# Representations for Shape and Motion in 4D Acquisition

Jean-Sébastien Franco

▶ **To cite this version:**

Jean-Sébastien Franco. Representations for Shape and Motion in 4D Acquisition. Computer Vision and Pattern Recognition [cs.CV]. Université Grenoble Alpes, 2021. tel-04117752

**HAL Id: tel-04117752**
**https://hal.science/tel-04117752**

Submitted on 5 Jun 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE GRENOBLE

## HABILITATION À DIRIGER DES RECHERCHES

Spécialité : **Informatique**

Présentée par

**Jean-Sébastien Franco**

# Representations for Shape and Motion in 4D Acquisition

soutenue publiquement le **10 juin 2021**,
devant le jury composé de :

**Christian Theobalt**
Professeur, Saarland University, MPI Informatik, Rapporteur
**Adrian Hilton**
Professeur, University of Surrey, Rapporteur
**Jean Ponce**
Directeur de Recherche, Inria, Rapporteur
**Cordelia Schmid**
Directrice de Recherche, Inria, Examinatrice
**Adrien Bartoli**
Professeur, Faculté de Médecine Clermont Ferrand, Examinateur
**Edmond Boyer**
Directeur de Recherche, Inria, Examinateur

# ABSTRACT

Acquiring and representing 3D shapes in motion from one or multiple views, coined 4D modeling, has been a topic of interest to computer vision and computer graphics for several decades, for 3D content production, virtual reality, telepresence and human motion analysis applications. In this time frame, the topic has gone from theoretical research subject to partial industrialisation, including nowadays many acquisition studios from major world companies, and a successful Inria startup with 15 years of activity. To achieve this, a set of hard problems had to be addressed, with advances in what can be classified as three subfields of shape acquisition and representation, motion retrieval and modeling, and appearance and texture estimation. This document summarizes my journey and contributions toward these goals with the students I was privileged to work with, with a particular focus on PhDs defended in the last five years. First I will discuss our work with multi-view stereo and monocular shape estimation. Second, I will present how weak but general priors on surface or volumetric rigidity, and using the a priori of an underlying human shape space, can be leveraged for surface tracking and alignment of human subjects in tight or loose clothing. Third, I will discuss the space-time statistical and representational models of appearance we proposed to estimate and store color map information of acquired subjects.

# RÉSUMÉ

Acquérir et représenter des formes 3D en mouvement à partir d'une ou plusieurs vues, aussi appellé modélisation 4D, a été un sujet d'intérêt pour la vision par ordinateur et l'infographie depuis plusieurs décennies, pour la production de contenu 3D, les applications liées à la réalité virtuelle, la téléprésence et à l'analyse du mouvement humain. Dans ce laps de temps, le sujet est passé de la recherche théorique à une industrialisation partielle, avec de nos jours de nombreux studios d'acquisition dans les entreprises majeures du domaine de l'informatique, et une startup Inria avec actuellement à son actif 15 ans d'activité. Pour atteindre ce palier majeur, un ensemble de problèmes difficiles a dû être résolu, avec des progrès dans ce qui peut être classé en trois champs d'investigations, l'acquisition et la représentation de forme, la modélisation et l'estimation de mouvement, et l'estimation de l'apparence et de la texture. Ce document résume mon parcours et mes contributions dans le sens ces objectifs avec le étudiants avec lesquels j'ai eu le privilège de travailler, avec un accent particulier sur les thèses soutenues ces cinq dernières années. J'aborderai tout d'abord notre travail avec

la stéréo multi-vues et l'estimation monoculaire de forme monoculaire 3D. Deuxièmement, je présenterai l'apport d'a prioris faibles mais très généraux sur la rigidité des surfaces ou en volume, et l'utilisation d'a priori d'un espace des formes humaines, et comment les mettre à profit pour le suivi de surface et l'alignement de sujets humains en vêtements amples comme serrés. Troisièmement, je discuterai des modèles statistiques spatio-temporelles et des représentation d'apparence que nous avons proposés pour estimer et stocker des cartes de couleur pour les sujets acquis.

# Table des matières

# CHAPITRE 1

# Introduction

## 1.1 RESEARCH PROBLEM : 4D MODELING

The past decades have seen an ever increasing interest in automated 3D dynamic content creation, for various applications such as 3D content production, advertizing, entertainment, virtual reality, telepresence. It is also becoming a popular research topic in the learning era, where works are increasingly showing that 3D constraints and inference can be built on deep learning methods to increase result quality and scene understanding.

This trend has been fueled by the increased availability of multi-camera systems, such as our 68-camera capture plaftorm Kinovis [kin], sometimes comprised of dozens or hundreds of cameras, that can be used for performance capture (*e.g.* [SH07, dAST+08, LDX10, CCS+15, JLT+15]). Such systems can readily produce video streams of the same subject from multiple viewpoints, allowing indirect and passive access to the 3D geometry, motion, and appearance of filmed subjects. This technology offers numerous promises with respect to rugged, but sparse previous generation motion capture techniques, where only a predefined set of points and no colorimetry was observed, and relies on instrumentation of the captured subjects with active or passive markers. Thanks to spatially and temporally dense color and sometimes depth observations now offered with multi-view platforms, a much richer set of possibilities arise toward estimation of full shapes in motion, with detailed surface reconstruction, motion tracking, and access to the appearance through the acquired images. This is all with the advantage of passive, non invasive protocols where the subject or actor can come in full clothing and expect to be captured with all his apparel, props, and ultimately interaction with other people or with the set and scenery. But with this richer data comes a vastly increased complexity, and extracting these problems is still a very active research field to this day.

**FIGURE 1.1 –** *Visualization of the geometry of a 4D aligned template model.*

This document is centered on the problem of 4D modeling, which is the process of producing this 3D content and appearance representations from temporal sequences obtained from a set of cameras. In most works in this topic, access to a controlled setup with pre-calibrated cameras is assumed, such that the focus is on obtaining the 3D model geometry, its motion and its appearance as texture map as opposed to egenral scene structure. 4D modeling differs from simple, one shot reconstructions over each frame of the sequence, by the main property that it exploits time continuity and redundancy to enhance the quality of the models, or produces models that are aligned, e.g. with identical surface topology and connectivity but deforming shape or, by means of varying model parameters or vertex positions.

## 1.2 CHALLENGES

Our general goal is to produce highest quality possible and easy to use models, and produce the best 4D modeling algorithms to this goal. This is a generally very challenging topic because, in its most general form, we are observing a moving shape or set of moving shapes with very indirect raw data, namely sets of image sequences $\mathrm{I} = \{\mathrm{I}_i^t\}_{i\in\{1,\cdots,n\}}^{t\in\{1,\cdots m\}}$ obtained from a set of calibrated $n$ cameras during $m$ temporal frames, and how to extract the shape, motion and appearance of these single underlying objects, in some sense a large scale multi-variate and interdependent regression problem. Some of the main obstacles to achieve these goals are the following :

***Dimensionality***

The key information is buried in several high frequency video streams that quickly comprise Terabytes of data. On a typical platform such as Kinovis, acquisitions produce data at a rate of 40Gb/s.

The dimensionality of the output is typically also very large, even if it exploits and removes key input redundancies, but still comprises large representations, *e.g.* mesh and textures, with full or partial updates that need to be periodically refreshed.

#### *Representation*

Finding the best intermediate and output representations for the 4D model is still to this day an open problem. Exactly what granularity shall be stored for the geometry, temporal evolution and appearance can lead to compact results, but the desirable properties of the representations may not be the same at different stages of the pipeline. Should we use surface meshes, or volumetric primitives for the inference ? How exactly should the color information be stored, as a separate per-frame texture map, or using a temporally refactored representation ? What is the best way of co-storing shape, geometry, motion and appearance ? These are some of the questions examined in this document.

#### *Choosing Priors*

The image set we observed is a noisy, partial observation of the quantities we wish to estimate, and our problem is an inverse and ill-posed problem. Choosing the correct priors and how to embed them in our method is going to be important and in fact we have explored many different directions.

First the assumptions inherent to the model itself involve relying on regularizing behaviors, and usually in their simplest form involve geometric, motion and appearance continuity over the full shape. But choosing the correct prior is non-trivial. Typically building the prior that the motion is from a human subject is a key tradeoff we examine in this work, on one hand dealing with very basic motion hypothesis or at the opposite of the spectrum having a strong constraining human model. One will favor generality and the other will be more stable but may not account for out-of-model situations such as object or multi-person interaction or loose clothing.

## 1.3 STRUCTURE AND CONTRIBUTIONS

As the previous challenges hint at, the 4D modeling problem requires examining a collection of subproblems together to target a single goal. To break the complexity of the algorithm, some kind of stratification is usually chosen. Throughout our works during the years, the stratification that we opted for has been relatively uniform, first extracting the shape at individual time frames, then performing motion analysis for tracking and alignment on the individual reconstructed shapes directly in 3D, and finally extracting appearance information. This is of course not the only possible stratification, but to this day remains a very popular way to break down the problem, with variants in how geometric and temporal tasks are connected, *e.g.* [CCS+15, MVK+20].

This manuscript follows these three main stages in their logical order of application in the pipeline, which in particular is quite different than the chronological order in which these works were actually executed.

In Chapter 2, I describe how we approached the problem of 3D shape estimation with one frame or small group of frames simultaneously :

— I first present two of our contributions toward designing a customized multi-view stereo pipeline particularly suitable for the context of performance capture and 4D modeling tasks. A careful analysis of our needs and acquisition scenarios lead us to build-in key features and exploit local temporal redundancy to increase precision ; we also examine how deep learning can be used to optimize the feature extraction and reconstruction decisions.

— Second, I present a method whereby we used an end-to-end deep learning to extract a full 3D from one viewpoint, with the idea of testing the limits of the monocular reconstruction setup, and as preliminary step to possible multi-camera applications as well.

In Chapter 3, I present several works focused on temporal analysis and alignment of the 3D models previously extracted

— I first summarize the works achieved on a series of models based on weak but quite general, quasi-rigid patch-based priors, using adaptive versions of patch-based constraints either at the surface or the volumetric level.

— Then I examine how a stronger human shape-space model can be used to constrain the estimation of shapes in motion, and estimate the underlying body shape of acquired subject under clothing.

In Chapter 4, the discussion is on two appearance estimation works we did to retrieve an appearance map measuring dense radiance information observed :

— First, we present a principled method by which the appearance estimation of a 3D model under small motions can be treated as a specialized 3D version of the well-known video superresolution problem

— Second, I discuss how this work can be extended to longer sequences by storing long term appearance variations as a linear combination of 3D mapped Eigen textures.

Before getting to the technical part of the presentation, I will describe some contextual information, in particular about the people and contexts I worked with.

## 1.4 PEOPLE

This manuscript is intended to give an overview of my research work as an associate professor (”*maître de conférences*”) in computer science at Grenoble INP - Ensimag (part of the Université Grenoble Alpes, France) with a focus on PhDs defended

in the last 5 years. During my career, I had the pleasure and privilege to work with a wide range of colleagues and students on the aforementioned topics. I will give a brief overview of these encounters in the following paragraphs.

### 1.4.1 Contexts

I obtained my PhD from the Institut National Polytechnique de Grenoble in 2005 with the INRIA MOVI team headed by Radu Horaud, supervized by Edmond Boyer.

I then joined UNC Chapel Hill as a post-doctoral research assistant to Marc Pollefeys in 2006-2007, during which I co-supervized PhD students Li Guan and collaborated with Sudipta Sinha on probabilistic visual hulls.

In 2007, I was appointed to my first associate professor position at Université Bordeaux 1 with Pascal Guitton and the IPARLA team, Inria Bordeaux Sud-Ouest. I co-supervized the PhDs of Yann Savoye and Robin Skowronski with Pascal Guitton, working respectively on non-rigid human tracking and calibration for drones. I also collaborated in the supervision of with PhD student Benjamin Petit with whom I interacted a lot with for the Dalia ANR project on telepresence between two multi-camera platforms Hemicyclia and GrImage.

I then moved to the Ensimag (School of Computer Science and Applied Mathematics, INP Grenoble University) as associate professor of Computer Science, and as a researcher at the Inria Grenoble Rhône-Alpes, France, with the Morpheo team in 2011 and have been co-supervising many PhD students since then, Abdelaziz Djelouah on multi-view segmentation with Patrick Perez and Francois LeClerc at Technicolor, Vagia Tsiminaki and Adnane Boukhayma on multi-view appearance estimation and superresolution, Benjamin Allain on shape tracking and temporal surface alignement, Vincent Leroy on probabilistic Multi-View Stereo, with Edmond Boyer ; and I have had a non-official supervizing role throughout Jinlong Yang's PhD on model-based human body tracking under clothing, a collaboration with Morpheo team members Stefanie Wuhrer and Franck Hetroy.

Out of the master students I supervized too numerous to enumerate here, one stands out recently as it opened a collaboration cycle with Inria Thoth team members Gregory Rogez and Cordelia Schmid, through the supervision of Valentin Gabeur. This collaboration would yield a major conference paper at ICCV 2019 and paved the way to several Morpheo-Naver Labs Europe collaborations after Gregory Rogez and Vincent Leroy joined Naver.

I am currently co-supervizing Mathieu Armando on mesh and appearance superresolution, Boyao Zhou on using deep learning for human space-time human shape reconstruction, with Federica Bogo and Edmond Boyer, as part of a Microsoft Grant ; I am also co-supervizing Abdullah Haroon Rasheed on learning based cloth physics estimation with Florence Bertails, Stefanie Wuhrer, and Mathieu Marsot on learning space-time human shape and motion generative models, with Stefanie Wuhrer, and Anne-Hélène Olivier from Inria Rennes who is a ANR project collaborator. These have all been exciting projects that are followups of the research projects described in this document, and have benefited from these previous experiences.

### 1.4.2 Student Supervision

## PhD students

*Current PhD students* :

— Mathieu Marsot. **ANR 3D MOVE**, starting October 2019. Learning generative models for 3D human motion. 50% supervision. Co-supervised with Stefanie Wuhrerm, Anne-Hélène Olivier.
— Boyao Zhou. **MS Research PhD grant**, starting Decembre 2018. Motion and trajectory analysis in an interactive environment. 30% supervision. Co-supervised with Edmond Boyer, Marc Pollefeys, Federica Bogo, Bugra Tekin.
— Abdullah-Haroon Rasheed. **ERC Grant Funding**. Since November 2017. 3D Deep Learning For Inference of Cloth Physical Parameters. Co-supervised with Florence Bertails, Stefanie Wuhrer. 30% supervision.
— Mathieu Armando. **MS Research PhD grant**, starting November 2017. Mesh and Texture superresolution. 50% supervision. Co-supervised with team leader Edmond Boyer.

*PhD students already defended* :

— Vincent Leroy. **ANR Achmov Project Grant**, starting October 2015, defended October 17th 2019. 50% supervision. 4D stereo surface extraction. Co-supervised with team leader Edmond Boyer. Publication in ICCV 2017, ECCV 2018. Now research scientist at Naver Labs Europe.
— Benjamin Allain. **RE@CT FP7 European Project Grant**, starting October 2012, defended March 31st, 2017. *Volumetric Tracking of 3D Deformable Shapes*. 60% supervision. Co-supervised with team leader Edmond Boyer. Publications in ECCV 2014, CVPR 2015, CVPR 2016. Now research scientist with startup Smart Me Up.
— Vagia Tsiminaki. **RE@CT FP7 European Project Grant**, started June 2012, defended December 14th, 2016. *3D model appearance extraction from multiple-view sequences*. 60% supervision. Co-supervised with team leader Edmond Boyer. Publications in CVPR 2014, ECCV 2016. Now post-doc at ETH Zürich with Marc Pollefeys.
— Abdelaziz Djelouah. **CIFRE Doctoral Grant**, industrial collaboration and co-funding with Technicolor, Rennes, started April 2011, defended March 17th, 2015. *Multi-view object segmentation with calibrated camera networks*. Co-supervised with Edmond Boyer, François Le Clerc, Patrick Perez. Publications in ECCV 2012, ICCV 2013, PAMI 2015. 40% supervision. Formerly Post-Doc at Inria Sophia Antipolis Méditerranée with George Drettakis, now research scientist with Disney Research, Zürich.

