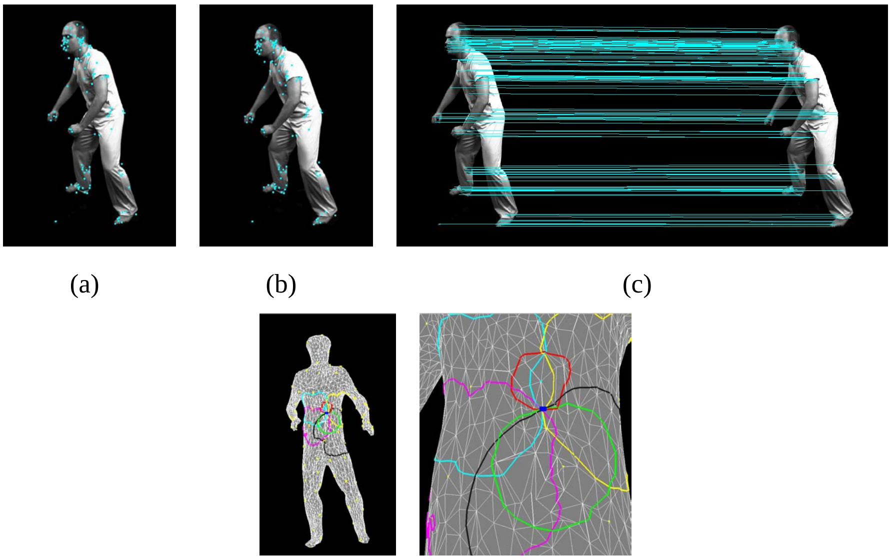
10.2 Spatio-Temporal Correspondence Finding

In the following we describe a triangle mesh as $\mathcal{M} = (\mathcal{V}, \mathcal{T}, \mathbf{p})$, where $\mathcal{V}$ denotes vertices and $\mathcal{T}$ their triangulation or *connectivity*. Hence, $(i, j, k) \in \mathcal{T}$ denotes a triangle, and with each vertex $\ell \in \mathcal{V}$ we associate positions $\mathbf{p}_\ell \in \mathbb{R}^3$ defining the surface's embedding in 3D. We consider N time-frames and thus write a sequence of meshes as $\mathcal{M}(t) = (\mathcal{V}(t), \mathcal{T}(t), \mathbf{p}(t)), t = 0, \ldots, N-1$, where $\mathcal{M}(t)$ approximates the (ideal) surface $\mathcal{S}(t)$.

Our algorithm propagates the connectivity of mesh $\mathcal{M}(0)$ by iteratively matching it against reconstructed visual hull meshes. In the following, we write $\mathcal{M}_0(t)$ for meshes with connectivity $(\mathcal{V}_0, \mathcal{T}_0) := (\mathcal{V}(0), \mathcal{T}(0))$ of $\mathcal{M}(0)$, i.e., $\mathcal{M}_0(t) = (\mathcal{T}(0), \mathcal{V}(0), \mathbf{p}(t))$ and in particular $\mathcal{M}(0) = \mathcal{M}_0(0)$. Then given a subsequent pair of meshes $\mathcal{M}_0(t)$ and $\mathcal{M}(t+1)$, where $\mathcal{M}_0(t)$ is $\mathcal{M}(0)$ aligned with $\mathcal{M}(t)$ during a previous iteration, our algorithm proceeds as follows:

In a first step, initial coarse correspondences are obtained by matching robust optical features between image-frames and mapping them to 3D-positions on the surfaces, Sect. 10.3. We use SIFT [Lowe99] for this purpose, yielding a sparse covering of the surfaces with feature points. In contrast to deformation transfer methods [Sumner04, Zayer05a], we can't choose ideal features, i.e. our sparse features alone generally don't carry enough information for direct correspondence or deformation-based alignment, see also Sect. 10.8.

Therefore, we estimate dense correspondences in a second step, which constitutes the core of our approach: with each feature point we associate a scalar, monotonic function with certain interpolation properties. Requirements for such functions will be discussed in detail in Sect. 10.4. Dense correspondences are found by pairing surface locations with similar function values.



(a) (b) (c)

(d) (e)

Figure 10.1: Detected SIFT features in two consecutive frames (a) and (b). Matched features are shown in (c). Obvious outliers, such as matches outside the silhouette, are filtered out during preprocessing. Intersecting iso-contours of harmonic functions centered on sparse correspondences (shown as colored lines) can be used to localize surface points. For clarity, (e) zooms in on a subregion of (d).

This way we can provide surface correspondences which are densely and faithfully distributed over the surface. We use these matching 3D surface points as constraints for deforming one mesh over time without resorting to involved deformation algorithms (see, e.g., [Botsch07]) that were necessary if correspondences were sparse. The result is an animation sequence with constant connectivity.

We remark that the approach is tailored to the particular animation setting: the acquisition and shape-from-silhouette reconstruction provides only moderately accurate and medium resolution geometry data, possibly contaminated with noise, but at the same time high-resolution texture information per image frame. The individual matching steps are detailed in the following subsections.

10.3 Coarse Correspondences

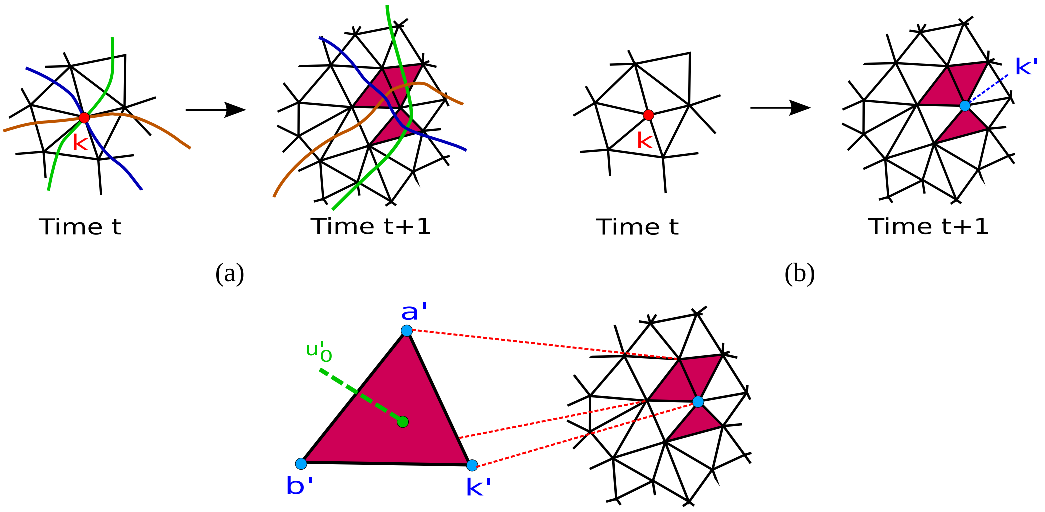
In order to establish coarse correspondences we find robust optical features between adjacent frames by localizing them in the input video frames and inferring their 3D positions by means of the available reconstructed model geometry. For localizing features we apply SIFT descriptors [Lowe99] as this technique has a number of advantageous properties for our video setting: identified features are largely invariant under rotation, scale and moderate change in viewpoint, and the rich descriptors also enable wide-baseline matching. In particular the latter property pays off in our setting as rapid scene motion may easily lead to large image disparities between subsequent frames. In such a scenario, alternative image matching approaches, such as KLT or general optical flow methods are more likely to fail [Barron92]. Also, as opposed to geometric feature matching [Gal06] we can maintain precision even if the reconstructions don't exhibit salient shape details.

We compute 2D SIFT feature locations for each input frame $I_c(t)$ at all time steps t and all camera views c in a preprocessing step. On a typical sequence we obtain between 300 and 500 features per time step (with multiple occurrences of the same feature across cameras discarded.

When aligning two subsequent meshes $\mathcal{M}_0(t)$ and $\mathcal{M}(t+1)$, we compute 3D feature positions at either time step by back-projection from images onto the 3D shapes. To preserve the highest possible feature localization accuracy independently of triangulation (from Marching Cubes after Poisson reconstruction), 3D positions of features are computed from linear interpolation rather than nearest vertex positions. To this end, we exploit the graphics hardware and assign to each feature an interpolated 3D position obtained via rasterizing the 3D shape's coordinates into the same camera view.

To facilitate later computation of dense correspondences, we intermediately enforce association of features with vertices by locally splitting each original triangle containing a feature into three triangles. This is achieved by inserting a new vertex at the interpolation point. By performing 3D localization and subdivision for all camera views at a each time step t and $t+1$, we create a set of possibly subdivided versions of the original reconstruction meshes $\mathcal{M}'_0(t)$ and $\mathcal{M}'(t+1)$. Each of these meshes possesses an associated set of feature vertex indices $\mathcal{F}(t)$ and $\mathcal{F}(t+1)$. Note that these meshes only serve as temporary helper structures to gain accuracy. Local splits will be rolled back later, and are neither used in the final output of our method nor induce any other side effects, see Sect. 10.6. Therefore, and to keep notation simple, we will continue to refer to $\mathcal{M}_0$ and $\mathcal{M}$.