— Yann Savoye. **Ministry Doctoral Grant MNERT**. *Dynamic Reconstruction of Human Shape and Motion*. Since October 1st, 2008, mesh tracking for shape modeling, 3D interaction and telepresence. 90% supervision, co-supervised with team leader Pascal Guitton. Doctoral School EDMI Bordeaux. Now lecturer at Liverpool John Moores University, UK.
— Robin Skowronski. **CIFRE Doctoral Grant**, industrial collaboration and co-funding with start-up Aérodrones, France. *Environment Perception and Application to Lightweight UAVs*, from March 1st, 2008, defended on November 3rd, 2011, computer vision and calibration from a UAV-mounted camera. 90% supervision, co-supervised with team leader Pascal Guitton. EDMI Bordeaux. Currently full-time R&D engineer with Aérodrones.
— Li Guan. **PhD at UNC Chapel Hill**. *Multi-view Dynamic Scene Modeling*. 90% supervision. Co-supervised with Marc Pollefeys from February 2006, defended on August 14th, 2009. Occlusion-robust shape modeling from multiple silhouettes, motion analysis. Publications in 3DPVT 2006, 2008, CVPR 2007, 2008, 2010. Formerly Research Scientist (Computer Vision) with Amazon, now with Zillow.

*Significant participation in the following PhD supervision and collaborations* :

— Jinlong Yang. **ANR Achmov Project Grant**, started October 2015, defended March 28th 2019. Learning shape space of dressed 3D human models in motion. Collaboration with supervisers Stefanie Wuhrer and Franck Hetroy, estimated 30% supervision. ECCV 2016, ECCV 2018. Now with Facebook at Occulus Research.
— Adnane Boukhayma, started October 2014, defended 6th December 2018. *Surface motion capture animation*. Collaboration with superviser Edmond Boyer. Estimated 20% supervision. ECCV ECCV 2016, CVPR 2017. Now Post-Doc at University of Oxford, UK, with Phil Torr.
— Benjamin Petit, 2007-2011, *Telepresence, immersion and interactions for real-time 3D reconstruction*, collaboration with supervisers Edmond Boyer and Bruno Raffin. Publications at VMV 2011 & 2013, IJDMB journal 2010, Multimedia 2010, 3DTV 2009, VRST 2008. Now general manager at Beam'Art. Estimated 20% supervision.

## Master students

— Briac Toussaint. March to September, 2021. Multi-view calibration for the Kinovis Platform.
— Jiabin Chen. March to September, 2019. 4D Motion Analysis. Co-supervized with Stefanie Wuhrer.
— Valentin Gabeur. February to September, 2018. Inference of 3D models from monocular image input. Co-supervized with Grégory Rogez and Cordelia Schmid.

— Ivan Iudintsev. February to June, 2018. A comparison of NN-generative models for 3D human shapes. Co-supervized with Stefanie Wuhrer and Abdullah-Haroon Rasheed.
— Abdullah-Haroon Rasheed. March to September, 2017. 3D Deep Learning For Shape from Silhouettes in a Multi-Camera Setup.
— Pau De-Jorge Aranda. March to August, 2017. Deep Neural Networks for 3D body shape and pose prediction in real images. Co-supervized with Gregory Rogez and Cordelia Schmidt.
— Hamza Jaffali, March to August, 2017. Optimization for Cloth Motion Inversion. Co-supervized with Florence Bertails.
— Yannick Marion. March to August, 2015. Shape space model for multi-camera 3D tracking analysis. Co-supervised with Stephanie Wuhrer.
— Antoine Fond. Centrale Nantes Masters. April 1st to September 30th 2014. Spatio-temporal shape and point trajectory analysis. Co-supervised with Franck Hetroy. Now PhD candidate at Inria Nancy Grand Est with Marie-Odile Berger.
— Lienhoa Nguyen. Mosig Masters, Grenoble Universities. February to June 2014. Optimal polyhedral boolean CSG. Co-supervised with Matthijs Douze and Bruno Raffin.
— Renato Oliveira. Mosig Masters, Grenoble Universities. February to June 2012. 3D shape tracking and refinement from multiple views. Now Software Developer at Hewlett-Packard.
— Hassan Kourad, Master Signal, Image, Parole et Telecom (SIPT), Grenoble Universities. EM algorithm parallelization on GPU. March to September 2012. Co-supervised with Dominique Houzet and Vincent Fristot (GIPSA lab).
— Olivier Augereau. ENSEIRB Bordeaux school of engineering Master. *A new Multitouch Interface for 3D Interaction*. February to June 2009. Hybrid multitouch and multi-camera systems. Co-supervised with Martin Hachet. Now research assistant professor at Osaka Prefecture University.
— Arash Kian. Bordeaux 1 University Master. *Opportunistic Music Control*, from February to June 2009. Computer stereovision and gesture detection for musical interaction. Co-supervised with Martin Hachet. Publication at JVRC 2009.
— Elric Delord. Bordeaux 1 University Master. *3D Immersive Interaction by Temporal Reconstruction of Human Motion*. February to June 2008. Bayesian modeling of scene flow in a multi-view sequence.
— Steven Gay. *Occupancy Grids and Silhouette fusion on a GPU*. June 18th to August 18th, 2005. Co-supervised with Edmond Boyer.

### 1.4.3 Main projects and collaborations

— **Participant, ANR-3DMOVE project**, JCJC from Stefanie Wuhrer, since October 2019. Generative models for 3D human Motion. **Supervision of a doctoral student**, Mathieu Marsot, starting in 2019.
— **Participant, Microsoft-Inria Joint-lab project**, since 2017. Learning human motion, surface geometry and appearance details. **Co-supervision of two PhD**

**students**, Matthieu Armando and Boyao Zhou, starting in October 2017 and January 2019 respectively.

— **Participant, ANR-Achmov project**, since November 2015. Shape space inference and tracking, generative-discrimative approaches, in collaboration with IMAR, Romania. **Supervision of a doctoral student**, Vincent Leroy, starting in 2015.

— **Participant and scientific coordinator for Morpheo team, for the RE@CT European project** (FP7-ICT-2011.1.5), since September 2012. 4D performance Capture, temporal sequence alignement and analysis, appearance and semantic acquisition. Rigid structure detection. Collaboration with OMG/Vicon, BBC, HHI, Artefacto, University of Surrey. **Supervision of two doctoral students**, Vagia Tsiminaki and Benjamin Allain, starting in 2012.

— **Participant** in **Equipex Kinovis** (French national *equipment of excellence* grant), since 2012. Participation in scientific panel and elaboration of the plateform, dedicated to the high definition acquisition of subjects in motion (68 4Megapixels cameras, 100m2 studio, 17 twelve-core PC cluster).

— **Participant and scientific coordinator for IPARLA team, for the ANR-DALIA project** (French Ministry of Research Project Grant), from September 1st, 2007 to July 2010. 50% participation. Telepresence and collaborative 3D interaction, in collaboration with MOAIS and Perception teams (INRIA Rhône-Alpes, France) and PRV (LIFO lab, Orléans, France). **Supervision of a research engineer** Benoît Bossavit, and co-administration of the multimedia platform Hémicyclia (10-node cluster and multi-camera platform) at LaBRI lab, Bordeaux, France.

— **ANR-SeARCH** project, 10% participation, from 2009 to 2010. Acquisition, modélisation et assemblage semi-automatique de modèles archéologiques des pièces du phare d'Alexandrie. Collaboration with Ausonius, and the Center of Alexandrine Studies.

— **ANR-InSTINCT**, 20% participation, from 2009 to 2010. 3D interaction using a hybrid system, computer vision platform and multitouch surfaces.

— Punctual participation as doctoral student various technologic transfer projects (2003-2006) : RNTL OCETRE, FP6-IST STREP HOLONICS (european commission), ACI Jeune Chercheur Cyber I and ACI *Masse de Données* Cyber II, (mixed-reality projects) at INRIA Rhône-Alpes, France.

### 1.4.4 Manuscript Focus

It will be apparent to the reader at this point that this document is in no way meant as an exhaustive review of works and student projects. The fact that I left them out of this manuscript says nothing about those works, and everything about having to keep this document contained in the interest of time and by focusing on key emblematic stages and developments we had with the understanding of the problems in the last five years.

I have in particular left out the excellent work of Li Guan [GSFP06, GFP07, GFP08a, GFP08b, GFBP10, GFP10], the first PhD student I co-supervized with Marc Pollefeys during my post-doc, of Abdelaziz Djelouah [DFB+12, DFB+13, DFB+15, DFB+16], which have in common the fact that they are focused on earlier silhouette-based extraction efforts.

I have also taken the stance of leaving out works that were not defended yet as PhDs, leaving this discussion at the legacy stage at the end of this work, this includes the excellent work of Mathieu Armando [AFB19, AFB20], Boyao Zhou [ZFB+20], and Abdullah Haroon Rasheed (CVPR 2020 oral paper [RRBD+20]).

I did include Valentin Gabeur's work on monocular reconstruction with Deep Learning [GFM+19], defended as master student, since it so well represents the Deep Learning leap and transition we took these last years, in continuity with the work of Vincent Leroy.

# CHAPITRE 2

# 3D Shape Estimation

## 2.1 INTRODUCTION

Among the first problems we examine in this document, and one of the first problems to solve in stratified 4D Modeling, is that of 3D shape estimation from images. This is in fact one of the fundamental problems of computer vision and has often been treated in generic form in the geometric era of computer vision [HZ00]. Using a pre-configured platform camera rig, we have been naturally strongly inclined to pursue research in methods that assume calibration is available. This was notably the case during my thesis and post-doc years where I pursued research in silhouette-based methods [FB03, FMBR04, FB05, FB09, GFP08a, GFBP10].

While this push to enhance silhouette-based methods yields quite useful results and has been the basis of many other works discussed in the other chapters, during the last five years, we also pursued research on enhancing 3D modeling, first in the direction of significantly improving model quality given the abundant input data acquired through our 68-camera Kinovis platform, and second gaining better understanding of how deep learning can contribute to improving 3D modeling methods, in particular toward any of the common objectives of increasing model quality, precision, robustness to input corruption and lack of texture detail routinely encountered in everyday, casual clothing. Dealing with fast motions has also been a desirable target due to the presence of motion blur when subjects execute fast movements typical to dancing or sports. This drive was to open new research avenues in the team, as they have proven largely successful.

To illustrate the push in these directions, I will discuss in this chapter the two most relevant efforts we pursued, first toward novel improving Multi-View Stereo in the context of performance capture applications, and second toward the problem of monocular 3D shape estimation, where Deep Learning has proven to be instrumental for both.

## 2.2 MULTI-VIEW STEREO EFFORTS

Innovation on the quality of the shape result was a major drive throughout the last decade in the team ; and multi-view stereo had the potential to offer these benefits. But multi-view stereo is both a widely studied and notoriously difficult subject to work with on a technical level, which has stumped a number of students we attempted to tackle this problem with. A testament to this is the endurance of several widely respected benchmarks that have survived as beacon references to the community for the better part of the last decades, such as the Middlebury, the DTU and more recently tanks and temples datasets [SCD+06, SvHG+08, JDV+14, SSG+17, KPZK17], which all have registered small-step but continuous improvements in the domain over large time frames.

Also, the technical variety of the frameworks and representations devised to tackle this problem is quite astounding spanning several decades, from level sets [FK98], voxel carving [KS00], depth map sweeping [GFM+07], Delaunay decomposition and graph cut optimization [LPK07], sparse points from image features [FP10] spatio-temporal integration [GM04, MKGH16], model-based integration [SH07, dAST+08], convex grid optimization [CK11] or probabilistic inference [UGB15], to name only a few.

It thus takes a special type of character to confront this problem, with perseverance, heterogeneous literature comprehension, and large programming and technical proficiency, which we found with PhD student Vincent Leroy. In this journey, we implemented two state of the art MVS pipelines with a classic plane-sweeping articulation but several key practical improvements, then substituted the photoconsistency core with a Deep Learning replacement that considered local photoconsistency volumes to compute a depth indicator function.

### 2.2.1 DAISY-Based Plane Sweep Stereo Pipeline with Local Temporal Integration

Our first effort in this direction was to implement a classic pipeline based on what stood out in the literature as desirable characteristics for our capture scenarios. First and foremost, we wanted to be able to enhance the quality of the surface thanks to temporal smoothing and refinement, which has been a longstanding goal [GM04, APSK07, MKGH16] seldom achieved for centrimetric detail in performance capture scenarios. Of particular inspiration both for the remarkable detail accumulation in dynamic capture environments and for the popularization of Truncated Signed Distance Functions (TSDF) as depth map volumetric fusion representation, were methods of the Kinect and Dynamic fusion family [NIH+11, NFS15, IZN+16, DKD+16], a key difference with our goals being that they process RGB-D streams instead of multi-RGB inputs.

Second, depth-map plane sweep methods have been notable for one very desirable feature : their spatially dense, per-pixel monotonous depth parameterization of visible surface geometry for every viewpoint, with first-point visibility built-in the parameterization [Col96, MAW+07]. This yields very efficient depth map extraction methods

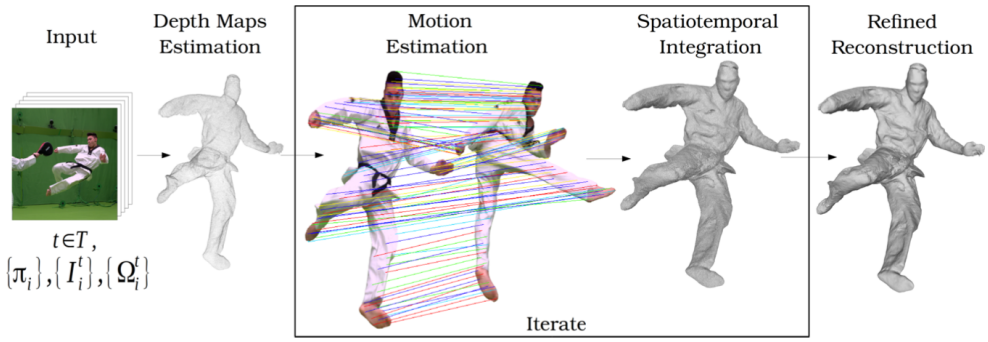


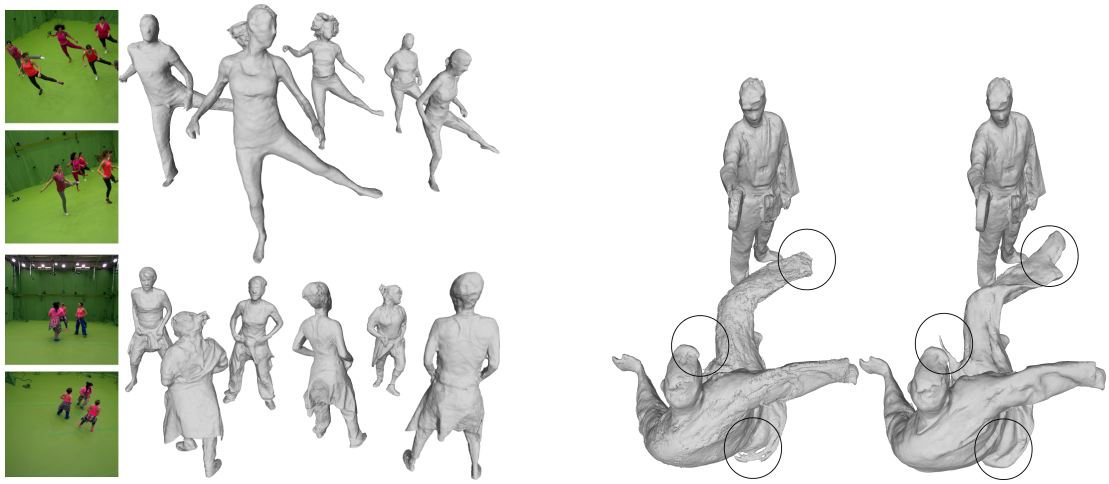

**FIGURE 2.1** – *Spatiotemporal refinement framework.*

which are cast as extracting the first visible point at each pixel, gracefully dealing with the visibility problem. Third, among the various classic photoconsistency characterizations based on normalized window correlation (*e.g.* ZNCC, SSD, SHD, etc) or gradient feature correlation [Low04, BTG06, MS03], DAISY features [TLF10] was the state of the art at the time for dense, wide-baseline stereo and computational efficiency, and as such was a natural contender for our photoconsistency function and evaluation.

### *Method and Contribution*

We proposed a pipeline illustrated in Fig. 2.1 implementing together these key characteristics, assuming images $\mathrm{I}_i$, silhouettes $\Omega_i$, and calibration matrices $\pi_i$ are given.

- First we compute per-view depth maps based on line searches that maximize a DAISY-based photoconsistency criterion, filtering correct matches using a visual hull of the subject obtained from the silhouettes
- Second we provide an initial estimate of a multi-view merged shape based on the fused multiple depth maps obtained at a single frame. Because of the detail density, classic TSDF on regular grids is particularly inefficient and memory heavy, so we proposed an implicit TSDF form which can be computed with sparse storage.
- Third, using this initial per-frame shape estimate for a time-window of neighboring frames, we construct sparse surface feature matches to neighboring frame shapes using MeshHOG features [ZBH12], which we densify by propagating them on the surface.
- We use these matches to fold the shape contributions to the center frame in each temporal window using a simple locally rigid deformation model.
- These folded shapes bring new sparse implicit TSDF constraints to the central frame, which can then be seamlessly integrated. The process is then iterated between alternate steps of matching and reconstructing, leading to a refined final shape.
- The final surface can be extracted from the implicit TSDF using Centroid Voronoi Tesselations [DFG99, LWL+09] of a carefully selected set of points samples in the vicinity of the final surface as hinted by the TSDF induced occupancy



(a) (b)

**FIGURE 2.2 –** *Obtained spatio-temporally refined results. (a) Detailed shapes obtained for complex, dynamic, multi-person performance captures scenes. (b) Comparison of the method without, and with the temporal integration proposed.*

function, and clipping the CVT tesselation to the zero level set of the implicit TSDF function to extract surface polygon geometry, as illustrated in Fig. 2.2

Our results showed that the quality of the reconstructions, measured with the widely adopted Chamfer-based metrics of surface accuracy and completeness [SCD+06, JDV+14], was significantly improved using the various components of our proposed approach. The approach also demonstrated measurable improvements with respect to state of the art methods in the context of performance capture. Details of the method and results can be found in the ICCV 2017 publication [LFB17].

### 2.2.2 Volume Sweeping : Learned Photoconsistency

During the time frame of Vincent Leroy's thesis, it became quite evident that the benefits of Deep Learning could reach beyond 2D vision problems which were the main focus in the early DNN years vision [KSH12]. As my personal and our team expertise in the subject was initially limited, our approach has been largely progressive and at first geared toward testing the contribution of learning techniques in well identified stages of time-honored vision pipelines.

In this respect, the MVS work with Vincent was an ideal candidate. As illustrated in Fig. 2.3 showing how we modified the classic MVS pipeline, the photoconsistency function $\rho$ used to characterize surface hits for every depth candidate along each pixel's viewing line surface depths has a well identified local support in the images at the projection of the candidate, which can be described with a small size network. The surface hit decision is typically a per pixel decision in plane sweep algorithms with optional image-domain regularization of the depth maps based on local per-pixel depth data terms. The photoconsistency function was thus ripe for this kind of trial as it also clas-

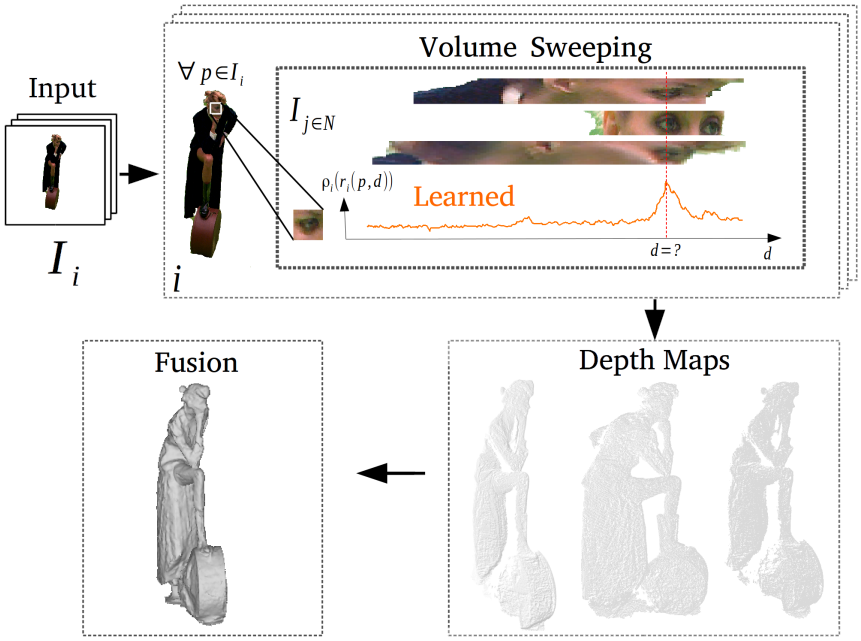



**FIGURE 2.3 –** *Volume sweeping method overview. Depth maps, for all input image $I_i$, are obtained by maximizing, along viewing lines, a learned function $\rho$ that measures photoconsistency at a given depth $d$ along the viewing line of a given pixel $p$. Depth maps are then fused into an implicit form from which the zero set surface is extracted.*

sically relies on a hand-crafted measure and empirically selected image features. This is typically where deep learning can contribute by automatically extracting relevant features from example data.

In fact at the time several works had started pushing in the direction of short-baseline MVS with symmetric combination of 2D learned features [HGH+17], or wide baseline sparse capture scenarios [HLC+18, GVCH18] that followed learning research on short baseline stereo [vL16, LSU16, ZK15, UZU+17].

### *Method and Contribution*

While some learning methods for MVS were starting to emerge, at the time little or none of them were particularly geared or tested for the specifics of the multi-camera performance capture scenario, where in particular the attention is on moving subjects that cover only a minority fraction of all frame pixels, which we informally referred to as the "mid-range scenario". This use case is in particular very different to the case where objects are fully framed in every view, for close static objects or full-views of architectural scenes which are typical used as benchmark cases [SCD+06, JDV+14, SvHG+08].

Our feeling as well was that, in our case with known calibration, we would be under-using the input data if we only used 2D patches to characterize matches in the learned photoconsistency function, as several methods were proposing [vL16, LSU16,

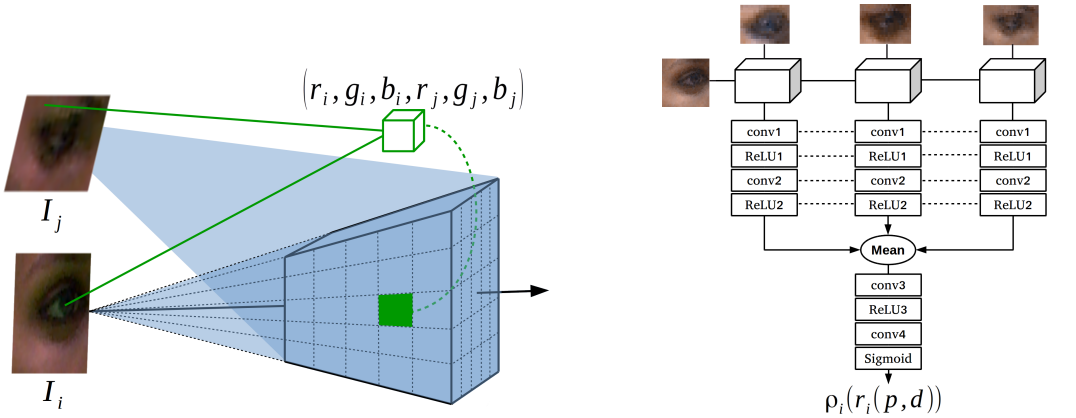



**FIGURE 2.4 –** *(left) The 3D volume used to estimate photoconsistency along rays from the reference image $I_i$. $k^3$ samples within the volume are regularly distributed along viewing rays and contain color pairs as back-projected from images $I_i$ and $I_j$. At a given depth along a ray from $I_i$ any image $I_{j\neq i}$ can define such a pairwise comparison volume. (right) CNN architecture. Each cube is a pairwise comparison volume with $k^3$ samples that contain 6 valued vectors of RGB pairs and over which 3D convolutions are applied. The output score $\rho_i(r_i(p,d)) \in [0..1]$ encodes the photoconsistency measure at depth $d$ along the ray from pixel $p$ in image $I_i$.*

ZK15, UZU+17]. Several methods were also starting to propose 3D inference in the volume for these kind of reasons [CXG+16, JGZ+17, KHM17]. Reasoning only on 2D patches deprives the network from information on relative orientation of views and the local geometric context that can be additionally useful to build a decision. This intuition turned out to be correct and verified in our results, where we compared 2D versus 3D volumetric support of the learning function.

To these goals, we proposed to formulate the inference on a small projective volume surrounding each query point of interest along a viewing line. Each voxel of this volume is assigned two R,G,B triplets from the reference view whose depth map we are computing, and another view used to check photoconsistency. The support region of the network, and its architecture is illustrated in Fig. 2.4. The architecture is intentionally simple and classically inspired [KSH12], with the main characteristic that the contribution of all pairs of views needs to be merged using a symmetric function computing a result independent of the number of views considered, with the mean giving the best empirical results in practice. The network was trained on a large set of samples built from the standard DTU dataset, which for this compact network can provide thousands of training samples per set of multi-view input frames given the local support of the learned function; we also notably used no view from our Kinovis acquisition platform, a testament to the generalization ability of the method.

We obtained impressive results that, quite frankly, exceeded our expectations by a large margin. As Fig. 2.5 illustrates, the method was able to infer more details and exhibited less errors and more completeness even in color-ambiguous, partially reflective regions (*e.g.* thigh, boots) than the classic version of our pipeline. More unexpected, the

**FIGURE 2.5 –** *Challenging scene captured with a passive RGB multi-camera setup [kin]. (left) one input image, (center) reconstructions obtained with classical 2D features [LFB17], (right) proposed solution. Our results validate the key improvement of a CNN-learned disparity to MVS for performance capture scenarios. Results particularly improve in noisy, very low contrast and low textured regions such as the arm, the leg or even the black skirt folds.*

method recovered fold details of the black dress in portions of the image with almost zero contrast. The results were shown to outperform several off-the-shelf available methods, including classic methods [FP10, TSF12, CVHC08] and recently published learning based methods [YLL+18, JGZ+17]. Interestingly, the method also recovers results of quality comparable to [CCS+15] who use two modalities RGB and infrared pattern projections, despite the fact that we were using only the RGB inputs of their datasets for our comparison.

Details are available in the ECCV 2018 publication [LFB18] but the more thorough comparisons and comments are published in IJCV 2021 [LFB21], which is the one I am providing in the appendix A.8. We refer to the supplemental video for more results [1]

## 2.3 MOULDING HUMANS : LEARNING 3D SHAPE ESTIMATION FROM A SINGLE IMAGE

The success of deep learning on a multi-view stereo pipeline naturally brought us to another research question : can we transfer these benefits to the less data rich case of monocular 3D shape estimation ?

The question of how much can be recovered from a single view has intrigued researchers for quite a while. From a purely geometric standpoint, the fundamental

1. https://hal.archives-ouvertes.fr/hal-01567758/file/1361-supp.mp4

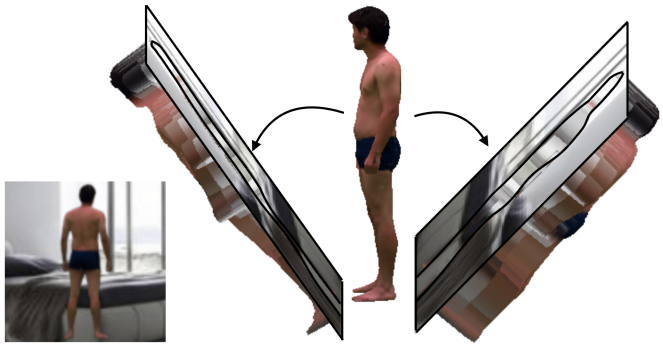


**FIGURE 2.6 –** *Our non-parametric representation for human 3D shape : given a single image, we estimate the "visible" and the "hidden" depth maps from the camera point of view. The two depth maps can be seen as the two halves of a virtual "mould". We show this representation for one of the images of our new dataset.*

projection ambiguity prohibits recovering absolute depths and scales in that situation [HZ00]. It has been however identified quite early that data-driven priors may provide the possibility of finding a solution, *e.g.* for the problem of recovering a human 3D [AT06]. In fact, estimating 3D poses is one of the key problems by which deep learning was thrown at monocular 3D estimation [CWL+16, RWS19] along with denser monocular 3D estimation applications such as dense surface correspondance [WHC+16, GNK18].

[RWS19] happens to be a Thoth team contribution down our corridor at Inria Rhone-Alpes with Gregory Rogez and Cordelia Schmid. The Thoth team was getting more and more interested in addressing 3D problems with human subjects, as evidenced by various other collaborative publications to estimate 3D shapes from images [VRM+17, VCR+18]. So we came to join forces with Gregory and Cordelia and co-supervized a master student together, Valentin Gabeur.

One of the issues with using DNNs for human 3D data is that straightforwardly extending CNNs from the 2D to the 3D domain, while relying on well defined extensions of the usual CNNs convolution stages in the 3D domain such as with [VCR+18], are subject to the curse of dimensionality as this creates a large memory and computational burden in the training process. This limits the amount of training data which can be processed in a single batch and also means a large number of training parameters, with potential for overfitting.

For this reason, we set our main quest as one of finding a lean representation that would allow us to retrieve human 3D models of comparable or better quality at a fraction of the cost. In fact the lower parameter dimensionality would probably help us achieve the quality goal. The idea emerged in the discussions of performing depth inference in the image domain, but to output a surface, we need to regress not one but two depth maps, one for the front view and one for the back view of the human subject, a representation we informally coined a "mould", for the reason apparent in Fig. 2.6.

Both depth maps assembled together create a point cloud which can yield a surface reconstruction using standard techniques [KH13]. Although technically only having two depth maps encoding doesn't cover all possible situations (such as self-occlusion which leads to four or more surface crossing and depth values along a given viewing ray), this first dual-depth map encoding, with the advantage of a fixed output inference representation, provided surprisingly good results already. The inference can then be expressed in terms of an image-wide regression task, for which we proposed a stacked hourglass architecture [NYD16], previously shown successful for human pose regression - but not full surfaces. While generally successful, we did encounter one significant issue in that the succession of reduced latent space at the bottlenecks of the hourglass networks were not sufficiently regularizing for the network to fully encode the "humanness" prior underlying the training set, yielding humans with additional limbs or no limbs for example. For this reason, we additionally introduced a discriminator network, trained in adversarial fashion, to allow this to happen. This proposed pipeline is illustrated in Fig. 2.7.

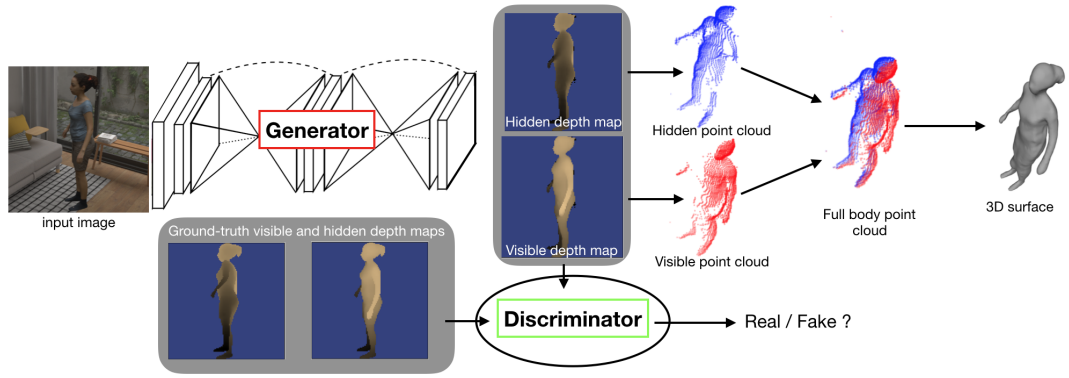
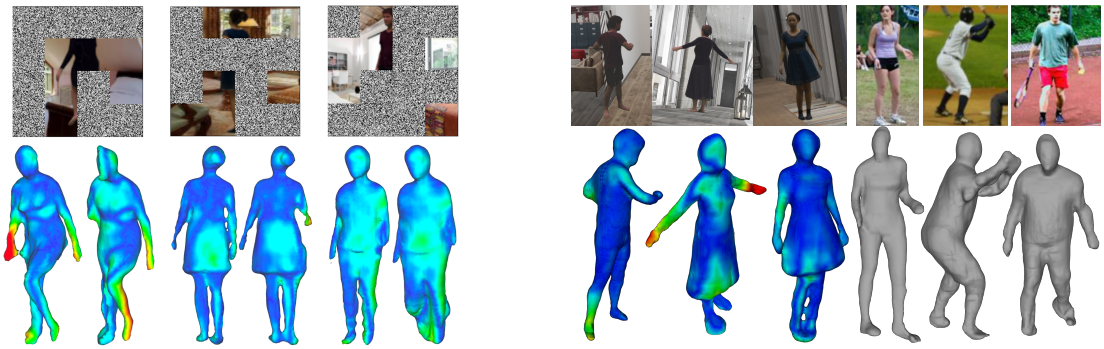



**FIGURE 2.7 –** *Overview. Given a single image, we estimate the "visible" and the "hidden" depth maps. The 3D point clouds of these 2 depth maps are combined to form a full-body 3D point cloud, as if lining up the 2 halves of a "mould". The 3D shape is then reconstructed using Poisson reconstruction [KH13]. An adversarial training with a discriminator is employed to increase the humanness of the estimation.*

We devised a training dataset, 3D HUMANS, with a large synthetic portion from [VRM+17], complemented by real captures from the Kinovis platform to capture some clothing and shape variability absent from the latter. We obtained very good to excellent results, improving over state of the art approaches, with shorter training and inference times. This includes on test datasets that had nothing to do with the training such as the DeepFashion dataset, with images of women in completely different clothing or postures, as evidenced in Fig. 2.9, Fig. 2.8. The contribution of the GAN can also be appreciated in these figures under severe input frame occlusion.