We find correspondences between SIFT feature vertices on either mesh by looking for pairs with similar descriptors. To this end, we compute the Euclidean distance $D_e(i,j)$ between the descriptors of all elements $i \in \mathcal{F}(t)$ and $j \in \mathcal{F}(t+1)$. A correspondence (i,j) is considered plausible and hence established if $D_e(i,j)$ is below a certain threshold. This way, possible outliers in all correspondence sets



(a)

(b)

(c)

Figure 10.2: (a) Vertex k (corresponding to u_0) and the iso-contours intersecting at it. For better visibility only $K = 3$ contours are shown. At time $t + 1$, the same iso-contours don't intersect in a single point. Each candidate triangle (shown in red) is intersected by two of the iso-contours. (b) A vertex k' from the candidate triangle set on $\mathcal{M}(t + 1)$ that is closest to k according to D_h criterion is selected. (c) Finding the surface point u'_0 within the best-matching triangle (a', b', k') (according to D_h) that is adjacent to k'.

are filtered out by discarding matches with implausible 3D distances. Erroneous matches outside the silhouette area are trivially discarded. Fig. 10.1(a-c) illustrates SIFT features.

10.4 Finding Dense Correspondences

The basic idea for establishing dense correspondence is to infer additional values from the given sparse features and the surface, and to then carefully analyze and compare these values over time. For this purpose we define bivariate scalar functions h_i on the surfaces, each function is associated with a particular feature $f_i \in \mathcal{F}$, $i = 0, \ldots, m$. In an ideal setting we could think of these as distance or coordinate functions: given three (feature) points a, b, c in the plane, any point in the plane can be characterized by its distance to each of a, b, c or in terms of its barycentric coordinates w.r.t. the triangle (a, b, c). Our choice of functions h_i resembles barycentric coordinates as we require *interpolation* $h_i(\mathbf{u}_i) = 1$ and

$h_i(\mathbf{u}_j) = 0$ for all $i \neq j$, and *monotonicity* of h_i with extrema at the interpolation points, where $\mathbf{u}_i \in \mathbb{R}^2$ denotes a surface point associated with f_i.

In order to be meaningful when evaluated for different t over the time-dependent surface $\mathcal{S}(t)$, we additionally require that h_i is taken from a class of functions which change their values only slightly under moderate surface deformations. For this reason we chose harmonic functions which satisfy

$$\Delta_{\mathcal{S}(t)}\, h_i = 0 \ , \tag{10.1}$$

where $\Delta_{\mathcal{S}(t)}$ denotes the Laplace-Beltrami operator. This is justified by the isometry-invariance of the operator, i.e., for isometric deformations of $\mathcal{S}$ into $\mathcal{S}'$ we have $\Delta_{\mathcal{S}} = \Delta_{\mathcal{S}'}$. We assume moderate deformations of $\mathcal{S}(t)$ to be largely isometric. This property has previously been exploited to compute signatures for shape matching and retrieval, see, e.g., [Elad03, Reuter06].

So far we assumed continuous functions. In practice, h_i are piecewise linear functions w.r.t. $\mathcal{M}(t)$, and an appropriate discretization of the differential operator $\Delta_{\mathcal{S}(t)}$ is required. In particular, we require independence of the triangulation, i.e. for different meshes approximating the same shape, the discrete solutions of (10.1) should yield the same or very similar results. We use the well-established cotangent discretization which provides this linear-precision property and is symmetric (see [Wardetzky07] for a comparison of alternative discretizations).

With functions h_i computed we proceed in several steps to find dense correspondence. Given a surface point $\mathbf{u}_0 \in \mathcal{S}(t)$ that corresponds to a vertex k of $\mathcal{M}_0(t)$, the goal is to find a matching point $\mathbf{u}'_0 \in \mathcal{S}(t+1)$ using h_i defined on the mesh $\mathcal{M}_0(t)$ and h'_i defined on $\mathcal{M}(t+1)$. Evaluation of the harmonic functions yields "coordinates" $\mathbf{h}(\mathbf{u}) := [h_0(\mathbf{u}), \ldots, h_m(\mathbf{u})]$ and $\mathbf{h}'(\mathbf{u}) := [h'_0(\mathbf{u}), \ldots, h'_m(\mathbf{u})]$ for both surfaces. As contributions of $\mathbf{h}$ are localized we restrict ourselves to the K coordinate values of largest magnitude at $\mathbf{u}_0$, i.e., we consider $\mathbf{h}_{\mathbb{K}}(\mathbf{u}_0) := [h_{i_1}, \ldots, h_{i_K}]$, $i_1, \ldots, i_K \in \mathbb{K}$, where $h_\ell(\mathbf{u}_0) \geq h_{\overline{\ell}}(\mathbf{u}_0)$ for all $\ell \in \mathbb{K}, \overline{\ell} \notin \mathbb{K}$. In our implementation we use $K = 10$. We can visualize the local influence of the h_i geometrically by the analog of a planar Voronoi diagram thinking of $1 - h_i$ as distance function. Then for each element in a "Voronoi cell", we expect significant or meaningful contribution only from functions associated with the cell and its immediate neighbor cells. We therefore chose K conservatively, as on average one will find 6 immediate neighbors. In an ideal setting, $\mathbf{h}(\mathbf{u}) = \mathbf{h}'(\mathbf{u})$, and retrieving $\mathbf{u}'$ can be imagined as intersecting iso-contours $h'_i(\cdot) = h_i(\mathbf{u}_0)$, $i \in \mathbb{K}$. Fig. 10.1(d),(e) illustrates this concept by visualizing several iso-contours on the surface of a visual hull mesh intersecting in a single vertex. In the presence of moderate deformations and given discrete meshes, the equality generally does not hold. Therefore, instead of exact intersections, we are interested in a set of *trian-*

gles $\mathcal{E} \subset \mathcal{T}(t{+}1)$, which are intersected by at least one of the iso-contours passing through $\mathbf{u}_0$. These are triangles in which $\mathbf{u}_0'$ potentially resides. To put this idea into practice, we add to $\mathcal{E}$ all those triangles that are intersected by the highest number of contours with iso-value $h_i(\mathbf{u}_0)$. This yields a (potentially) 1-to-many match from $\mathbf{u}_0$ to a set of candidate triangles, see Fig. 10.2(a). To handle possible localization inaccuracies, in practice we build $\mathcal{E}$ conservatively and also include all candidate triangles for the vertices in a 1-ring around $\mathbf{u}_0$ which are identified by the same procedure.

To determine the final position of $\mathbf{u}_0'$ on $\mathcal{M}(t+1)$, we first identify the vertex $k' \in \mathcal{V}_{t+1}$ that is closest to $\mathbf{u}_0'$. We extract this vertex k' from the set $\mathcal{E}$ by computing a distance measure between $\mathbf{h}_{\mathbb{K}}(\mathbf{u}_0)$ and $\mathbf{h}_{\mathbb{K}}'(\mathbf{u}_\ell')$ for all vertices ℓ out of $\mathcal{E}$, see Fig. 10.2(b) for illustration on a simplified setting. (Note that the set $\mathbb{K}$ is determined w.r.t. $\mathbf{h}$ on $\mathcal{M}_0$.)

Through experiments we found the following measure to work very satisfactorily. Let $\mathbf{d}_{\mathbb{K}} := \mathbf{h}_{\mathbb{K}}(\mathbf{u}_0) - \mathbf{h}_{\mathbb{K}}'(\mathbf{u}_\ell')$. We define the distance $D_h(\mathbf{u}_0, \mathbf{u}_\ell')$ as

$$D_h(\mathbf{u}_0, \mathbf{u}_\ell') \;=\; \mathbf{d}_{\mathbb{K}} \;\left(\mathbf{I} - \mathrm{diag}(\mathbf{h}_{\mathbb{K}}'(\mathbf{u}_\ell'))\right)^3 \;\mathbf{d}_{\mathbb{K}}^\top$$

Let $\mathcal{E}_{\mathcal{V}}$ contain all vertices shared by triangles in $\mathcal{E}$. We select that vertex $k' \in \mathcal{E}_{\mathcal{V}}$ with minimal distance, i.e. $D_h(\mathbf{u}_0, \mathbf{u}_{k'}') \leq D_h(\mathbf{u}_0, \mathbf{u}_\ell')$ for all $\ell \neq k', \ell \in \mathcal{E}_{\mathcal{V}}$.

The final step in finding $\mathbf{u}_0'$ is to localize its position at sub-discretization accuracy since, in general, $\mathbf{u}_0'$ is an arbitrary surface point and won't coincide with a vertex location. To achieve this purpose, we first identify the triangle (a', b', k') in the 1-ring of k' for which the average of $D_h(\mathbf{u}_0, \mathbf{w})$ (with $\mathbf{w} \in \{\mathbf{u}_{a'}, \mathbf{u}_{b'}, \mathbf{u}_{k'}\}$) is minimal. The best-matching surface point is expressed linearly as $\mathbf{u}_0' = \lambda_{a'}\,\mathbf{u}_{a'} + \lambda_{b'}\,\mathbf{u}_{b'} + \lambda_{k'}\,\mathbf{u}_{k'}$. We determine $\mathbf{u}_0'$ within (a', b', k) as

$$\underset{\lambda_{a'},\lambda_{b'},\lambda_{k'}}{\arg\min} \, ||\mathbf{d}_{\mathbb{J}}||^2 \;,$$

where $\mathbf{d}_{\mathbb{J}} := \mathbf{h}_{\mathbb{J}}(\mathbf{u}_0) - \mathbf{h}_{\mathbb{J}}'(\mathbf{u}_0')$ and $\mathbb{J} \subset \mathbb{K}$ contains the indices of the three largest coordinate values at $\mathbf{u}_0$. Intuitively, we thereby place $\mathbf{u}_0'$ as close as possible to either of the three highest-value iso-contours within the area of (a', b', k'), ideally at their intersection point. Fig. 10.2(c) illustrates this last step.