(a) (b)

**FIGURE 2.8 –** *(a) Performance on 3D-HUMANS dataset in presence of severe occlusions on three frames : (top) input images, (left) with GAN, (right) without GAN. Errors above 15cm are shown in red. The GAN helps increase the "humanness" of the predictions. (b) Generalisation to previously unobserved data. We apply our pipeline to images with 3D realistically rendered backgrounds (left), and with 3 real-world images from the LSP dataset (right). These poses, in particular the baseball player, have not been seen at training time but our model still generalizes well.*

More details and results can be found in the attached paper A.7 (Appendix A.7), which was published at ICCV 2019 [GFM+19] with the master student, and the supplemental video [2].

2. https://hal.inria.fr/hal-02242795/file/Moulding-Humans-ICCV2019.mp4

**FIGURE 2.9 –** *Comparison between HMR [KBJM18] (left), Bodynet [VCR+18] (middle) and our method (right). Unlike [KBJM18, VCR+18], we do not train on in-the-wild images but estimate 3D shapes of clothed subjects.*

# CHAPITRE 3

# 3D Motion Estimation

## 3.1 INTRODUCTION

In this chapter, we examine the problem of tracking shapes and estimating surface alignments over complete temporal sequences of shapes seen from multiple calibrated cameras.

When estimating 3D shapes per frame, one is left with a representation that has no temporal consistency, meaning the geometric representations have no primitive correspondence. In our case with the Kinovis platform multi-RGB video streams, and applying the algorithms of the type described in chapter 2, that means we retrieve a set of meshes per time frame that are all independent and possess completely different, unmatched sets of vertices and triangles. [1]

These per-frame representations can be used as-is for some applications, such as streaming and displaying the models in sequence, but this has a number of limitations. The first one is that it is memory and bandwidth inefficient. All resources must be transferred from scratch per frame, typically to the GPU for displaying or through the network for streaming. Pre-loading the sequences doesn't circumvent this bottleneck as it introduces a transfer latency at the beginning of a sequence. Additionally, this may not be possible on a GPU because each mesh can be several megabytes depending on resolution and a sequence of meshes can saturate GPU memory quite fast. All of these problems are drastically worse when one starts attaching attribute data to the mesh, typically appearance texture data which allows to rendering perceived detail on the surface of the shape.

1. Due to the chronology of this work, at the time we investigated the alignment methods described in this chapter, we were using mostly polyhedral visual hulls obtained with algorithms from my thesis [FB03, FB09], as these were fast, rugged, and implemented with the platform software suite we were using with Kinovis ; but this doesn't alter the essence of this discussion.

More generally, all temporal-related tasks requiring surface tracks or speed or a motion representation, such as motion editing and motion-driven interactive setups, are not possible with this inadequate frame-to-frame representation.

Fundamentally, throwing away the previous reconstructed shape information is sub-optimal : in typical acquisition scenarios involving human or non-human subjects, the underlying shape observed in motion has a fixed physical envelope and topology, or a set of discrete *geometric* topologies, depending on whether self-contacts are considered as merges of the geometic envelope of the object [LB12]. Using the hypothesis of a common underlying surface means recovering the sequence as a moving shape, whose topology is at least temporarily stationary. In concrete terms one can then store a single shape topology or sparse set of topologies for the whole sequence, often referred as template, and store frame updates as a set of relative or absolutes motions, expressed as raw vertex displacements, local piecewise transforms using a shape decomposition such as patches [CBI10b], or updates in a parametric or kinematic representation [dATSS07].

This provides a wide set of benefits, notably reducing the memory and bandwidth requirements of the representation, since only the motion updates need to be transferred per frame. This is especially true when one attaches additional surface attribute information, such as a appearance, to the fixed-topology templates rather than each time frame [CCS+15], thus avoiding needless duplication. From a broader standpoint, subject and human motion analysis then become possible, such as monitoring specific attributes of motion, *e.g.* trajectories of specific landmarks, building motion statistics, using the model to compute dynamic collisions with virtual objects in a immersive digital world, with a wide range of applications for sports, medical, virtual reality, entertainment and interactive systems. This type of representation is also a pre-requisite to make the motion sequence *editable* [SdATS07, MT02], to use it in a content production pipeline where artists or designers wish to use real-life full shape captures as the starting point for 3D asset creation, in which modification and stylisation of the motion are targeted.

### 3.1.1 Alignment Problem

The benefits are clear, but the problem itself is challenging due to various aspects. Looking at it from a global perspective, it can be seen as retrieving from a set of image sequences $\mathrm{I} = \{\mathrm{I}_i^t\}_{i\in\{1..n\}}^{t\in\{1..m\}}$ a single, time-independent set of parameters of a *shape model* $\mathcal{S}$ of the subject on one hand, and a set of (most often per-frame $t$) motion parameters of a *motion model* $\Theta = \{\Theta^t\}^{t\in\{1..m\}}$ on the other hand, such that $\{\mathcal{S}, \Theta\}$ optimally explain the full set of image sequences I. The shape representation $\mathcal{S}$ will contain geometric attributes of the shape, describing its surface, volume, pose, etc., and may also include non-geometric attributes such as appearance, or other features.

Although some methods do treat the problem this way as a single, global sequence retrieval problem [GM04], in particular in some recent publications [NMOG19], it is very complex and technically challenging for a number of reasons. First the size of the image sequence data I, and shape and motion parameter dimensionality, usually

prohibits global optimization update steps relating directly all images $\mathrm{I}$ to all motion parameters $\Theta$ and the shape $\mathcal{S}$. Second because in several ways this is a chicken and egg problem : updates to the shape require updates to the motion parameters, and vice versa. Matching shape surface points to identified 2D projection in the viewpoints in the different temporal frames requires updating the motion model, and vice versa. Third it is intrinsically difficult to formulate spatio-temporal surfaces directly in terms of multi-view and temporal image content variation, which this view of the problem implies.

Interestingly some formulations close to the latter standpoint exist, *e.g.* [GM04] where the moving shape is described as an spatiotemporal isosurface optimized from image variations. While quite elegant, as in many methods, the formulation however needs to make practical compromises, *e.g.* it forgoes describing the single underlying shape $\mathcal{S}$, and in practice makes partial updates in frame batches that are slow and susceptible to local minima.

Because of these difficulties, the vast majority of estimation methods in the literature propose stratified approaches, breaking the problem down into more tractable sub-problems, and this is still the case today with state of the art approaches, *e.g.* [BHKH13, MVK+20]. This can be done along various dimensions, *e.g.* Simon *et al.* [SVMS14] consider spatiotemporal priors on point trajectories, putting more emphasis on temporal connections than shape connectivity. But a majority of techniques first pre-process whole shapes in individual frames, yielding a set of independent shapes $\{\mathcal{S}^t\}^{t\in\{1,\cdots,m\}}$ substituted for inputs of the 3D temporal alignment stage, which is the strategy we follow. This is also a natural path to tackle the problem by stepping up from our expertise in 3D reconstruction techniques from a calibrated set of images.

### 3.1.2 Model-Based Approaches

In the stratified view of the problem relying on per-frame reconstructions, we must first select and conceive shape and motion models. Various strategies exist to this end, but in this document and in the work described in this chapter we took interest in so-called *model-based approaches*. These adopt a further simplification to the general paradigm described earlier in §3.1.1, by restricting the shape $\mathcal{S}$ either to a family of parametric shapes [ASK+05, HAR+10, NH13, PWH+17] or to a single *template* shape [SH03, BC08, VBMP08, GSA+09, LGS+13], to name only a few. In our work, we explored both possibilities. This simplification thus trades the complexity of matching of spatio-temporal 4D representations to input images, for the problem of fitting a model by optimizing its deformation parameters $\Theta^t$ such that it best explains per-time step 3D reconstructions.

Examining the literature, one can see that most such approaches mostly share a common general canvas : once the motion model is chosen, they all require defining a 3D matching cost or loss to provide a metric measuring the disparity between the deformation estimate and the 3D reconstructions used as input of the fitting. This loss can be either directly minimized for certain forms of the loss function, *e.g.* Chamfer which considers maximum surface to surface distances, or the minimization can be interlea-

ved with an association step, where shape primitives are matched to the primitives of per-frame reconstructions in ICP fashion [BM92, GP02, MCA07].

I will here discuss various template-based approaches we explored based on motion expressed on detecting rigid parts for surface or volume patch decomposition of a single template shape to track (§3.2), and an approach based on the human shape space S-SCAPE which is equipped with a multi-linear deformation framework for identity and kinematic motion (§3.3), all of which illustrate particularly well our research interests and progress on this subject.

## 3.2 Surface and Volume Patch-Based Shape Templates

With the first model, our goal was to explore different improvements of a fixed template object $\mathcal{S}$ that allow more robust and more precise fitting. A very popular strategy which we follow in this work is to use a particular capture or scan of the exact subject to be tracked as template shape, popularized by various works in the 2000s [dAST+08]. With this simplification advantage of a fixed shape $\mathcal{S}$, the essence of the alignment approach is to define a deformation model and its motion parametrization $\Theta^t$.

Many such parametrizations exist in prior works. One is to rig a generic human model using kinematic chain which can be pre-fitted to the template [BC08, VBMP08, GSA+09, LGS+13] or in some cases automatically extracted [BP07]. Another is to relax this strong kinematic prior by using local surface rigidity constraints, with the idea that looser skin or clothing can then also be fitted by the model. Typically, the idea is to preserve local intrinsic surface properties, e.g. isometric deformations [MS04, BBK06, OMMG10, SY10, BPC13], conformal deformations [BGC+15], inextensibility of the template mesh [SMNLF08], or general neighborhood preservation through functional warps [DB11]. Others use properties related to local rigidities, *e.g.* as-rigid-as-possible deformations or elasticity minimization in [WSSC11, ZWG+13, BGC+15]. Among the methods allowing to express rigid cohesion of the template surface, and a particular topic of interest to us, patch-based methods [CBI10a, CBI10b, BHH11] offer an interesting compromise as they decorrelate the surface support of the deformation model from the geometry resolution, by formulating elasticity-like constraints between a set of pre-computed surface patches.

To summarize, on one hand at the time we had a large set of methods with very strong priors on human articulation, which couldn't deal with non rigid surface aspects such as clothing or objects held by the subject, on the other hand we have a set of generic surface fitting methods that use weak rigidity constraints that can deal with many kinds of surfaces but are completely agnostic to the underlying characteristics of the shape, assuming instead uniformly distributed local cohesion properties.

This raised my curiosity and sparked a research effort aimed at finding a middle ground between the two families, *i.e.* a deformation model that could be applied to most situations, *e.g.* for humans with or without loose clothing or even holding objects, able to model both rigid and non-ridid aspects in a single method. One question this

**FIGURE 3.1 –** *Illustration of the convergence of algorithm [FB11] for a single frame of the* LOCK *sequence, courtesy [SH07]. Each rigid patch assignment is given by a different color.*

raised for me was whether an alignment inference method could become automatically aware of the distribution of rigid regions on the template surface.

### 3.2.1 Preliminary Effort on Rigidity Learning

My first attempt toward this goal [FB11] was to devise a method that introduced rigid segmentation parameters $\mathrm{K}_v \in \{\emptyset, 1..k\}$ associated to every vertex $v$ on the template surface. Each of the $k$ rigid components was assigned a set of rigid motion parameters $\Theta_k^t$ per time step $t$ and rigid component $k$. An additional robustness outlier class label $\emptyset$ was included for every $\mathrm{K}_v$ to allow for a vertex to also be drawn from a free uniform motion component unexplained by the $k$ rigid components. Each input observed surface point $o$ at time $t$ was also given an association variable $\mathrm{V}_o^t$ basically attaching a matched template vertex $v$ to every observed point of the input reconstructions at time $t$.

A joint probability distribution $p(\mathrm{K}_v, \mathrm{V}_o^t, \Theta^t)$ was then expressed on the full set of statistical variables $\mathrm{K}_v$, $\mathrm{V}_o^t$ and $\Theta^t$ to allow for a Maximum A Posteriori (MAP) to be computed, using Gaussian priors for the distribution of labels on the template mesh, and Gaussian noise distributions on the difference between observed surface points $o, t$ and their deformed template associated vertex given by $\mathrm{V}_o^t$. By treating $\mathrm{K}_v$ and $\mathrm{V}_o^t$ as hidden variable sets in the inference, one could then write a formal Expectation Maximization algorithm akin to a specialized GMM, that basically extracted a locally optimal point estimate of the rigid-component transforms, a set of per-vertex probability tables over $\mathrm{K}_v$ that amounted to fuzzy and automatic rigid-guided patch-segmentation extension of [CBI10a], and sparse probability tables over $\mathrm{V}_o^t$ for each observed vertex $o$ at time $t$, giving probabilistic associations to the data points at each time frame, which essentially builds EM-ICP in the method. This was published in CVPR 2011 [FB11].

**FIGURE 3.2 –** *Tracking excerpts from the* DANCER *dataset. Colors code patches.*

The method had some excellent qualities, in simultaneously exhibiting a data-driven solution to rigid segmentation of the template, rigid motion and closest point assignments with closed form updates resembling those of a standard GMM, essentially giving a principled solution to the template-based multi-rigid EM-ICP problem. However, this initial attempt was also undermined by possible losses of tracking for our practical data, with complete stretching and disconnection of the rigid patches. This is because it had not built strong inter-vertex cohesion constraints in the method and the general surface cohesion provided by the rigid components was too weak to prevent a rigid component to go astray.

### 3.2.2 Detecting Rigidities on a Patched-Based Surface Template

We were still convinced that a more performance capture-suitable analysis was possible and that it would yield improvements in results quality with respect to purely agnostic models. So we included this as PhD topic in European Project React that was aimed at creating new vision-driven workflows for 3D digital content creation, which sparked a 4-year effort on this topic with PhD student Benjamin Allain. The first work was also motivated by discussions with Tony Tung of Kyoto University at the time, and the paper became a collaboration.

Our first effort was a better attempt to a surface-based template with patch-based deformation. To this goal, we opted for a much simpler model than previously. First we computed a fixed patch decomposition of the template surface into $k$ uniformly distributed patches (see Fig. 3.2), exactly as in the work of Cagniart *et al.* [CBI10a, CBI10b]. However, where Cagniart opted for a uniform expression of the inter-patch deformation energy - that is, regardless of the location on the template surface - we introduced a set of binary-valued *rigitidy deformation variables* $\mathrm{C} = \{\mathrm{C}_{k,l}\}_{(k,l)\in\mathcal{N}}$,

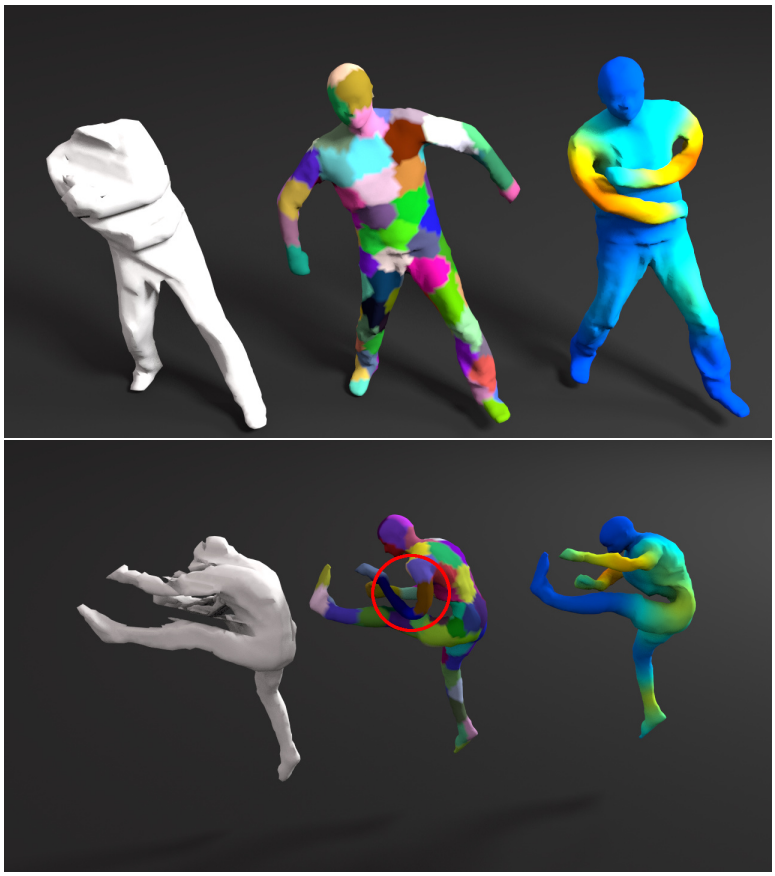


**FIGURE 3.3 –** *(left) Input mesh, (middle) tracked mesh without rigidity inference - Cagniart et al. [CBI10b] and (right) with rigidity inference. The absence of the head on the top-left input mesh is imputable to the visual hull computation method, which assumes full visibility for each camera. It can be observed that our method corrects substantial artifacts.*

with each $C_{k,l} \in \{0, 1\}$, for every pair $k, l$ in the set $\mathcal{N}$ of immediately adjacent patches on the fixed template geometry. $C_{k,l}$ is meant as a time-independent inference variable describing how rigidly correlated two neighboring patches are.

In some sense this can be seen as a simpler way to implement the initial idea of §3.2.1 [FB11], first because the rigidity variables $C_{k,l}$ can together express arbitrary rigid aggregations of the initial and smaller "primordial" $k$ patches. Second, while the number of rigid components were fixed in the previous effort [FB11], the arbitrary rigid aggregations afforded by the $C_{k,l}$ variables allow support for any composed set of $k'$-component patch aggregates by this method, as long as $k' <= k$. Third, the strong surface cohesion previously missing is now built-in the inter patch distance prior in the expression of a surface prior term which basically quadratically penalizes neighboring patches that come apart. When $C_{k,l} = 0$, that is when patches $k$ and $l$ are non-rigidly correlated, the surface term is chosen to allow neighboring patches to loosely hinge. When $C_{k,l} = 1$, that is when pacthes $k$ and $l$ belong to the same rigid group the surface

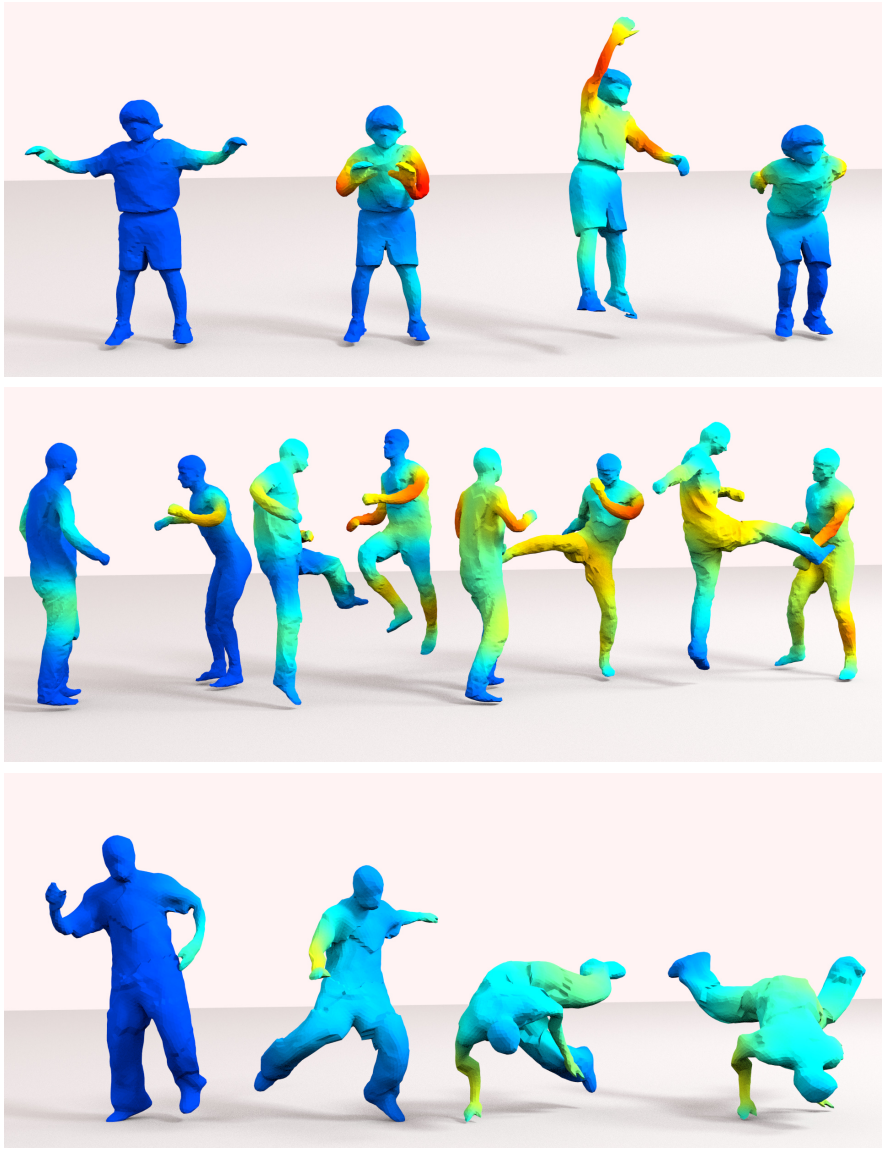


**FIGURE 3.4 –** *Tracking excerpts from* GOALKEEPER, MARKER *[LGS+13] and* FREE *datasets showing a heatmap coded values of the* $C_{k,l}$ *rigidity variables : red when very likely non rigid, blue when very likely rigid. Note that the inference is performed on a time window, i.e. the rigid inference corresponds to rigidities as observed in that window.*

intrinsic penalty term penalized any non rigid movement between the two patches. The probability assigned ot each $\mathrm{C}_{k,l}$ is then expected to be estimated based on the data at inference time thanks to these prior constraints.

An additional contribution of this model is in how it computes inter-patched distances. We had noticed a bias in how Cagniart computed inter-patch penalties [CBI10a] that basically favored estimates close to the default template pose. For this reason, we included a mean pose $\bar{\Theta}$ as a point estimate to be retrieved during the optimization for a full sequence, which allows the model to significantly deviate from the default template pose, and also acts as a regularizer to stabilize estimates patch pose etsimates in the sequence.

Similarly to §3.2.1, each (now constant) patch $k$ is again given a set of rigid transform parameters $\Theta_k^t$ at each time frame $t$ of the processed sequence of reconstructions, and each observed point $o$ on the input 3D meshes at each time step $t$ is again given a template vertex assignment index $\mathrm{V}_o^t$. The inference can then again be expressed as a MAP over the joint probability $p(\mathrm{C}, \mathrm{V}, \Theta, \bar{\Theta})$, treating $\mathrm{C}_{k,l}$ as hidden. Using Expectation Maximization, one then alternates between refining point estimates of the patch transforms $\Theta^t$ and mean patch pose $\bar{\Theta}$ in the maximization step, and computing probabilities over each patch pair $(k, l)$'s rigidity variable in $\mathrm{C}_{k,l}$ on one hand, and EM-ICP-like template vertex assignment probabilities $\mathrm{V}_o^t$ on the other hand, both as part of the E-step.

The results obtained with this simple idea confirmed the expected regularizing behavior of sharing a common rigidity inter-patch property over all frames of a sequence (see Fig. 3.2 and Fig. 3.4), leading to better temporal alignment and more robustness, quite apparent in Fig. 3.3. The features we built in the method also acted as a form of damping and stabilization in the inference when applied in sliding-window fashion across longer sequences. Detailed results are available in the original ECCV 2014 publication [AFBT14a] (available as Appendix A.2) and supplemental video [2].

### 3.2.3 Volume-Patch Rigidity-Enforced Template

Although encouraging, it appeared from previous results that one of the remaining limitations of the previous method was in fact a common limitation of the surface-based family of methods : the method exhibits some stretching, squeezing and generally rubbery looking artifacts in non rigid zones.

As with LBS, this was prominent at kinematic joints, where a non smooth split between two neighboring rigid or quasi-rigid is expected; however the instrinsic surface priors we chose, when weighted by an intermediate, undecided probability of rigid correlation in the neighboring patches, tend to smoothen the energy required to follow the limb joint across limbs rather than putting all the deformation at the joint location (see which highlights these cases).

Our hope was in some sense that the rigidity inference and regularizing effect would at least partially mitigate some aspects of the problem, which turned out to be

2. https://hal.inria.fr/hal-01016981/file/Allain_ECCV2014_On_Mean_Pose_and_Variabiliy.mp4

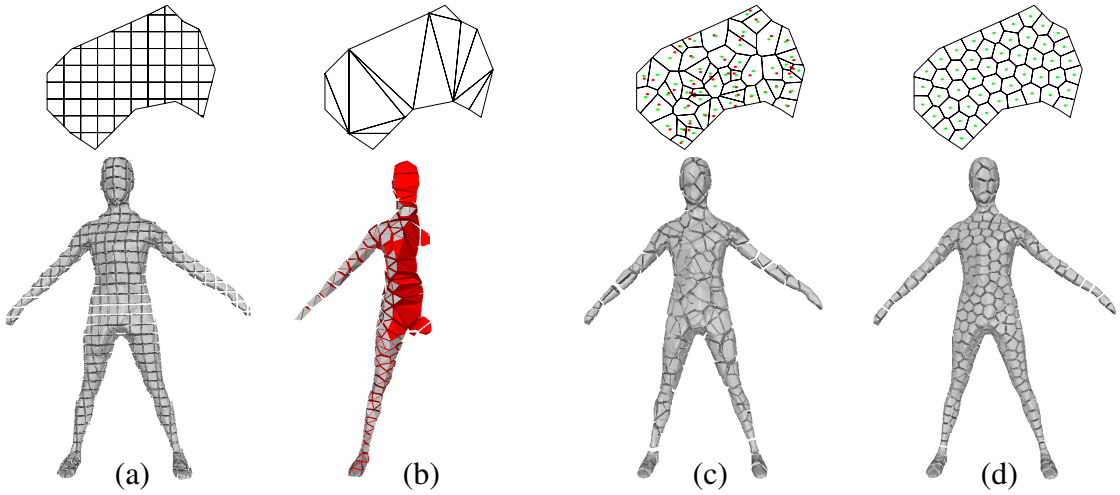



**FIGURE 3.5 –** *Possible volumetric decompositions of the template and observed shapes. (top) 2D schematic view. (bottom) 3D decomposition example. (a) Voxels on a regular grid. (b) A sliced Constrained Delaunay tetrahedrization showing the elongated inner tetrahedra generated. (c) Voronoi cells with random centroids shown in red, center of mass of each cell in green. (d) Centroidal Voronoi tesselation cells, where the center of mass and cell centroid coincide.*

correct, but we also sensed that this could be better enforced. One idea could have been to use a different norm as support for the surface tension expression, which in fact has later been explored at the time of this writing, *e.g.* [GXW+15] in the case of depth camera streams. Another possible path which we ended up following was to enforce rigidity constraints at the volumetric level, where volume preservation could then be built in the model to alleviate rubber and squeezing effects. A source of inspiration was to observe what had been done in the graphics community to express volumetric deformations and to mitigate problems of the LBS model [ACOL00, ZHS+05, BPWG07], which in fact had already been used for performance capture applications [dATSS07, dAST+08, BH10].

As illustrated in Fig. 3.5(a), those approaches are however based on tesselations of surface points of the input reconstructions and do not introduce any inner vertices. This allows some volumetric constraints to be taken into account for the deformation energy, but prevent a full volumetric treatment of the alignment problem. Other tesselations of the volume are possible (Fig. 3.5(b) and (c)), such as a regular voxel grid, but it is non isotropic and biased along the axis directions. Using Voronoi cells of points randomly drawn inside the volume is insufficient as it yields irregular inner cells.

Instead we propose a fully dense an regular volumetric treatment of the volume using a CVT of the volume (Fig. 3.5(d)). Informally CVT are a particular type of Voronoi tessellation where the samples are iteratively repositioned to coincide with the center of mass of their cell, which achieves the desired properties [DFG99] : isotropy, rotational invariance, uniform cells of compact and regular form factor, regular intersection of boundary cells and surface, independent cardinality and practical computa-

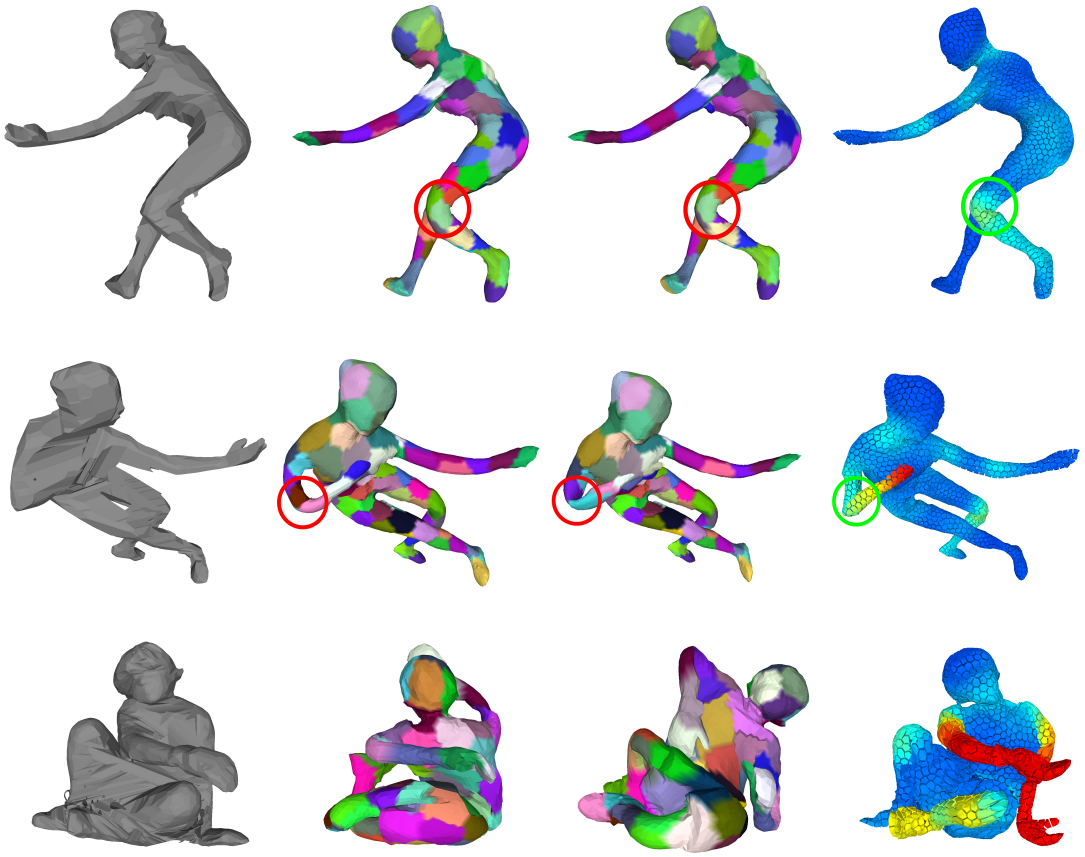



**FIGURE 3.6 –** *Frames of the* BALLET *(top and middle) and* GOALKEEPER *datasets (bottom). (a) Visual hull input. (b) Tracking result of Cagniart et al. [CBI10b]. (c) Allain et al. [AFBT14b]. (d) Our method. Note the improved angular shapes on the dancer's knee (top) and elbow (middle), and the improved robustness (bottom).*

tion. On can also choose the number of final cells desired, thus adapting the complexity to the desired problem granularity.

The tracking methodology and deformation model is then strictly analogous to §3.2.2, the main initialization difference being that both the inputs and template shape are now tesselated into a set of CVT cells, on which the whole algorithm operates. Namely, the initial patch decomposition is now expressed on volumetric cell groupings instead of surface vertices, and observation-to-template matches are also expressed from input volumetric cells to template volumetric cells. We keep the rigidity inference on $\mathrm{C}_{k,l}$ variables based on pairs of adjacent volumetric patches in the template as a complementary regularizer of rigid behavhior over each inferred sequence as in the previous method.