10.5 Remarks on Practical Implementation

10.5.1 Computation of Coordinate Functions

Numerically, h_i can be computed for every $\mathcal{M}(t)$ very efficiently by factoring a sparse matrix and then applying $m + 1$ back-substitutions. As a result we obtain $m+1$ linear functions h_i , i.e., for every vertex $j \in \mathcal{V}$ we have $h_i(\mathbf{u}_j)$. In practice, we compress this data efficiently by storing only the K largest values together with associated feature indices $\mathbb{I}_j = \{i_1, \ldots, i_K\} \subset \mathcal{F}$. Hence, for every vertex j we store $h_\ell(\mathbf{u}_j)$, $\ell \in \mathbb{I}_j$, where $h_\ell(\mathbf{u}_j) \geq h_{\overline{\ell}}(\mathbf{u}_j)$, $\overline{\ell} \notin \mathbb{I}_j$. Consequently, we implicitly assume $h_{\overline{\ell}}(\mathbf{u}_j) = 0$, which is reasonable and induces only small error as the values of h_i fall off quickly and significant contribution is localized. This way, we never require more storage than for $(K + 1) \times \#\mathcal{V}$ values and indices for the cost of $\#\mathcal{V}$ K-element sorts after each solution of the Laplace equation.

10.5.2 Intersection with Iso-contours

The intersections of triangles with an iso-contour $h_i(\mathbf{u}) = c$ can be implemented by a local search without additional data structures: Starting from the vertex associated with the feature f_i, i.e. where $h_i(\mathbf{u}_i) = 1$, we apply a gradient descent (h_i is monotone) on an arbitrary triangle attached to this vertex. We keep descending neighboring triangles until we hit a triangle that is intersected by the iso-contour. We then iteratively traverse all neighboring triangles which are also intersected.

10.5.3 Prefiltering of SIFT Features and Adaptive Refinement

Coarse correspondences identified in Sect. 10.3 may be distributed unevenly on the surface and can therefore be redundant if concentrated in certain areas. We can exploit this redundancy and reduce computation time by prefiltering keeping only a well-distributed subset. To identify the active feature subset, we partition the surface into patches with similar geodesic radius or geometric complexity. For each resulting surface cell, we maintain only one coarse feature (colored regions in Fig. 10.3(a)). In local sub-regions this reduction of coarse correspondences may lead to too few adjacent “cells” to yield meaningful coordinates. There we raise the number of coarse correspondences, thereby adaptively increase the patch density and then proceed iteratively as described above. Fig. 10.3(b) shows that –

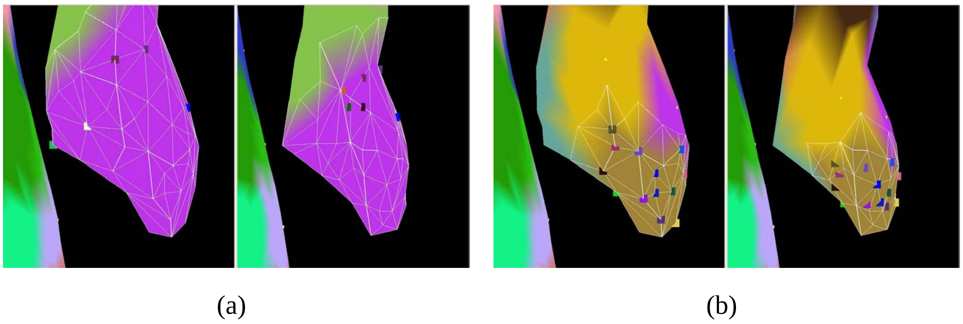

(a) (b)

Figure 10.3: Feature prefiltering and refinement. (a) zoom-in onto hand region of the model at two subsequent time steps. Colored areas represent surface regions. Due to sparse distribution of coarse features, the correspondences (colored dots) are not correct. (b) Adaptively increasing the number of coarse features leads to accurate correspondences.

on this particular data set – the latter greatly improves matching robustness in the hand region of the reconstructed human.

10.6 Alignment by Deformation

One intriguing advantage of our approach is that in the ideal case the dense correspondence field specifies the complete alignment of $\mathcal{M}_0(t)$ and $\mathcal{M}(t+1)$. To register the two meshes, we can therefore trivially move vertex locations without having to resort to involved deformation schemes. In practice, we find it advantageous to apply a fast and simple Laplacian deformation scheme rather than to perform vertex displacements only. This setting allows for trivial enforcement of surface smoothness during alignment hence smoothing out noise and mismatches. We refer to the recent survey [Botsch07] and the references therein for information on the method and its many variants. Laplacian deformation helps us to cure local reconstruction inaccuracies which may occur in surface regions for which feature localization was non-trivial, e.g. due to texture uniformity. Also, we take care that no loss of volume is introduced by the latter deformation approach: in rare cases where this becomes necessary, we force vertices of $\mathcal{M}_0(t)$ back onto $\mathcal{M}(t+1)$ along the shortest distance. This way we effectively deform $\mathcal{M}_0(t)$ to time-step $t+1$, and as we iterate the whole matching process over time, we track a single consistent mesh over the whole sequence, see Fig. 9.1 and Fig. 10.6

10.7 Results

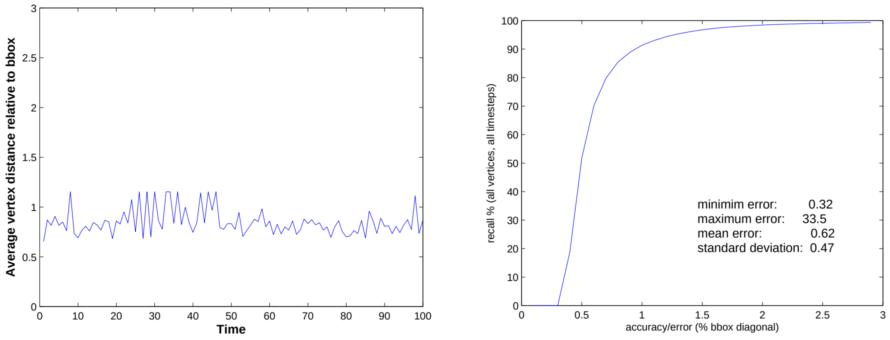
To demonstrate the performance of our reconstruction approach, we recorded two real-world motion sequences in our multi-camera system. The first sequence comprising of 105 frames shows a walking subject, Fig. 9.1(a)-(d), and the second sequence comprising of 100 frames shows a human performing a simple capoeira move, Fig. 10.6. As shown in these images as well as the accompanying video [C08b], our method enables faithful reconstruction of spatio-temporally coherent animations from this footage. A side-by-side comparison of the original input sequence and the reconstructed mesh sequence shows that our method delivers coherent scene geometry with low tangential distortion. When texturing our result with a fixed checkerboard, coherence and low distortion properties become very obvious, see Fig. 9.1(e),(f). We chose this visualization as texturing with the input video images would hide any geometric distortions.

Our algorithm is computationally more efficient than most deformation-based registration methods (see Sect. 9.1). Even if very detailed meshes comprising of roughly 10,000 vertices are reconstructed (Fig. 10.6(a)-(d)) and almost 600 coarse features are used, correspondences between pairs of frames can be computed in approximately 2 minutes on a Pentium IV 3.0 GHz. Prefiltering and adaptive refinement down to 120 coarse matches reduces alignment time to 1 minute per frame. In the more likely and practical case that mesh complexity is around 400 vertices, two frames can be aligned in as fast as 2 seconds even without prefiltering.

Even if surface triangulations are very coarse, our method produces high-quality coherent mesh animations and the advantages of the coherent mesh representation become even more evident. In the non-coherent version large triangulation differences between adjacent frames, Fig. 10.6(g),(h), lead to strong temporal noise which is practically eliminated in the coherent reconstructions, Fig. 10.6(e),(f).

10.8 Validation

In order to measure the accuracy of our algorithm we created a synthetic ground truth video sequence by texturing a virtual human character model (skeleton+surface mesh) with a constant noise texture, animating the model with captured motion data, and rendering it back into 16 virtual camera views. By this means, we obtain for each time step a ground truth 3D model with constant triangulation, as well as respective image data. To compare our results against



(a) (b)

Figure 10.4: (a) Average vertex distance (in $\mathbb{R}^3$) over time. (b) Recall accuracy (geodesic) for all vertices in complete sequence. Errors given w.r.t. ground truth sequence in % of bounding box size (1% error ~ 1.8 cm)

ground truth, we reconstruct visual hull meshes for all frames of the synthetic input and align the ground truth 3D model of the first frame with all subsequent ones. Fig. 10.4(a) shows that the average vertex distance between the ground truth and the coherent reconstruction remains at a very low level of 1% of the bounding box dimension over time. The plot also shows no significant error drift which underlines the robustness of our algorithm. Fig. 10.4(b) shows recall accuracy: for more than 90% of the vertices (all time-steps) we are within 1% bounding box diagonal (< 2cm) error radius.

By comparing the overlap between the coherent animations and the input silhouette images, we can assess the reconstruction quality of real sequences. On average, around 2.4% of the input silhouette pixels do not overlap with the reprojection which corresponds to an almost perfect match between input and our result, see Fig. 10.5(b). This comparison also clearly shows that dense correspondences are indeed needed to achieve this quality level as a deformation based on coarse features alone leads to a high residual alignment error, Fig. 10.5(a).

Our visual and quantitative results confirm effectiveness and efficiency of our method. In the following we discuss some properties and limitations inherent to the approach.

As we reconstruct shape from silhouette in every frame, the quality of results depends on the quality of the input video data and may suffer from artifacts attributed to the visual hull method itself. Some of the apparent phantom volumes in the results are solely due to the inability of shape-from-silhouette method to reconstruct concavities, and they are not introduced by our correspondence method. The focus of this method is not improving per-time step shape-reconstruction itself, and

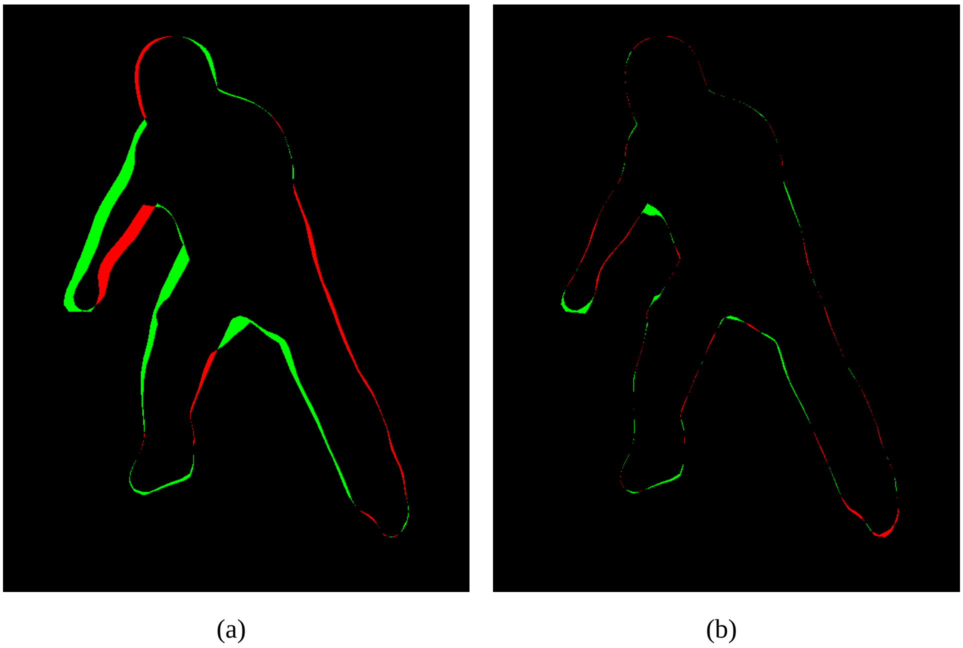

(a) (b)

Figure 10.5: Overlap of silhouettes of input and reprojected reconstructions in one camera view (red: non-overlapping pixels of input silhouette; green: non-overlapping pixels of reconstruction). (a) Coarse correspondences alone don't lead to a satisfactory alignment. (b) Dense correspondences, however, lead to an almost perfect alignment.

our method could be used in just the same way with more advanced reconstruction methods that also enforce photo-consistency, such as space carving.

Comparing to related work by Starck et al. [Starck05], our approach is more flexible (handles surfaces of arbitrary genus) and more efficient [Starck] as it does not rely on spherical parametrization, which is a non-trivial problem in its own. For their recent follow-up paper [Starck07a], we first remark that their goal is different in that wide time-frames are taken into account to solve a global problem. Hence, it is natural that our local approach is much more efficient. At the same time is accurate (they report typical errors of 5–10cm in their setting) and provides a map for *any* surface point.

Also, some video sequences show a fair amount of motion blur, and hence some reconstruction errors appear which could be easily overcome with faster cameras. Despite these unfaithful reconstructions our tests show the robustness of our method.

Our approach does not require surface parametrization. However, it shares one limitation with most practical parametrization methods, namely the absence of guarantees to obtain a valid one-to-one mapping: this means local fold-overs may occur when triangles are mapped between surfaces [Hormann07]. In practice, the alignment by means of Laplacian deformation smoothes out such local mismatches. This fact and experiments back the assumption of nearly isometric deformations.

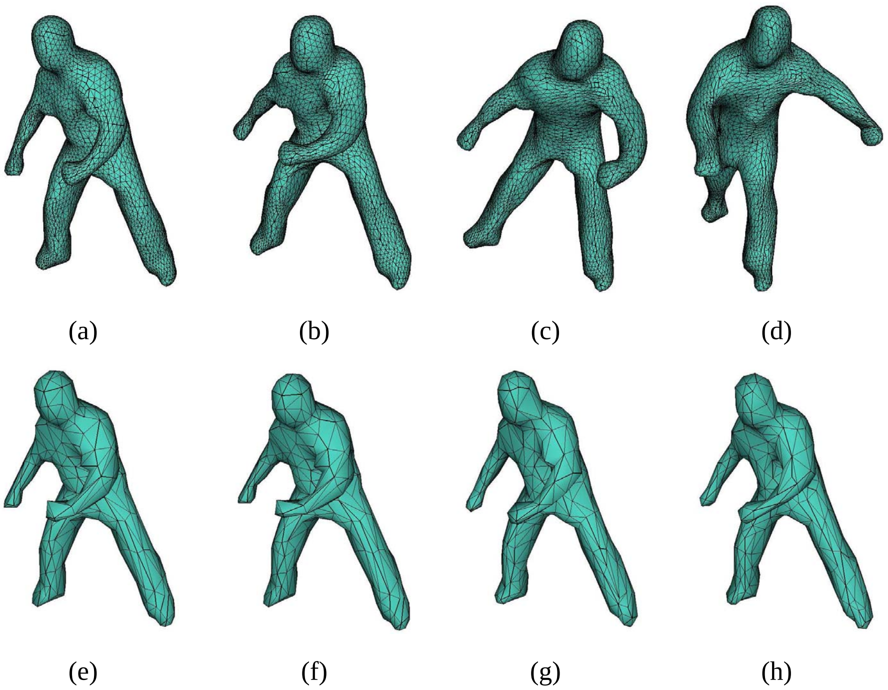
From a theoretical point of view our method is not proven to handle changes of the surface topology over time: "coordinate" functions might be locally unrelated in this situation, hence there is no guarantee that results are meaningful in the affected surface regions. Note that similar arguments are true for *any* method relying on local isometry which is not given under topology changes. In practice however, our method performs robustly towards typically observed topology changes (such as arms and legs merging in the visual hulls) similarly to [Starck07a]. To illustrate this robust handling, the video contains two synthetically generated example sequences (similar to the sequence used for accuracy measurement) in which arms and legs merge with the rest of the body. Generally, our goal is spatio-temporally coherent reconstruction, hence, topology changes should be avoided or corrected during the initial reconstruction step.

We gave intuitive motivation for selecting suitable "coordinate" functions and applying appropriate matching of surface points. We should remark that several aspects of our approach are based on heuristics which are justified only empirically, in particular the choice of distance measure D_h. An alternative approach might be based on learning techniques which compute perfectly parametrized distance functions for training sets.

Despite these limitations we have presented a robust and efficient dense correspondence finding method that enables spatio-temporally coherent animation reconstruction from multi-view video footage.

10.9 Conclusion

We presented a method to establish dense surface correspondences between originally unrelated shape-from-silhouette volumes that have been reconstructed from multi-view video. Our approach relies on sparse robust optical features from which dense correspondence is inferred in a discretization-independent way and without the use of parametrization techniques. Dense correspondences serve as mapping between surfaces to align a mesh with constant connectivity to all per-time-step reconstructions. Our experiments confirm efficiency and robustness of



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

(e) (f) (g) (h)

Figure 10.6: (a)-(d) Sample frames from a spatio-temporally coherent reconstruction of a capoeira move. Note that the actor's shape is faithfully reconstructed and triangle distortions are low. Remaining geometry artifacts are solely due to limitations of shape-from-silhouette methods. – The advantage of our reconstruction becomes very apparent in case of coarse triangulations (~ 750 **triangles). (e), (f) show subsequent frames from our reconstruction, and (g),(h) the same frames from the non-coherent input. The triangulation in the former models remains very consistent while in the latter case the triangulation dramatically changes even from one time step to the next.**

our approach, even in the presence of topology changes. As results we reconstruct animations from video as a deforming mesh with constant structure and low tangential distortion. This kind of input is required by subsequent higher-level processing tasks, such as analysis, compression, reconstruction improvement, etc.

Our method allows us to reconstruct spatio-temporally coherent geometry of arbitrary scenes directly from multi-view video data. Earlier in this thesis, we presented solutions to different problems that made use of the template mesh for capturing the shape and motion of the human actor. In order for those methods to work correctly with different subject, e.g. animals, the availability of the correct template geometry was necessary. We can completely replace the template geometry along with shape and motion capture with our method, which provides spatio-temporally coherent dynamic geometry of arbitrary scenes. This allows us to apply variety of video-based rendering, relighting or modeling algorithms over a wide range of multi-view video sequences, even if no template is available. It should be noted that a template model would not suffer by the limitations induced by the geometry reconstruction methods, e.g. concavities, low level of detail etc. On the other hand, our method allows a higher degree of flexibility in comparison to using a template model at the cost of the lower overall accuracy.

Chapter 11

Conclusions and Future Work

In this thesis we presented algorithmic solutions for a number of problems that arise in the reconstruction of high quality animation of humans from multi-view video data. Although the solution to each problem has been treated individually, they could also be combined to constitute a complete animation pipeline for acquisition, reconstruction and rendering of high quality virtual actors from multi-view video data. Even though the focus of the methods is reconstructing animation of real-world human actors, their fundamental principals can be applied to a larger class of real-world scenes.