We observed significant qualitative and quantitative improvements on Kinovis and BBC React acquired datasets [3], as measured with both the silhouette reprojection error for template-fitted sequences and sparse ground truth marker RMS error for some sequences which had been simultaneously captured with multi-RGB and sparse markers [LGS+13], as illustrated Fig. 3.6. One can notably observe the more natural folds at kinematic joints, and the increase robustness even with truly nightmarish data, *e.g.* GOALKEEPER has some frames where the subject gets up after leaning on the floor, with a very messy visual hull reconstruction, and the volumetric approach does not lose the main anatomic features. This paper, provided in appendix A.3 was presented as an oral at CVPR 2015 [AFB15]. More results can be seen in the supplemental video [4].

### 3.2.4 Volume-Based Tracking-by-Detection

For the record, I will briefly mention here an interesting ramification of this work demonstrating the usefulness of a CVT-based volume approach. Previously, the deformation model was the focus of our discussion and work, and associating observations to the template was performed with EM-ICP type terms. We showed, with a collaboration with Chun-Hao Huang, Federico Tombari, Slobodan Ilic and Nassir Navab, that is also possible to use the volumetric CVT template decomposition as the basis for the association component of the loop. For this, we directly took inspiration from work on the Vitruvian Manifold [TSSF12], where the depth-anchored, learned data-driven association model based on Random Forests [SFC+11] was generalized to continuous regression of depth points to a generic template surface, then applied as a one-shot association component to drive a non-rigid human body model fitting scheme.

We proposed in [HAF+16, HAB+18] to substitute into a Vitruvian-like pipeline CVT-based representations for the volumetric feature extraction and regression stages with Random Forests, this time driving the volume-based deformation model of volumetric templates described in the previous section, with of course some technical adjustments. The results showed drastically improved stability of the results and better trainability for a CVT-decomposition of the shape versus a training based on a regular grid volumetric feature support. We show results where the alignment is recovered even with truly challenging situations where the subject is a performing a full cartwheel, without any tracking loss. The PAMI version of this article explains the original methods, as well as additional temporal stability improvements, is provided as appendix A.6, with more results in the supplemental video [5].

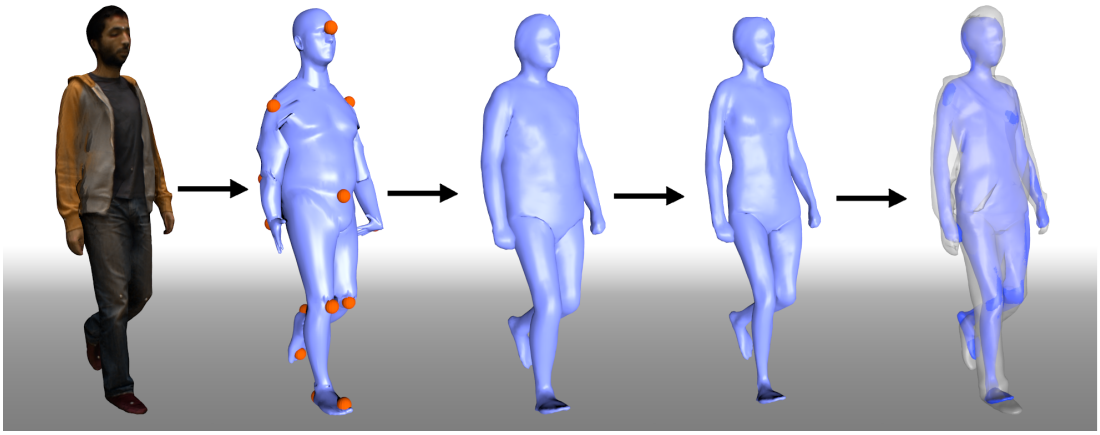


**FIGURE 3.7 –** *Overview of the proposed pipeline. From left to right : input frame, automatically computed landmarks using [ZB15], result after estimation of initial identity and posture, final result, and overlay of input and final result.*

## 3.3 INFERRING SHAPE UNDER CLOTHING USING SHAPE SPACES

In the work discussed above, we have used priors over rigidity but not complete priors of observing humans, only using the template shape but no constraint on human kinematics. This has both advantages and disadvantages. On one hand the generic nature of the deformation and fitting model allows uniform treatment of a wide variety of scenarios, such as people in clothing, people holding objects, and some hair detail, as long as the geometry of these features are pre-observed as part of the template used for tracking the sequence. Those added feature would elude methods purely based on a generic human shape template with no clothing for example. On the other hand, one can argue that, since we are observing humans, it only makes sense to use that prior in some way.

Various methods have gone in the direction of either explicitly including a clothing model, for example a specific skirt or type of clothing [RKP+08]. More recently more and more layered models have been proposed that use the underlying body and kinematic structure as a shape and motion prior guiding an outer cloth layer [ZPBPM17].

Before these latest works, in 2016, we worked on model that estimates shape *under* clothing, that is by only using the outer envelope over a fully observed sequence as a guide constraining the underlying human shape possibilities and narrows them down to the most plausible data given the indirect observations of it.

But this only works if we are given a low dimensional representation describing the full set of plausible human body shapes. This is why we rely on S-SCAPE, a mul-

3. Video available at https://hal.inria.fr/hal-01141207
4. https://hal.inria.fr/hal-01141207/file/Allain_CVPR2015_Volumetric_Tracking.avi
5. https://hal.inria.fr/hal-01300191/file/video.mp4

tilinear variant of SCAPE that is easier to manipulate and optimize. This was an ideal topic to start collaboration and PhD co-supervision of student Jinlong Yang, with then newly arrived Inria Researcher Stefanie Wuhrer, who joind the team in 2015 and had prior expertise in this field as a co-author of the S-SCAPE model. The well-known SCAPE family of methods has been introduced in the 2000s as a means to encode body shapes with two sets of parameters, a set of kinematic parameters governing the body pose observed, and a set of intrinsic body shape parameters. The kinematic set is the same controlling as other kinematic model based methods, but the body shape parameters are essentially a set of coordinates in a PCA-shape basis, which is learned from a database of human shapes with different body characteristics, such as height, gender and corpulence. The immense advantage of such a model is its ability to describe a specific body shape instance, also called "identity", with a few dozen parameters, and a description of shape and pose together only spanning around a hundred parameters.

We cast the fitting problem as a minimization procedure over the set of identity parameters $\beta$, which basically encode the estimated intrinsic shape $\mathcal{S}$ of the subject, and the set of $m$ per-frame poses $\{\Theta^t\}_{t\in\{1,\cdots,m\}}$, which in this case are the set of joint parameter of the S-SCAPE LBS-based kinematic model. We define a loss to minimize over the set of observed input 3D shapes over time $\{\mathcal{S}\}_{t\in\{1,\cdots,m\}}$, with

$$E_{sequence}(\beta,\Theta^t,\mathcal{S}) = \sum_{t=1}^{m} \omega_{lnd}E_{lnd}(\beta,\Theta^t,\mathcal{S}) + \omega_{data}E_{data}(\beta,\Theta^t,\mathcal{S}) + \omega_{cloth}E_{cloth}(\beta,\Theta^t,\mathcal{S}), \tag{3.1}$$

where $E_{lnd}(\beta,\Theta^t,\mathcal{S})$, $E_{data}(\beta,\Theta^t,\mathcal{S})$ and $E_{cloth}(\beta,\Theta^t,\mathcal{S})$ are energy terms weighted by scalars $\omega_{lnd},\omega_{data}$ and $\omega_{cloth}$. The landmark term $E_{lnd}(\beta,\Theta^t,\mathcal{S})$ measures the distance between a set of automatically computed landmarks on $\mathcal{S}$ and their corresponding anatomical points on $\mathbf{c}(\beta,\Theta^t)$. This energy allows to obtain a rough estimate of the body shape and posture at each frame. The data term $E_{data}(\beta,\Theta^t,\mathcal{S})$ measures the distance of points $\mathbf{c}(\beta,\Theta^t)$ to the nearest neighbors on scan $\mathcal{S}$ and serves to pull the estimate towards the observed scan surface. This term allows to obtain a good estimate for the identities $\beta$. The clothing term $E_{cloth}(\beta,\Theta^t,\mathcal{S})$ accounts for the loose clothing by encouraging $\mathbf{c}(\beta,\Theta^t)$ to be located inside the observation $\mathcal{S}$. Since the cloth term is applied to all frames, it allows to take advantage of the motion cues observed throughout the sequence ; as the clothing moves close to localized regions of the body in different frames it restricts the underlying shape §encoded by $\beta$ to essentially lie inside the observed cloth for all frames.

To acquire our solution over full sequence which may span hundreds of frames, as illustrated in Fig. 3.7, we begin by extracting the matching landmarks $\mathbf{c}(\beta,\Theta^t)$. Second, we initialize the identity parameters $\beta$ by minimizing (3.1) for a small subsequence at the beginning of our full sequence. Third we solve (3.1) for the pose parameters $\Theta^t$ sequentially over the full sequence, while fixing $\beta$. Fourth, we refine the identity parameters $\beta$.

To evaluate the proposed framework, we have captured a database of six subjects (three male, three female) performing three different motions and wearing three clo-

**FIGURE 3.8 –** *Six representative examples of frames of our motion database. From left to right, a female and male subject is shown for tight, layered, and wide clothing each.*

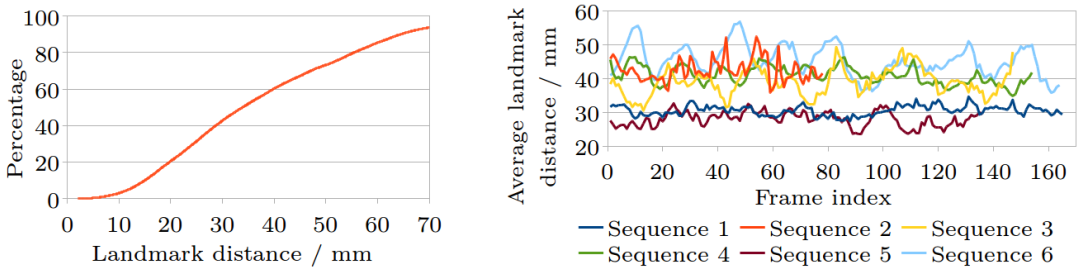



**FIGURE 3.9 –** *Accuracy of posture estimation over the walking sequences of all subjects in tight clothing. Left : cumulative landmark errors. Right : average landmark error throughout each sequence. Figure from [YFHWW16].*

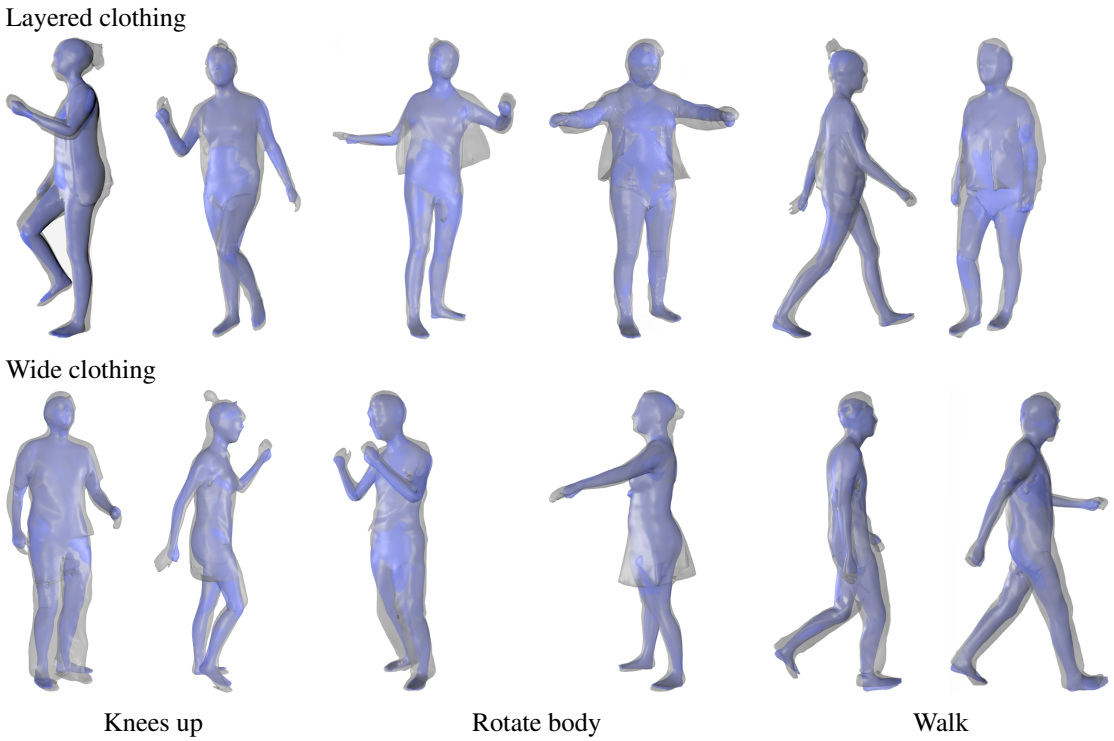



**FIGURE 3.10 –** *Overlay of input data (grey) and our result (blue). Figure adapted from [YFHWW16].*

thing styles each using the Kinovis platform, which has become popular with the community, (*e.g.* [ZPBPM17]), the conference paper [YFHWW16]. We make it available as Appendix A.4 and refer to the supplemental video for more results [6].

Figure 3.8 shows six representative frames of the database. Figure 3.9 evaluates the posture estimation on manually placed markers for walking sequences captured in tight clothing. Note that the use of S-SCAPE as statistical prior prevents drift. Figure 3.10 shows further qualitative results of our method. Due to these very interesting results and originality of the work, the paper became quite popular inspiring for instance several great followups in other teams, *e.g.* [ZPBPM17]. We also had a followup collaboration with Jinlong Yang and Stefanie Wuhrer examining the statistics of the outer shape as a vertex displacement layer, which I'm not discussing here in detail in the interest of space [YFHWW18].

6. https://hal.inria.fr/hal-01344795/file/supplementaryVideo.mp4

# CHAPITRE 4

# Appearance Estimation and Refinements

## 4.1 INTRODUCTION

Multi-camera capture platforms such as the 68-camera Kinovis produce a lot of observations on the same scene and from the same surfaces. Thanks to the sequence alignments described in the previous chapter, we have access to an additional source of observation redundancy. A natural research goal that has been the focus of our attention in the last decade, is how to best exploit this redundancy to enhance the quality of the models produced. This is a particularly important aspect of our work since, as was previously discussed, the subjects we acquire cover only a fraction of the input image pixels, limiting the actual amount of appearance data present in each input frame.

We discussed several of our works toward this goal concerning geometry acquisition in chapter 2, but another aspect to enhance the perceived realism and appeal of our output models is to acquire the fine-grain color appearance and texture of those subjects. To this goal we can leverage the multiple RGB streams relating to the same underlying surface.

We here discuss two families of works in this direction that we carried, to acquire appearance data, first in the direction of multi-view superresolution (§4.2), second in how to efficiently describe and estimate a temporally evolving appearance representation (§4.3).

## 4.2 HIGH RESOLUTION 3D SHAPE TEXTURE FROM MULTIPLE VIDEOS

Retrieving appearance information from all views of the subject raises interesting questions. In the case of a single video, a vast literature on 2D superresolution methods

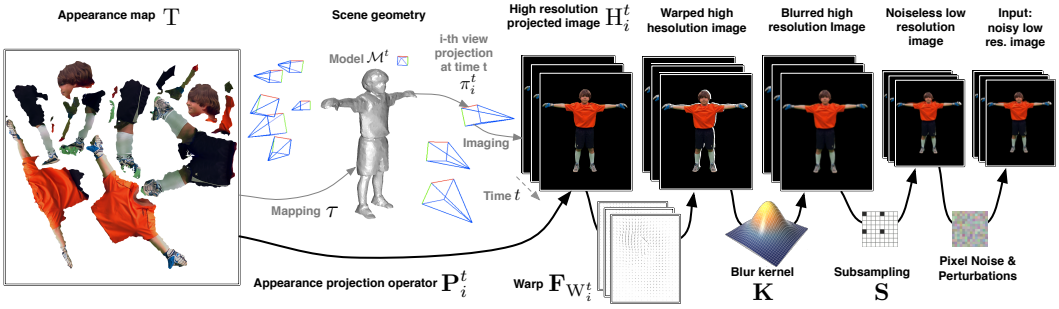



**FIGURE 4.1 –** *Summary of image formation model and problem notation.*

exists, which by analogy raises the possibility of retrieving appearance detail from the multi-view streams, with a quality beyond that of any given input frame in isolation.