In part I of this thesis, we described the fundamental components that are commonly used in all the algorithmic solutions described in the thesis. We described how to model the shape, appearance and kinematics of a human. We also described methods to animate the digitized human body model using both the kinematic skeleton and deformation. Either of the two animation techniques has been used throughout the solutions. In Chapter 3 we described our multi-view video studio, which is used to acquire high quality synchronized multi-view video streams under calibrated cameras and lighting. The acquired multi-view video streams are used in all of the presented algorithmic solutions.

In part II of this thesis we presented an automatic model-based approach to generate a personalized avatar from multi-view video streams showing a moving person. This solution is tailored for a specific scenario where the complexity of the model is limited by the available resources. We create high quality personalized human avatars with simplest of model descriptions. The avatar's geometry is generated by shape adapting a template human body model. Its surface texture is assembled from multi-view video frames showing arbitrary different body poses. The generated static texture can be used to render the complete human animation with just a single texture. Personalized human avatars allow for the photo-realistic rendition of real-world humans with a minimal model description. We demonstrated that they can be easily incorporated in virtual environments.

In part III of this thesis, we described methods that allow us to reconstruct high quality relightable free-viewpoint video from multi-view video data. We extended the earlier work in the area of dynamic surface reflectance estimation. First, we first described a method to improve spatio-temporal registration of the dynamic texture. The new method detects and compensates the shifting of the apparel over the body's surface of the actor. Exact texture registration is one of the main requirement for the accurate estimation of surface reflectance. Our second contribution was a spatio-temporal reflectance sharing method that reduces the bias in the estimated dynamic reflectance. This method assures that the estimated reflectance properties are not biased towards the recording environment. We validated our methods both visually and quantitatively.

In part IV of this thesis, we presented one of the first passive methods to reconstruct geometry of large dynamic scenes showing moving actors at unprecedented detail and accuracy from video only. Methods presented earlier in the thesis, used models that did not have embedded high resolution dynamic shape details. For relightable free-viewpoint video, we measured dynamic surface normal field parameterized over the smooth template. This was sufficient for rendering relightable free-viewpoint video from many angles apart from grazing ones, which require the details in the geometry. Also, the conversion of potentially noise-contaminated normal field parameterized over an arbitrarily shaped smooth surface into highly-detailed time-varying scene geometry, is a difficult problem in itself. We make use of the previous work in relightable free-viewpoint video, and improve the original reflectance estimation and normal estimation approach by employing robust statistics to handle sensor noise faithfully. Later, we applied a new MRF-based spatio-temporal surface deformation approach that converts the geometric details encoded in the normals into true dynamic 3D displacements.

Adding time-varying details to the geometry results in very high quality animations. Moreover, our method is completely passive and does not require any additional information other than multi-view video streams recorded under calibrated light sources. Our method can reconstruct subtle dynamic geometry details, such as wrinkles and folds in clothing. Our spatio-temporal reconstruction method outputs displaced geometry that is accurate at each time step of the video and temporally smooth, even if the input data are affected by noise.

In part V of this thesis, we described a method to establish dense surface correspondences between originally unrelated shape-from-silhouette volumes that have been reconstructed from multi-view video. The method uses sparse robust features that are used as the anchor points from which the dense correspondence is established. Dense correspondences are not only discretization-independent but also do not use any parameterization technique. This assures that the method does not suffer from any parameterization induced limitations, e.g. points at singularities. The method establishes dense correspondence between two volumes that are reconstructed from adjacent frames of the video. Dense correspondences allow trivial deformation of one volume to the other. Starting from the first two frames, the dense correspondences are propagated over the whole sequences, with the starting volume being deformed at each time steps. This results in a spatio-temporally coherent animation as a deforming mesh with low tangential distortion.

Spatio-temporally coherent scene geometry is an important and highly required property in captured animations. The output from our method can be directly used in the solutions presented earlier in this thesis. All the earlier method used a template model for capturing the shape and tracking the motion of the human actor. Thus the availability of the template and its accuracy was one of the limitations in all of the methods. Spatio-temporally coherent geometry of arbitrary scenes removes the template induced limitations and allows direct application of the video-based algorithms over multi-view video data. Additionally, spatio-temporal coherence is an explicit requirement for some tasks, such as surface reflectance estimation, compression, motion analysis, editing, reconstruction improvements, etc. Our proposed method is not only very efficient but is also very robust even in the presence of topology changes. Even though we have only used the recordings of human actors for our experiments, the method can be applied on any subject as long as the high spatial details in the input video are present.

The methods presented in the thesis demonstrate that it is now possible to passively reconstruct high quality 3D animation from multi-view video data. We would like to note that the arrival of high quality and high accuracy acquisition equipments have also played a part in the development of these methods. We can envision that in the future, with even higher resolution and accuracy video acquisition, the reconstructions quality or at least the final renditions should be even better. The methods presented in this thesis represent some early steps in the direction of high quality 3D animation reconstruction from video. They confirm that this goal is not only achievable but can already be put into practice with their use.

Bibliography

[Agarwal03] SAMEER AGARWAL, RAVI RAMAMOORTHI, SERGE BELONGIE, AND HENRIK WANN JENSEN. Structured importance sampling of environment maps. In *SIGGRAPH '03: ACM SIGGRAPH 2003 Papers*, pages 605–612, New York, NY, USA, 2003. ACM.

[Agrawal06] AMIT K. AGRAWAL, RAMESH RASKAR, AND RAMA CHELLAPPA. What Is the Range of Surface Reconstructions from a Gradient Field? In *Proc. of ECCV*, pages 578–591, 2006.

[Ahmed05] NAVEED AHMED, EDILSON DE AGUIAR, CHRISTIAN THEOBALT, MARCUS MAGNOR, AND HANS-PETER SEIDEL. Automatic generation of personalized human avatars from multi-view video. In *VRST '05: Proceedings of the ACM symposium on Virtual reality software and technology*, pages 257–260, New York, NY, USA, 2005. ACM.

[Ahmed07a] NAVEED AHMED, CHRISTIAN THEOBALT, MARCUS A. MAGNOR, AND HANS-PETER SEIDEL. Spatio-Temporal Registration Techniques for Relightable 3D Video. In *ICIP (2)*, pages 501–504. IEEE, 2007.

[Ahmed07b] NAVEED AHMED, CHRISTIAN THEOBALT, AND HANS-PETER SEIDEL. Spatio-temporal Reflectance Sharing for Relightable 3D Video. In Andr Gagalowicz and Wilfried Philips, editors, *MIRAGE*, volume 4418 of *Lecture Notes in Computer Science*, pages 47–58. Springer, 2007.

[Ahmed08a] NAVEED AHMED, CHRISTIAN THEOBALT, PETAR DOBREV, SEBASTIAN THRUN, AND HANS-PETER SEIDEL. Robust Fusion of Dynamic Shape and Normal Capture for High-quality Reconstruction of Time-varying Geometry. In *CVPR*, Anchorage, Alaska, 2008. IEEE Computer Society.

[Ahmed08b] Naveed Ahmed, Christian Theobalt, Christian Rössl, Sebastian Thrun, and Hans-Peter Seidel. Dense Correspondence Finding for Parametrization-free Animation Reconstruction from Video. In *CVPR*, Anchorage, Alaska, 2008. IEEE Computer Society.

[Alliez07] P. Alliez, G. Ucelli, C. Gotsman, and Marco Attene. Recent Advances in Remeshing of Surfaces. In *Shape Analysis and Structuring*. Spinger, 2007.

[Anguelov04] D. Anguelov, D. Koller, P. Srinivasan, S. Thrun, H.-C. Pang, and J. Davis. The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces. In *Proc. NIPS*, 2004.

[Baran07] Ilya Baran and Jovan Popović. Automatic rigging and animation of 3D characters. *ACM Trans. Graph.*, 26(3):72, 2007.

[Barron92] J.L. Barron, D.J. Fleet, S.S. Beauchemin, and T.A. Burkitt. Performance Of Optical Flow Techniques. In *CVPR*, pages 236–242, 1992.

[Bernardini01] Fausto Bernardini, Ioana M. Martin, and Holly Rushmeier. High-quality texture reconstruction from multiple scans. *IEEE TVCG*, 7(4):318–332, 2001.

[Boivin01] Samuel Boivin and André Gagalowicz. Image-Based Rendering of Diffuse, Specular and Glossy Surfaces From a Single Image. In *Proc. of ACM SIGGRAPH 2001*, pages 107–116, 2001.

[Borovikov00] E. Borovikov and L. Davis. A Distributed System for Real-Time Volume Reconstruction. In *Proceedings of Intl. Workshop on Computer Architectures for Machine Perception*, page 183ff, 2000.

[Botsch07] M. Botsch and O. Sorkine. On linear variational surface deformation methods. *IEEE TVCG*, 2007.

[Brostow04] Gabriel J. Brostow, Irfan Essa, Drew Steedly, and Vivek Kwatra. Novel Skeletal Representation For Articulated Creatures. In *ECCV04*, pages Vol III: 66–78, 2004.

[Byrd95] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ci You Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(6):1190–1208, 1995.

[C08a] http://www.mpi-inf.mpg.de/∼nahmed/CVPR08b.wmv .

[C08b] http://www.mpi-inf.mpg.de/∼nahmed/CVPR08a.wmv .

[Carranza03] J. CARRANZA, C. THEOBALT, M.A. MAGNOR, AND H.-P. SEIDEL. Free-Viewpoint Video of Human Actors. In *Proc. of SIGGRAPH'03*, pages 569–577, 2003.

[Chang07] JU YONG CHANG, KYOUNG MU LEE, AND SANG UK LEE. Multiview normal field integration using level set methods. In *CVPR*, 2007.

[Cheung00] K. M. CHEUNG, T. KANADE, J.-Y. BOUGUET, AND M. HOLLER. A Real Time System for Robust 3D Voxel Reconstruction of Human Motions. In *Proc. of CVPR*, volume 2, pages 714 – 720, June 2000.

[Cheung03] G. CHEUNG, S. BAKER, AND T. KANADE. Shape-from-silhouette for articulated objects and its use for human body kinematics estimation and motion capture. In *Proc. CVPR*, 2003.

[de Aguiar04] E. DE AGUIAR, C. THEOBALT, M. MAGNOR, H. THEISEL, AND H.-P. SEIDEL. Marker-free Model Reconstruction and Motion Tracking from 3D Voxel Data. *Proc. IEEE Pacific Graphics 2003*, Seoul, South Korea, pages 101–110, October 2004.

[de Aguiar05] EDILSON DE AGUIAR, CHRISTIAN THEOBALT, MARCUS MAGNOR, AND HANS-PETER SEIDEL. Reconstructing Human Shape and Motion from Multi-View Video. In *2nd European Conference on Visual Media Production (CVMP)*, pages 42–49, London, UK, December 2005. The IEE.

[de Aguiar07a] E. DE AGUIAR, C. THEOBALT, C. STOLL, AND H.-P. SEIDEL. Rapid Animation of Laser-scanned Humans. In *IEEE Virtual Reality 2007*, pages 223–226, 2007.

[de Aguiar07b] EDILSON DE AGUIAR, CHRISTIAN THEOBALT, CARSTEN STOLL, AND HANS-PETER SEIDEL. Marker-less Deformable Mesh Tracking for Human Shape and Motion Capture. In *Proc. CVPR*, pages 1–8. IEEE, 2007.

[de Aguiar08] E. DE AGUIAR, C. STOLL, C. THEOBALT, N. AHMED, H.-P. SEIDEL, AND S. THRUN. Performance Capture from Sparse Multi-view Video. In *ACM TOG (Proc. SIGGRAPH)*, 2008.

[Debevec97] PAUL E. DEBEVEC AND JITENDRA MALIK. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 369–378, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

[Debevec00] P. DEBEVEC, T. HAWKINS, C. TCHOU, H.-P. DUIKER, W. SAROKIN, AND M. SAGAR. Acquiring the Reflectance Field of a Human Face. *Proc. of ACM SIGGRAPH'00*, pages 145–156, 2000.

[Einarsson06] PER EINARSSON, CHARLES-FELIX CHABERT, ANDREW JONES, WAN-CHUN MA, BRUCE LAMOND, IM HAWKINS, MARK BOLAS, SEBASTIAN SYLWAN, AND PAUL DEBEVEC. Relighting Human Locomotion with Flowed Reflectance Fields. In *Rendering Techniques*, pages 183–194, 2006.

[Elad03] A. ELAD AND R. KIMMEL. On Bending Invariant Signatures for Surfaces. *IEEE Trans. PAMI*, 25(10):1285–1295, 2003.

[Farin99] GERALD FARIN. *Curves and Surfaces for CAGD: A Practical Guide*. Morgan Kaufmann, 1999.

[Franco03] JEAN-SÉBASTIEN FRANCO AND EDMOND BOYER. Exact Polyhedral Visual Hulls. In *Proc. of BMVC*, pages 329–338, 2003.

[Frankot88] ROBERT T. FRANKOT AND RAMA CHELLAPPA. A Method for Enforcing Integrability in Shape from Shading Algorithms. *IEEE Trans. PAMI*, 10(4):439–451, 1988.

[Fua94] PASCAL FUA AND YVAN G. LECLERC. Using 3-Dimensional Meshes To Combine Image-Based and Geometry-Based Constraints. In *Proc. of ECCV*, pages 281–291, 1994.

[Fua98] P. FUA, A. GRUEN, R. PLANKERS, N. APUZZO, AND D. THALMANN. Human body modeling and motion analysis from video sequences. In *Proc. Int. Symp. on Real-Time Imaging and Dynamic Analysis*, 1998.

[Gal06] RAN GAL AND DANIEL COHEN-OR. Salient geometric features for partial shape matching and similarity. *ACM TOG*, 25(1):130–150, 2006.

[Gardner03] A. GARDNER, C. TCHOU, T. HAWKINS, AND P. DEBEVEC. Linear light source reflectometry. *ACM Trans. Graphics. (Proc. of SIGGRAPH'03)*, 22(3):749–758, 2003.

[Georghiades03] ATHINODOROS S. GEORGHIADES. Recovering 3-D Shape and Reflectance From a Small Number of Photographs. In *Eurographics Symposium on Rendering*, pages 230–240, 2003.

[Gibson01] SIMON GIBSON, TOBY HOWARD, AND ROGER HUBBOLD. Flexible Image-Based Photometric Reconstruction using Virtual Light Sources. *Computer Graphics Forum*, 20(3), 2001.

[Goesele00] M. GOESELE, H. LENSCH, W. HEIDRICH, AND H.-P. SEIDEL. Building a Photo Studio for Measurement Purposes. In *Proc. of VMV2000*, pages 231–238, 2000.

[Goldman04] D. GOLDMAN, B. CURLESS, A. HERTZMANN, AND S. SEITZ. Shape and Spatially-Varying BRDFs From Photometric Stereo. In *Proc. of ICCV*, pages 341–448, 2004.

[Gross03] MARKUS H. GROSS, STEPHAN WÜRMLIN, MARTIN NÄF, EDOUARD LAMBORAY, CHRISTIAN P. SPAGNO, ANDREAS M. KUNZ, ESTHER KOLLER-MEIER, TOMÁS SVOBODA, LUC J. VAN GOOL, SILKE LANG, KAI STREHLKE, ANDREW VANDE MOERE, AND OLIVER G. STAADT. blue-c: a spatially immersive display and 3D video portal for telepresence. *ACM Trans. Graph. (Proc. of SIGGRAPH'03)*, 22(3):819–827, 2003.

[Hähnel03] D. HÄHNEL, S. THRUN, AND W. BURGARD. An Extension of the ICP Algorithm for Modeling Nonrigid Objects with Mobile Robots. In *Proc. of IJCAI*, 2003.

[Hartley00] R. HARTLEY AND A. ZISSERMAN. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[Hawkins04] T. HAWKINS, A. WENGER, C. TCHOU, A. GARDNER, F. GÖRANSSON, AND P. DEBEVEC. Animatable Facial Reflectance Fields. In *Proc. of Eurographics Symposium on Rendering*, pages 309–319, 2004.

[Heikkila96] J. HEIKKILA AND O. SILVEN. Calibration Procedure for Short Focal Length Off-the-shelf CCD Cameras. In *Proc. of 13th ICPR*, pages 166–170, 1996.

[Hernández04] C. HERNÁNDEZ AND F. SCHMITT. Silhouette and Stereo Fusion for 3D Object Modeling. *Computer Vision and Image Understanding, special issue on 'Model-based and image-based 3D Scene Representation for Interactive Visualization'*, December 2004.

[Hernández07] CARLOS HERNÁNDEZ, GEORGE VOGIATZIS, GABRIEL J. BROSTOW, BJOÖRN STENGER, AND ROBERTO CIPOLLA. Non-Rigid Photometric Stereo with Colored Lights. In *Proc. of ICCV*, 2007.

[Hilton99] ADRIAN HILTON, DANIEL BERESFORD, THOMAS GENTILS, RAYMOND SMITH, AND WEI SUN. Virtual People: Capturing Human Models to Populate Virtual Worlds. In *CA '99: Proceedings of the Computer Animation*, page 174, Washington, DC, USA, 1999. IEEE Computer Society.

[Hormann07] KAI HORMANN, BRUNO LEVY, AND ALLA SHEFFER. Mesh Parameterization: Theory and Practice. In *SIGGRAPH Course Notes*, 2007.

[Huber03] DANIEL HUBER AND MARTIAL HEBERT. Fully automatic registration of multiple 3D data sets. *IVC*, 21(7):637–650, July 2003.

[Huber04] P.J. HUBER. *Robust Statistics*. Wiley, 2004.

[Jain95] R. JAIN, R. KASTURI, AND B. G. SCHUNCK. *Machine Vision*. McGraw Hill International, 1995.

[Jones06] ANDREW JONES, ANDREW GARDNE, MARK BOLAS, IAN MCDOWALL, AND PAUL DEBEVEC. Performance Geometry Capture for Spatially Varying Relighting. In *Proc. of CVMP*, 2006.

[Kakadiaris95] I. A. KAKADIARIS AND D. METAXAS. 3D human body model acquisition from multiple views. In *ICCV '95: Proceedings of the Fifth International Conference on Computer Vision*, page 618, Washington, DC, USA, 1995. IEEE Computer Society.

[Kanade97] TAKEO KANADE, PETER RANDER, AND P. J. NARAYANAN. Virtualized Reality: Constructing Virtual Worlds from Real Scenes. *IEEE MultiMedia*, 4(1):34–47, 1997.