When we took on this project with PhD student Vagia Tsiminaki, co-supervized with Edmond Boyer for the React European project on content production, the literature on the subject mostly looked at multi-view texture estimation without looking at the temporal redundancy, by aligning the various contributing images onto a single texture to avoid ghosting [TAL+07, LHS01], or building a texture patchwork of single-view contributions optimized with graph cuts [LI07], extended to the temporal domain as one of the rare methods addressing time redundancy in the multi-view case [JP09] at the time. Only a handful of particularly relevant works has started to examine this explicitly as a multi-view superresolution problem, but without considering temporal frames [GC09, GAKC13].

The literature in the monocular video case is however abundant. An interesting thing we learned from reviewing those works is that a well identified generative model of low resolution image formation had emerged, as a geometric warping, blurring and sub-sampling process of the initial high-resolution image [BK02]. Of particular interest to us are that this model can be represented by a stack of linear transforms and that Bayesian noise models have been developed to explicit the noise dependencies and statistical priors over the image and warps to estimate [FSG07], some using the L1-norm based priors and total variation (TV-)minimization popular for image restoration tasks [LS11]. Notably, super-resolving multiple videos of a moving subject was examined in a performance capture context, but only for the input viewpoints [Tun08].

### *Proposed Methodology and Contributions*

Essentially no multi-view superresolution technique existed that used the time-tested elements of 2D superresolution generative models to retrieve a common appearance map for a 3D model, nor was the temporal aspect really examined in this context either. Yet the building blocks for that were there, including for example papers such as [EDM+08], which hinted at how optical-flow-driven warps could be substituted for the 2D optical flow driven warps in 2D superresolution pipelines.

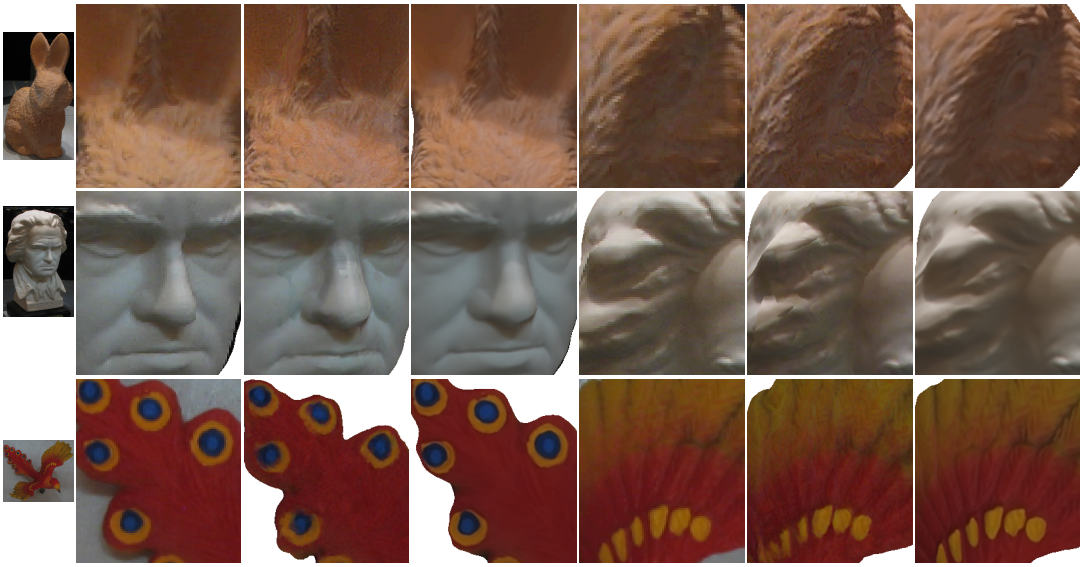


**FIGURE 4.2 –** *Comparison on* BUNNY, BEETHOVEN *and* BIRD *datasets. Left column : input images. Middle : output of [GAKC13]. Right : our algorithm.*

We looked at the problem as one of inferring a single appearance map $\mathrm{T}$ attached to a reconstructed geometry pre-aligned with [AFBT14a], from all input frames $\mathrm{I} = \{\mathrm{I}_i^t\}_{i\in\{1,\cdots,n\}}^{t\in\{1,\cdots,m\}}$. We proposed to generalize the linear generative pipeline of 2D superresolution technique [LS11], which explains the low resolution input views as warped, blurred, subsampled versions of an underlying high resolution (HR) image; by filling the gap in explaining the high resolution images as a projected version of the high resolution common texture. This framework is illustrated in Fig. 4.1.

To generate an input image from each HR image $\mathrm{H}_i^t$, the HR image is first warped according to the different sources of variability apparent in the image - calibration error, distortion, model geometry error - using a dense warp field $\mathrm{W}_i^t$. This warp results in a linear operator over the HR image, which we note $\mathbf{F}_{\mathrm{W}_i^t}$. The image then traverses the optical system, where it is blurred and captured by the CCD which performs light integration at every photosite. Following 2D super-resolution literature [BK02, FSG07] this is generally modeled using a Point Spread Function (PSF) with the form of a Gaussian blur kernel, followed by an image subsampling stage. Both operations can be written as linear operators, the image-wide blur operator $\mathbf{K}$ and subsampling operator $\mathbf{S}$, which are applied to the HR image to obtain a view's observed image $\mathrm{I}_i^t = \mathbf{SK}\mathrm{H}_i^t$. Remarkably, in its noiseless form, the full image formation model can thus be noted as a single, sparse linear operator $\mathbf{A}_i^t = \mathbf{SKF}_{\mathrm{W}_i^t}\mathbf{P}_i^t$ for each view $\{i,t\}$, with $3\times w_{\mathrm{T}}\times h_{\mathrm{T}}$ rows and $3\times w_{\mathrm{I}_i^t}\times h_{\mathrm{I}_i^t}$ columns, such that $\mathrm{I}_i^t = \mathbf{A}_i^t\mathrm{T}$ for each view $\{i,t\}$. This linear model is then used as the foundation to describe the noisy dependencies in the model and ultimately make this a Bayesian generative model, whose MAP we estimate by alternating the minimization of the common appearance map $\mathrm{T}$ and the warps $\mathrm{W}_i^t$, in analogous fashion to the original 2D algorithm.

**FIGURE 4.3 –** *Top :* GOALKEEPER. *Middle :* BACKPACK. *Bottom :* ACTOR. *The figure illustrates various temporal improvements. Top : Input is compared to Frame 1 and Frame 3. Middle : Input on left, Frame 1 and Frame 3 comparisons. Details are revived on the backpack, T-shirt and pants. Bottom : Input, Frame 1 and 2 comparisons.*

***Results***

We compared a static version of our algorithm to the then state of the art static multi-view resolution technique of [GAKC13], with results significantly improved, with less noise and artifacts, as illustrated in Fig. 4.2. The method was also successfully tested on several datasets acquired with the Kinovis platform and with the BBC React platform, shown in Fig. 4.3, on small temporal windows, in particular comparing the results obtained with 1, 2 or 3 frames with each time a measurable improvement. A more detailed version of the framework, which was published at CVPR 2014 [TFB14], and additional results and comments are available as Appendix A.1 and the supplemental video [1].

## 4.3 EIGEN APPEARANCE MAPS

In the latter work, we were able to retrieve high quality textures, by performing a principled fusion of the appearance data present in the different views and the different time frames. This however can only be valid over a relatively short temporal window before the perceived appearance changes. This is because the images provide a reading of the radiance of the surface patches each pixel sees and not an intrinsic value; and this radiance changes with the surface motion and orientation, and with significant changes in the lighting conditions that may occur over a larger temporal window. Appearance changes may also encode desirable appearance variations, such as change in lighting or personal expression of the subject, that have a negligible impact on geometry and could not be encoded at that level. One strategy is therefore to store estimated textures for as long as the texture doesn't significantly change, *e.g.* as in [CCS+15]. But how to represent this information for longer sequences is still an open problem and needs to account for all these different sources of variability. With PhD students Vagia Tsiminaki and Adnane Boukhayma, and given previously encouraging results described in the previous section, it came as a natural extension of their work to tackle this challenge and examine how to retrieve and represent appearance maps over full acquired 4D sequences.

We proposed to advance this aspect by providing a view-independent appearance representation and estimation algorithm, to encode the appearance variability of a dynamic subject, observed over one or several temporal sequences. Compactly representing image data from all frames and viewpoints of the subject can be seen as a non-linear dimensionality reduction problem in image space, where the main non-linearities are due to the underlying scene geometry. Our strategy is to remove these non-linearities with state-of-the-art geometric and image-space alignment techniques, so as to reduce the problem to a single texture space, where the remaining image variabilities can be straightforwardly identified with PCA and thus encoded as Eigen texture combinations. To this goal, we identify two geometric alignment steps. First, we co-

1. https://hal.inria.fr/hal-00977755/file/CVPR2014-HR-3D-Shape-Texture-from-Videos.mp4

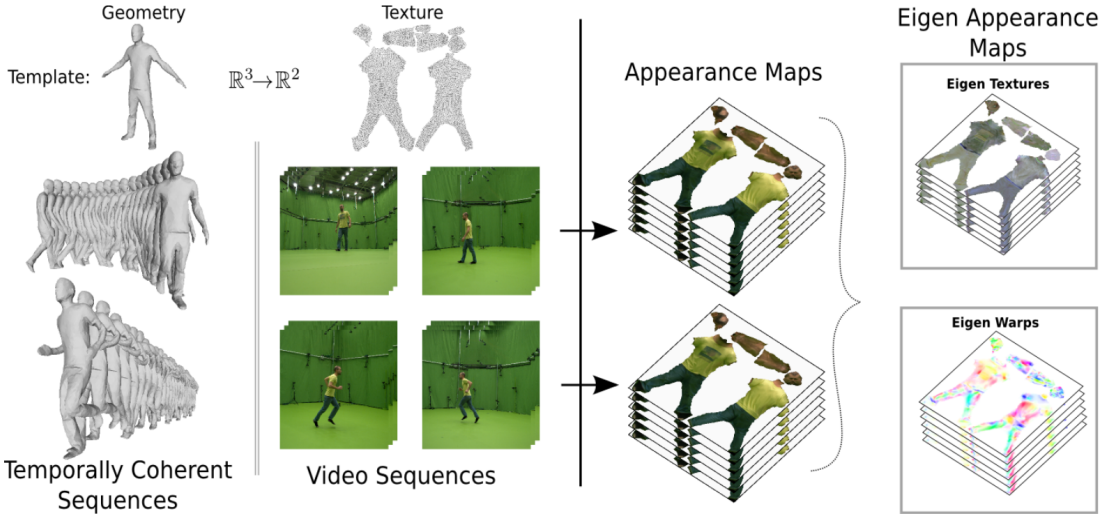



**FIGURE 4.4 –** *Overview : Time consistent shape modeling provides datasets of appearance maps. Our proposed method exploits the manifold structure of these appearance information through PCA decomposition to generate the Eigen appearance maps relative to a shape.*

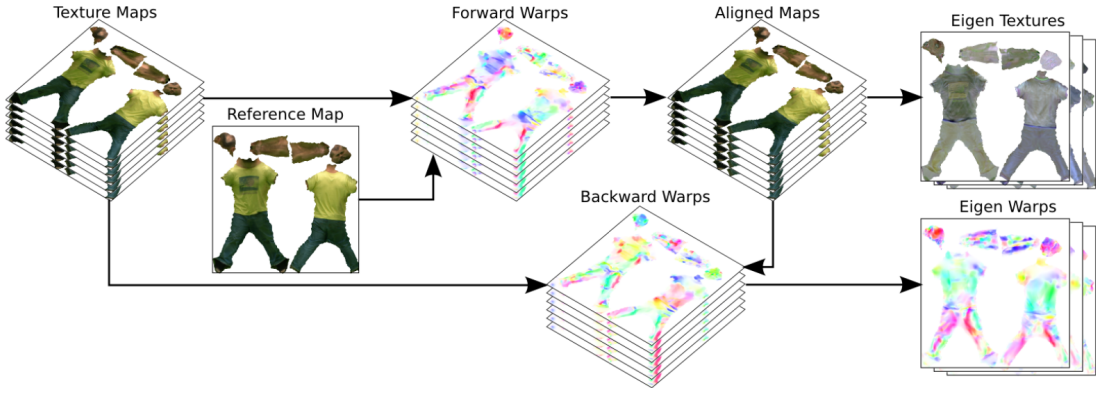



**FIGURE 4.5 –** *Method pipeline from input textures (left) to eigen maps (right).*

arsely register geometric shape models of all time frames to a single shape template, for which we pre-computed a single reference surface-to-texture unwrapping. We use the techniques of chapter 3 for this task. Second, to cope with remaining fine-scale misalignments due to registration errors, we estimate realignment warps in the texture domain. Because they encode low-magnitude, residual geometric variations, they are also advantageously decomposed using PCA, yielding Eigen warps. The full appearance information of all subject sequences can then be compactly stored as linear combinations of Eigen textures and Eigen warps. Our strategy can be seen as a generalization of the popular work of Nishino et. al. [NSI01], which introduces Eigen textures to encode appearance variations of a static object under varying viewing conditions, to the case of fully dynamic subjects with several viewpoints and motions.

The pipeline is shown to yield effective estimation performance. In addition, the learned texture and warp manifolds allow for efficient generalizations, such as texture interpolations to generate new unobserved content from blended input sequences, or completions to cope with missing observations due to *e.g.* occlusions.

### *Method*

To eliminate the main geometric non-linearity, we first align sequence geometries to a single template shape and extract the texture maps of a subject over different motion sequences in a common texture space using the previously described method in §4.2 [TFB14]. From these subject specific textures, Eigen textures and Eigen warps that span the appearance space are estimated. The main steps of the method below are depicted in Fig. 4.5.

1. Texture deformation fields that map input textures to, and from, their aligned versions are estimated using optical flows. Given the deformation fields, Poisson reconstruction is used to warp textures.
2. PCA is applied to the aligned maps and to the texture warps to generate the Eigen textures and the Eigen warps that encode the appearance variations due to, respectively, viewpoint, illumination, and geometric inaccuracies in the reference model.

Hence, The main modes of variation of aligned textures and deformation fields, namely Eigen textures and Eigen warps respectively, span the appearance space in our representation.

Note that due to texture space discretization, the warps between textures are not one-to-one and, in practice, two separate sets of warps are estimated. Forward warps map the original texture maps to the reference map. Backward warps map the aligned texture maps back to the corresponding input textures (see Fig. 4.5).

Given the Eigen textures and the Eigen warps, and as shown in Fig. 4.6, a texture can be generated by first creating an aligned texture by linearly combining Eigen textures and second de-aligning this new texture using another linear combination of the Eigen warps.

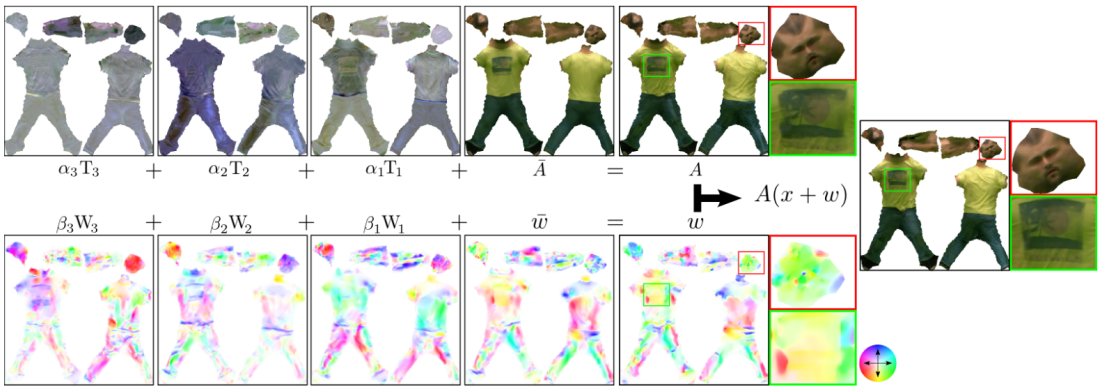



**FIGURE 4.6 –** *Texture map generation by linear combination.*

### *Results*

We evaluated the method on a number of Kinovis acquired datasets and measure the quality using the SSIM metric [WBSS04] that is more tolerant to small shifts, both in image space after reprojection and texture space before reprojection by comparing it to the short term maps estimated with §4.2 [TFB14]. The sequences are previously tracked using the method from §3.2.3 with a single template that is used as support for the common texture space. The results show that our strategy successfully encodes 2048x2048 datasets of 200 or 300 frames with virtually no error (0.98 SSIM) with 50 PCA components. We also show that using an equivalent number of parameters, a simple baseline PCA strategy without eigen warps achieves much lower performance, substantiating that the eigen warps successfully correct small geometry induced texture slippage. We also show the applicability of our method for two tasks : interpolation between two template poses, by interpolating in both eigen spaces (warp and texture) and texture completion to correct artifacts in estimated textures *e.g.* due to poor visibility, by projecting the textures to a pre-established space of our eigen representation, in both cases with encouraging results as illustrated in figures Fig. 4.7 and Fig. 4.8. More details are available in the original ECCV 2016 publication [BTF+16] (Appendix A.5) and with the supplemental video [2].