[Kanade98] T. KANADE, H. SAITO, AND S. VEDULA. The 3D Room: Digitizing Time-Varying 3D Events by Synchronized Multiple Video Streams. Technical Report CMU-RI-TR-98-34, Robotics Institute - Carnegie Mellon University, 1998.

[Kazhdan06] M. KAZHDAN, M. BOLITHO, AND H. HOPPE. Poisson Surface Reconstruction. In *Proc. SGP*, pages 61–70, 2006.

[Kück04] H. KÜCK, W. HEIDRICH, AND C. VOGELGSANG. Shape from Contours and Multiple Stereo - A Hierarchical, Mesh-Based Approach. In *CRV*, pages 76–83, 2004.

[Kutulakos00] KIRIAKOS N. KUTULAKOS AND STEVEN M. SEITZ. A Theory of Shape by Space Carving. *Int. J. Comput. Vision*, 38(3):199–218, 2000.

[Lafortune97a] E. LAFORTUNE, S. FOO, K. TORRANCE, AND D. GREENBERG. Non-Linear Approximation of Reflectance Functions, August 1997.

[Lafortune97b] ERIC P. F. LAFORTUNE, SING-CHOONG FOO, KENNETH E. TORRANCE, AND DONALD P. GREENBERG. Non-linear approximation of reflectance functions. In *Proc. of SIGGRAPH'97*, pages 117–126. ACM Press, 1997.

[Lange99] HOLGER LANGE. Advances in the Cooperation of Shape from Shading and Stereo Vision. *3dim*, 00:0046, 1999.

[Lee00] W-S. LEE, J. GU, AND N. MAGNENAT-THALMANN. Generating Animatable 3D Virtual Humans from Photographs. In M. Gross and F. R. A. Hopgood, editors, *Computer Graphics Forum (Eurographics 2000)*, volume 19(3), 2000.

[Lensch03] HENDRIK P. A. LENSCH, JAN KAUTZ, MICHAEL GOESELE, WOLFGANG HEIDRICH, AND HANS-PETER SEIDEL. Image-Based Reconstruction of Spatial Appearance and Geometric Detail. *ACM Transactions on Graphics*, 22(2):27, 2003.

[Lensch04] HENDRIK P. A. LENSCH. *Efficient, Image-Based Appearance Acquisition of Real-World Objects*. PhD thesis, Universität des Saarlandes, Göttingen, Germany, March 2004.

[Levoy96] MARC LEVOY AND PAT HANRAHAN. Light field rendering. In *in Proc. of ACM SIGGRAPH'96*, pages 31–42, 1996.

[Li02] MING LI, HARTMUT SCHIRMACHER, MARCUS MAGNOR, AND HANS-PETER SEIDEL. Combining Stereo and Visual Hull Information for On-line Reconstruction and Rendering of Dynamic Scenes. In *Proc. of IEEE Multimedia and Signal Processing*, pages 9–12, 2002.

[Lowe99] DAVID G. LOWE. Object Recognition from Local Scale-Invariant Features. In *Proc. IEEE ICCV*, volume 2, page 1150ff, 1999.

[Lucas81] B. LUCAS AND T. KANADE. An iterative image registration technique with an application to Stereo Vision. In *Proc. DARPA IU Workshop*, pages 121–130, 1981.

[Luck02] J. LUCK AND D. SMALL. Real-Time Markerless Motion Tracking Using Linked Kinematic Chains. In *Proc. of CVPRIP*, 2002.

[Marschner98] S. MARSCHNER. *Inverse Rendering for Computer Graphics*. PhD thesis, Cornell University, 1998.

[Matsuyama02] T. MATSUYAMA AND T. TAKAI. Generation, Visualization, and Editing of 3D Video. In *Proc. of 1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT'02)*, page 234ff, 2002.

[Matsuyama04] T. MATSUYAMA, X. WU, T. TAKAI, AND S. NOBUHARA. Real-time 3D shape reconstruction, dynamic 3D mesh deformation and high fidelity visualization for 3D video. *CVIU*, 96(3):393–434, 2004.

[Matusik00] W. MATUSIK, C. BUEHLER, R. RASKAR, S.J. GORTLER, AND L. MCMILLAN. Image-Based Visual Hulls. In *Proceedings of ACM SIGGRAPH 00*, pages 369–374, 2000.

[Matusik01] W. MATUSIK, C. BUEHLER, AND L. MCMILLAN. Polyhedral Visual Hulls for Real-Time Rendering. In *Proceedings of 12th Eurographics Workshop on Rendering*, pages 116–126, 2001.

[Matusik03] W. MATUSIK, H. PFISTER, M. BRAND, AND L. MCMILLAN. A data-driven reflectance model. *ACM Trans. Graph. (Proc. SIGGRAPH'03)*, 22(3):759–769, 2003.

[Matusik04] WOJCIECH MATUSIK AND HANSPETER PFISTER. 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Trans. Graph. (Proc. of SIGGRAPH'04)*, 23(3):814–824, 2004.

[Moezzi97] SAIED MOEZZI, LI-CHENG TAI, AND PHILIPPE GERARD. Virtual View Generation for 3D Digital Video. *IEEE Multi-Media*, 4(1):18–26, 1997.

[Motion] ORGANIC MOTION. http://www.organicmotion.com.

[Narayanan98] P. J. NARAYANAN, P. RANDER, AND T. KANADE. Constructing Virtual Worlds using Dense Stereo. In *Proc. of ICCV'98*, pages 3–10, 1998.

[Nehab05] DIEGO NEHAB, SZYMON RUSINKIEWICZ, JAMES DAVIS, AND RAVI RAMAMOORTHI. Efficiently Combining Positions and Normals for Precise 3D Geometry. *ACM TOG*, 24(3), 2005.

[Nesys] NESYS. http://www.nesys.de/.

[Nishino01] K. NISHINO, Y. SATO, AND K. IKEUCHI. "Eigen-Texture Method: Appearance Compression and Synthesis based on a 3D Model". *IEEE Trans. PAMI*, 23(11):1257–1265, nov 2001.

[Paquette96] STEVEN PAQUETTE. 3D Scanning in Apparel Design and Human Engineering. *IEEE Comput. Graph. Appl.*, 16(5):11–15, 1996.

[Phong75] B.-T. PHONG. Illumnation for Computer Generated Pictures. *Communications of the ACM*, pages 311–317, 1975.

[Poppe07] R.W. POPPE. Vision-based Human Motion Analysis: An Overview. *CVIU*, 108:4–18, 2007.

[Ramamoorthi01] R. RAMAMOORTHI AND P. HANRAHAN. A Signal-Processing Framework for Inverse Rendering. In *Proceedings of SIGGRAPH 2001*, pages 117–128. ACM Press, 2001.

[Reuter06] M. REUTER, F.-E. WOLTER, AND N. PEINECKE. Laplace-Beltrami spectra as 'Shape-DNA' of surfaces and solids. *Computer-Aided Design*, 38(4):342–366, 2006.

[Rushmeier97] H. RUSHMEIER, G. TAUBIN, AND A. GUÉZIEC. Applying Shape from Lighting Variation to Bump Map Capture. In *Eurographics Workshop on Rendering*, pages 35–44, June 1997.

[Rusinkiewicz00] S. RUSINKIEWICZ AND S. MEASUREMENT MARSCHNER. Measurement I - BRDFs. Script of course CS448C: Topics in Computer Graphics, held at Stanford University, October 2000.

[Rusinkiewicz05] S. RUSINKIEWICZ, B. BROWN, AND M. KAZHDAN. 3D Scan Matching and Registration. In *ICCV short courses*, 2005.

[Sato97] YOICHI SATO, MARK D. WHEELER, AND KATSUSHI IKEUCHI. Object Shape and Reflectance Modeling from Observation. In *Proc. of SIGGRAPH'97*, pages 379–388, 1997.

[Shinya04] MIKIO SHINYA. Unifying Measured Point Sequences of Deforming Objects. In *Proc. of 3DPVT*, pages 904–911, 2004.

[Starck] JONATHAN STARCK. personal communication.

[Starck05] J. STARCK AND A. HILTON. Spherical Matching for Temporal Correspondence of Non-Rigid Surfaces. *IEEE ICCV*, pages 1387–1394, 2005.

[Starck06] J. STARCK, G. MILLER, AND A. HILTON. Volumetric Stereo with Silhouette and Feature Constraints. *Proc. of BMVC*, 3:1189–1198, 2006.

[Starck07a] J. STARCK AND A. HILTON. Correspondence labelling for wide-timeframe free-form surface matching. In *IEEE ICCV*, 2007.

[Starck07b] J. STARCK AND A. HILTON. Surface Capture for Performance Based Animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007.

[Stoll06] C. STOLL, Z. KARNI, C. RÖSSL, H. YAMAUCHI, AND H.-P. SEIDEL. Template Deformation for Point Cloud Fitting. In *Proc. SGP*, pages 27–35, 2006.

[Sumner04] ROBERT W. SUMNER AND JOVAN POPOVIC. Deformation transfer for triangle meshes. *ACM TOG (Proc. SIGGRAPH)*, 23(3):399–405, 2004.

[Theobalt03] C. THEOBALT, M. LI, M. MAGNOR, AND H.-P. SEIDEL. A Flexible and Versatile Studio for Multi-View Video Recording. In Peter Hall and Philip Willis, editors, *Vision, Video and Graphics 2003*, pages 9–16, Bath, UK, July 2003. Eurographics, Eurographics.

[Theobalt04] C. THEOBALT, J. CARRANZA, M. MAGNOR, AND H.-P. SEIDEL. Combining 3D Flow Fields with Silhouette-based Human Motion Capture for Immersive Video. *Graphical Models*, 66:333–351, September 2004.

[Theobalt05a] CHRISTIAN THEOBALT. *From Image-based Motion Analysis to Free-Viewpoint Video*. PhD thesis, Universität des Saarlandes, December 2005.

[Theobalt05b] CHRISTIAN THEOBALT, NAVEED AHMED, EDILSON DE AGUIAR, GERNOT ZIEGLER, HENDRIK P. A. LENSCH, MARCUS MAGNOR, AND HANS-PETER SEIDEL. Joint Motion and Reflectance Capture for Creating Relightable 3D Videos. Research Report MPI-I-2005-4-004, Max-Planck-Institut fuer Informatik, Saarbruecken, Germany, April 2005.

[Theobalt07] CHRISTIAN THEOBALT, NAVEED AHMED, HENDRIK P. A. LENSCH, MARCUS MAGNOR, AND HANS-PETER SEIDEL. Seeing People in Different Light - Joint Shape, Motion and Reflectance Capture. *IEEE TVCG*, 2007.

[Tsai86] R. Y. TSAI. An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision. In *Proc. of CVPR*, pages 364–374, 1986.

[Wand07] MICHAEL WAND, PHILIPP JENKE, QIXING HUANG, MARTIN BOKELOH, LEONIDAS GUIBAS, AND ANDREAS SCHILLING. Reconstruction of deforming geometry from time-varying point clouds. In *Proc. SGP*, pages 49–58, 2007.

[Ward92] G. J. WARD. Measuring and Modeling Anisotropic Reflection. In *Proc. of SIGGRAPH*, pages 265–272, 1992.

[Wardetzky07] M. WARDETZKY, S. MATHUR, F. KLBERER, AND E. GRINSPUN. Discrete Laplace operators:No free lunch. In *Proc. SGP*, pages 33–37, 2007.

[Waschbüsch05] M. WASCHBÜSCH, S. WÜRMLIN, D. COTTING, F. SADLO, AND M. GROSS. Scalable 3D Video of Dynamic Scenes. In *Proc. of Pacific Graphics*, pages 629–638, 2005.

[Waschbüsch07] M. WASCHBÜSCH, S. WÜRMLIN, AND M. GROSS. 3D Video Billboard Clouds. In *Proc. Eurographics*, 2007.

[Weng90] J. WENG, P. COHEN, AND M. HERNIOU. Calibration of Stereo Cameras Using a Non-Linear Distortion Model. In *ICPR*, pages 246–253, 1990.

[Wenger05] A. WENGER, A. GARDNER, C. TCHOU, J. UNGER, T. HAWKINS, AND P. DEBEVEC. Performance Relighting and Reflectance Transformation with Time-Multiplexed Illumination. In *ACM TOG (Proc. of SIGGRAPH'05)*, volume 24(3), pages 756–764, 2005.

[Woodham89] ROBERT J. WOODHAM. Photometric method for determining surface orientation from multiple images. pages 513–531, 1989.

[Würmlin02] S. WÜRMLIN, E. LAMBORAY, O.G. STAADT, AND M.H. GROSS. 3D Video Recorder. In *Proc. of IEEE Pacific Graphics*, pages 325–334, 2002.

[Würmlin03] S. WÜRMLIN, E. LAMBORAY, O. G. STAADT, AND M. H. GROSS. 3D Video Recorder: a System for Recording and Playing Free-Viewpoint Video. *Comput. Graph. Forum*, 22(2):181–194, 2003.

[Yu98] Y. YU AND J. MALIK. Recovering Photometric Properties of Architectural Scenes from Photographs. In *Proceedings of ACM SIGGRAPH'98*, pages 207–218, 1998.

[Yu99] Y. YU, P. DEBEVEC, J. MALIK, AND T. HAWKINS. Inverse Global Illumination: Recovering Reflectance Models of Real Scenes From Photographs. In *Proc. of ACM SIGGRAPH'99*, pages 215–224, August 1999.

[Zayer05a] RHALEB ZAYER, CHRISTIAN RÖSSL, ZACHI KARNI, AND HANS-PETER SEIDEL. Harmonic Guidance for Surface Deformation. *Computer Graphics Forum*, 24(3):601–609, 2005.

[Zayer05b] RHALEB ZAYER, CHRISTIAN RÖSSL, AND HANS-PETER SEIDEL. Discrete Tensorial Quasi-Harmonic Maps. In *Proc. of Shape Modeling International*, pages 276–285. IEEE, 2005.

[Zhang99] RUO ZHANG, PING-SING TSAI, JAMES CRYER, AND MUBARAK SHAH. Shape from Shading: A Survey. *IEEE Trans. PAMI*, 21(8):690–706, 1999.

[Zhang04] LI ZHANG, NOAH SNAVELY, BRIAN CURLESS, AND STEVEN M. SEITZ. Spacetime Faces: High-Resolution Capture for Modeling and Animation. In *ACM TOG*, pages 548–558, 2004.

[Zickler05] TODD ZICKLER, SEBASTIAN ENRIQUE, RAVI RAMAMOORTHI, AND PETER N. BELHUMEUR. Reflectance Sharing: Image-based Rendering from a Sparse Set of Images. In *Proc. of Eurographics Symposium on Rendering*, pages 253–264, 2005.

[Ziegler] G. ZIEGLER, H. LENSCH, N. AHMED, M. MAGNOR, AND H.-P. SEIDEL. Multi-Video Compression in Texture Space,.

[Zitnick04] C. LAWRENCE ZITNICK, SING BING KANG, MATTHEW UYTTENDAELE, SIMON WINDER, AND RICHARD SZELISKI. High-quality video view interpolation using a layered representation. *ACM TOC (Proc. SIGGRAPH'04)*, 23(3):600–608, 2004.

Appendix A

List of Publications

[A] N. Ahmed, C. Theobalt, P. Dobrev, H.P. Seidel, S. Thrun: *Robust Fusion of Dynamic Shape and Normal Capture for High-quality Reconstruction of Time-varying Geometry*. In Proc. of CVPR 2008, Anchorage, USA.

[B] N. Ahmed, C. Theobalt, C. Rössl, H.P. Seidel, S. Thrun: *Dense Correspondence Finding for Parametrization-free Animation Reconstruction from Video*. In Proc. of CVPR 2008, Anchorage, USA.

[C] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel and S. Thrun: *Performance Capture from Sparse Multi-view Video*. In Proc. of ACM SIGGRAPH 2008, Los Angeles, USA.

[D] M. Eisemann, B. de Decker, M. Magnor, P. Bekaert, E. de Aguiar, N. Ahmed, C. Theobalt and A. Sellent: *Floating Textures. In Proc. of EUROGRAPHICS 2008 (Computer Graphics Forum, vol. 27 issue 2), Crete, Greece.*

[E] C. Theobalt, N. Ahmed, G. Ziegler, H.P. Seidel: *High-quality Reconstruction of Virtual Actors from Multi-view Video Streams*. In IEEE Signal Processing Magazine, 2007.

[F] N. Ahmed, C. Theobalt, M. Magnor, H.P. Seidel: *Spatio-Temporal Registration Techniques for Relightable 3D Video*. In Proc. of ICIP 2007, San Antonio, USA.

[G] N. Ahmed, C. Theobalt, H.P. Seidel: *Spatio-temporal Reflectance Sharing for Relighatble 3D Video*. In Proc. of Mirage 2007, Paris, France.

[H] C. Theobalt, N. Ahmed, H.Lensch, M. Magnor, H.P. Seidel: *Seeing People in Different Light: Joint Shape, Motion, and Reflectance Capture*. In IEEE Transactions on Visualization and Computer Graphics, 2007.

[I] N. Ahmed, E. de Aguiar, C. Theobalt, M. Magnor, H.-P. Seidel: *Automatic Generation of Personalized Human Avatars from Multi-view Video. In Proc. of the ACM VRST '05, p. 257-260. Monterey, USA. 2005.*

[J] C. Theobalt, N. Ahmed, E. de Aguiar, G. Ziegler, H. Lensch, M. Magnor, H.-P. Seidel: *Joint Motion and Reflectance Capture for Relightable 3D Video. Technical Sketch, ACM SIGGRAPH, Los Angeles, 2005.*

[K] G.Ziegler, H. Lensch, N. Ahmed, M. Magnor, H.P. Seidel: *Multi-Video Compression in Texture Space*. In Proc. of ICIP 2004, Singapore.

Appendix B

Curriculum Vitae – Lebenslauf

Curriculum Vitae

1979	Born in Karachi, Sindh, Pakistan
1988-1995	Chiniot Islamia Public School, Karachi, Pakistan
1995-1997	D.J. Science College, Karachi, Pakistan
1998-2001	B.S. in Computer Science, University of Karachi, Karachi, Pakistan
2003-2004	MSc. in Computer Science, Saarland University, Saarbrücken, Germany
2005-	Ph.D. Student at the Max-Planck-Institut für Informatik, Saarbrücken, Germany

Lebenslauf

1979	Geboren in Karachi, Sindh, Pakistan
1988-1995	Chiniot Islamia Public School, Karachi, Pakistan
1995-1997	D.J. Science College, Karachi, Pakistan
1998-2001	B.S. in Computer Science, University of Karachi, Karachi, Pakistan
2003-2004	MSc. in Computer Science, Universität des Saarlandes, Saarbrücken, Germany
2005-	Promotion am the Max-Planck-Institut für Informatik, Saarbrücken, Germany