2. https://hal.inria.fr/hal-01348837/file/EigenAppearance.mp4

**FIGURE 4.7 –** *Interpolation examples using linear interpolation (left) and our pipeline (right). From left to right : Input frames, Interpolated models, and a close-up on the texture maps (top) and the rendered images (bottom).*

**FIGURE 4.8 –** *Completion examples. From left to right : Input and completed models, close-up on input and completed texture maps (top) and rendered images (bot- tom).*

# CHAPITRE 5

# Conclusion

In this chapter, we give a short recap of the contributions, discuss some insights, and mention some of the interesting followups related to this work that are happening in the team or that are envisioned as future work.

## 5.1 SUMMARY AND INSIGHTS

This manuscript presents some of the prominent works with my students, for PhDs defended in the last five years. With the general goal of advancing 4D modeling to digitize captured subjects, in the form of high quality time evolving representations estimated from multi-viewpoint video, we advanced the field in the following directions.

In chapter 2, we first examined estimating 3D shapes in the static case, with several simultaneous images inputs from one or several cameras. We have crafted a classic MVS pipeline tuned for performance capture applications, but also successful with more general data, then successfully demonstrated the benefit of learning the photo-consistency function and substituting it in this pipeline. We then extended our experience of learning-based approaches to end-to end monocular 3D models estimations, which is currently the basis for new works in the team for monocular sequence analysis and new multi-view reconstruction algorithms. All contributions presented became state of the art in reconstruction quality at the time of their publication and to this day serve as input to subsequent methods explored in the team. Some main limitations are the relatively slow training and processing times for *e.g.* Vincent Leroy's algorithms, due to the very fine level of detail targeted and particular training procedure, and this also drives are future research to improve efficiency.

In chapter 3, we have presented several principled motion models for 3D sequence tracking and alignment, using surface and volume rigidity priors, and volume support for improved data-driven shape matching, with state of the art results at the time of

publication. We also presented human-prior driven tracking under wide clothing. They are all grounded in the previous extraction of 3D models per time step discussed in the previous chapter. While these have been great achievements and have definitely contributed to advance understanding of motion models, improving the precision, robustness, in particular breaking the centimetric error barrier in broad acquisition situations, are still prominent challenges. Even better cooperation between the reconstruction, 3D sequence alignment and appearance modeling layers, over longer terms, are probably a key to break this deadlock, as well as better accounting for clothing in our models.

In chapter 4, we presented a principled approach to estimate appearance maps from multi-view sequences over small time windows, treating it as a generalized multi-camera superresolution problem. We further show how this estimation can be used a building block to build an Eigen space of varying textures to encode the appearance of a 4D sequence. These contributions were made possible by the advances with the surface reconstruction and alignment previously presented. One of the challenges facing us is to make the superresolution model more practical as it is currently very compute-intensive. We believe finer detail is to be accessed with better cooperation between the geometric and appearance stages as well. This has largely been the basis for new work with PhD student Matthieu Armando on new encodings for appearance on the shape [AFB19], and fine detail correction and restoration on the shape surface [AFB20], for which we are obtaining very encouraging results.

## 5.2 FUTURE DIRECTIONS

The work has brought me to very different fields and was rich with new discoveries, which helped forge some intuitions.

***Breaking stratified assumption and representations.***

Mostly one stratification approach was examined in this document, however others are possible. For example, some works examine the 3D motion problem and priors in trajectory space [AKSK08, SVMS14]. I think there are powerful priors to be learned by examining the problem in such orthogonal directions, not necessarily abandoning the existing geometric priors in fact, but using those in complementary fashion.

Learning offers new paths to consider input data and output of shape, motion and data jointly representation. Graph-convolution [MBM+17, VBV18] networks offer a promising avenue of research that we have began exploiting in the team, *e.g.* [AFB20]. Some methods also of great interest to us that redefine shape representation based on implicit networks [CAPM20] *e.g.* the recent occupancy flow works [NMOG19], which echo some of our own works to a similar goal [GFBP10], or NERF [MST+20]. Thsi is driving some of our current explorations, *e.g.* with PhDs student Mathieu Marsot and Boyao Zhou [ZFB+20].

***Better priors for motion and clothing.***

Modeling clothing is definitely becoming a hot topic, because our research is has hit the precision limits associated to body-only models. New models are increasingly accounting for the actually observed surface layer which sits on top of the other body and flesh layers. We have started examining this to some extent with Jinlong Yang's work, as presented in this document, with one of the team followups focused on predicting a vertex deformation layer of a template based on the underlying human motion parameters [YFHWW18]. We've personally made a low level-oriented effort in collaboration with Stefanie Wuhrer, and Physical and Structural Mechanics specialist Florence Bertails and Arnaud Lazarus, in co-supervizing PhD student Abdullah Haroon Rasheed, to better understand how cloth simulations could be used in vision systems to build new priors [RRBD+20].

One can measure the growing popularity of the topic with the number of papers and citations surrounding this work, and seeing interesting new work picking up where we left off with Jinlong Yan's thesis, *e.g.* [LPMG17, PMPHB17]. And I believe much more can be done in this direction.

***From old make new.***

It would be quite lame and self-evident to state that Deep Learning will continue reshaping our field, 4D modeling being no exception. But our experience, and the literature shows that one of the key limits fo Deep Learning models, explainability, can be at least partially mitigated by focusing the learning method on well understood sub-problems of time-tested classical pipelines, wit a well identified support domain, which has guided some of our earlier works, notably with the MVS photoconsistency function. While more and more methods are able to look at the whole problem end-to-end, I think this is still a valid approach to gain understandings on new idea and new problems.

Many times we can use the classic pipeline as a basis for a data-driven method to build on. This is in some sense what we were doing with our volumetric tracking by detection method in §3.2.4 [HAF+16] by alternating a data-driven discriminative detection step with a classic generative model-based error minimization. Interestingly, during her post-doc after her thesis with us in Morpheo, Vagia Tsiminaki demonstrated another way to achieve this, by using a network to compute the fine detail residual of appearance maps with respect to the generative appearance map model presented here [RCO+19]. These are all useful alternative ways to look at our problems.

Another aspect I am noticing and increasingly interested in integrating in my future learning approaches, is that some time time-tested pipeline structuring ideas that have proven to be essential for classic vision algorithms to work in practice and achieve top performance, such as coarse-to-fine processing, are finding their way in the way deep learning pipelines are structured. It is not a coincidence that coarse-to-fine pyramid networks are top scorers in today's benchmarks, such as PWC-net for optical flow [SYLK18] or for the MVS problem [YMAL20]. This is exciting as it offers new

opportunities to apply our classic vision knowledge and really put added value in how we craft future neural-network driven architectures.

# ANNEXE A

## Selected papers

I hereby provide a selection of my publications relevant to this manuscript, in chronological order.

— **High Resolution 3D Shape Texture from Multiple Videos**
Vagia Tsiminaki, Jean-Sébastien Franco, Edmond Boyer, CVPR 2014 - IEEE International Conference on Computer Vision and Pattern Recognition, Jun 2014, Columbus, OH, United States

— **On Mean Pose and Variability of 3D Deformable Models**
Benjamin Allain, Jean-Sébastien Franco, Edmond Boyer, Tony Tung, ECCV 2014 - European Conference on Computer Vision, Sep 2014, Zurich, Switzerland. Springer

— **An Efficient Volumetric Framework for Shape Tracking**
Benjamin Allain, Jean-Sébastien Franco, Edmond Boyer CVPR 2015 - IEEE International Conference on Computer Vision and Pattern Recognition, Jun 2015, Boston, United States.

— **Estimation of Human Body Shape in Motion with Wide Clothing**
Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, Stefanie Wuhrer European Conference on Computer Vision 2016, Oct 2016, Amsterdam, Netherlands

— **Eigen Appearance Maps of Dynamic Shapes**
Adnane Boukhayma, Vagia Tsiminaki, Jean-Sébastien Franco, Edmond Boyer ECCV 2016 - European Conference on Computer Vision, Oct 2016, Amsterdam, Netherlands.

— **Tracking-by-Detection of 3D Human Shapes : from Surfaces to Volumes**
Chun-Hao Huang, Benjamin Allain, Jean-Sébastien Franco, Nassir Navab, Slobodan Ilic, Edmond Boyer. IEEE Transactions on Pattern Analysis and Machine Intelligence, Institute of Electrical and Electronics Engineers, 2017,

— **Moulding Humans : Non-parametric 3D Human Shape Estimation from Single Images**
Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, Gregory Rogez. ICCV 2019 - International Conference on Computer Vision, Oct 2019, Seoul, South Korea. pp.1-10

— **Volume Sweeping : Learning Photoconsistency for Multi-View Shape Reconstruction**
Vincent Leroy, Jean-Sébastien Franco, Edmond Boyer. International Journal of Computer Vision, 2020

$\therefore$

## A.1 High Resolution 3D Shape Texture from Multiple Videos

Vagia Tsiminaki, Jean-Sébastien Franco, Edmond Boyer, CVPR 2014 - IEEE International Conference on Computer Vision and Pattern Recognition, Jun 2014, Columbus, OH, United States

# High Resolution 3D Shape Texture from Multiple Videos

Vagia Tsiminaki Jean-Sébastien Franco Edmond Boyer
Inria Grenoble Rhône-Alpes, LJK - Grenoble Universities, France
first.last@inria.fr

## Abstract

*We examine the problem of retrieving high resolution textures of objects observed in multiple videos under small object deformations. In the monocular case, the data redundancy necessary to reconstruct a high-resolution image stems from temporal accumulation. This has been vastly explored and is known as image super-resolution. On the other hand, a handful of methods have considered the texture of a static 3D object observed from several cameras, where the data redundancy is obtained through the different viewpoints. We introduce a unified framework to leverage both possibilities for the estimation of an object's high resolution texture. This framework uniformly deals with any related geometric variability introduced by the acquisition chain or by the evolution over time. To this goal we use 2D warps for all viewpoints and all temporal frames and a linear image formation model from texture to image space. Despite its simplicity, the method is able to successfully handle different views over space and time. As shown experimentally, it demonstrates the interest of temporal information to improve the texture quality. Additionally, we also show that our method outperforms state of the art multi-view super-resolution methods existing for the static case.*

## 1. Introduction

Gathering appearance information of objects through multi-camera observations is a challenging problem, of particular interest for multi-view capture systems. In such systems, typically, a geometric model is reconstructed, tracked or refined to be as close as possible to reality. Adding an appearance or texture layer to this geometric information plays an essential part in the realism of the result, and is often more important than geometric detail to convey the object's visual aspect. Applications of this acquisition pipeline, such as broadcast, special effects or entertainment, among others, are very highly demanding in terms of quality. Yet, even with state of the art multi-camera studio equipment, simply reprojecting texture from any one of the high resolution video streams used in the acquisition process is not enough to guarantee good texture coverage and high quality renderings or close-ups. Because several such input video streams are available in this context, and in order to take advantage of all the information they carry, we naturally turn to the various sources of data redundancy to boost texture quality. Following 2D monocular super-resolution techniques that successfully regain details from low resolution images, we consider here a similar framework for multi-viewpoint videos.

Figure 1. Input view $768 \times 576$ resolution with up-sampling by factor of three, BEETHOVEN dataset. Super-resolved $2304 \times 1728$ output of our algorithm rendered from identical viewpoint.

Such a framework is however significantly different from 2D super-resolution. First, dealing with multiple video streams is a different problem than using only one, where little parallax is usually assumed to occur. In a multi-view scenario, the intrinsic appearance of a single 3D object is only partially visible in each view, and observed only after being perspective projected, distorted by 3D geometry, and self-occluded. The 3D geometry itself is subject to reconstruction error and thus uncertain. Seamlessly blending and super-resolving the different input contributions into one single coherent texture space, while accounting for all such sources of variability is thus quite challenging. In fact it has only recently started to be addressed as such [10] for static objects.

But to fully exploit data redundancy, temporal accumulation of all views also needs to be examined. Not only is it an additional source of data, but interestingly temporal accumulation might make it possible to obtain high quality results with a sparser set of viewpoints than in the static case. This is not without its own source of difficulties. More often than not the subjects of interest are of arbitrarily deformable nature, such as human actors. This means that consistent temporal accumulation of texture data can only be done by realigning the relevant parts of the texture from one temporal frame to another, and accounting for sources of geometric variability. Fortunately, recent progress in non-rigid surface tracking methods [3] offer a path to resolve such issues, which we open with this work.

**Overview.** Generalizing existing multi-view appearance super-resolution work to the temporal domain requires a robust model of variability. As the appearance of subjects may drastically change over the long run, we focus in this paper on small non-rigid motions of the subject around a stable pose and observed appearance. We propose to deal with the largest non-rigid motion component using a surface-tracking method [3], and to compensate for any remaining geometry perturbations with a per-view, per-time frame warp. This per-view registration popularized in various rendering techniques [20, 8] has the large advantage of uniformly dealing with all sources of error, calibration, reconstruction, temporal misalignments and ghosting for our texture super-resolution, and is one of the major contributions of the paper. This paper is also the first, to our knowledge, to deal both with multiple viewpoints and temporal frames to build one common super-resolved texture, as opposed to [21] which enhance the input views directly, and [10] which only deals with the static multi-view aspect. Warping is done on an intermediate, high-resolution projected proxy of the model texture, where variability can be appropriately densely compensated (§3.1). We also expose a straightforward model and algorithm for this task, illustrated in Fig. 2. We notably show that some linear models [17] of the image formation can be generalized to the multi-view, multi-frame case (§3), as well as the monocular noise models (§4). We exhibit a two-stage iterative algorithm (§5), whose convergence is illustrated in experiments (§6). Our validation protocol also includes favorable comparison with the closest state of the art method [10], at the intersection of the validity domains of the methods (static, multi-view texture resolution case). Furthermore, we quantitatively demonstrate the convergence and temporal improvement of our method over using the same number of views in the static case.

## 2. Related Work

**View-Dependent Texturing.** Various strategies exist to retrieve and render the appearance of objects from input views and given a viewpoint, a geometric reconstruction being assumed available in general. One of the first proposed is to reproject and blend view contributions according to visibility and viewpoint-to-surface angle [7]. View-dependent techniques have been generalized to model and approximate the plenoptic function for the scene object, capturing view dependent shading effects [2] but this requires many dense views. Imperfect proxies and other geometric errors create rendering misalignments (ghosting), which various techniques correct with an additional image-space registration step [8], building a local basis of appearance variability [5], or refining the geometry proxy [19]. By nature, these methods are not targeted to capture intrinsic, view-independent texture properties and generally do not exploit viewpoint redundancy to super-resolve visual quality, nor do they easily extend to the time domain for deformable objects as proposed.

**Multi-View Texture Estimation.** To store intrinsic details of the acquired object and later render them, numerous methods build an image-based texture atlas to store appearance information, where each texel blends contributions from each view. Realignment is often proposed again to avoid ghosting [20, 15], but a second strategy exists which instead builds the texture as a mosaics of unique-view contributions, whose seam locations are optimized to minimize appearance change between fragments [14]. Interestingly, this strategy was extended to the temporal domain [13]. Only a handful of particularly relevant works examine how to super-resolve fine appearance detail from viewpoint redundancy at a single time frame [12, 10]. We propose an improved, unified model to deal with geometric variability due to reconstruction error and small deformation across time for multi-view super-resolution.

**Video Super-Resolution** While very few works exist concerning super-resolution techniques applied in a multi-view context, the problem has been extensively studied in the monocular case. The image formation model is well identified, as a geometric warping, blurring and sub-sampling process of the initial high-resolution image [1]. Two features of particular interest to us are that this model can be represented by a stack of linear transforms, and that Bayesian models have been developed to explicit the noise dependencies and priors over the target image and estimated warps [9, 17]. L1-norm based priors and total variation (TV-)minimization are increasingly popular [17] for their image restoration qualities. Notably, super-resolving multiple videos of a moving subject was examined in a performance capture context, but only for the input viewpoints [21]. Our model proposes temporal and multi-view super-resolution, yet super-resolves a single, intrinsic appearance map which can be re-used to render new viewpoints.

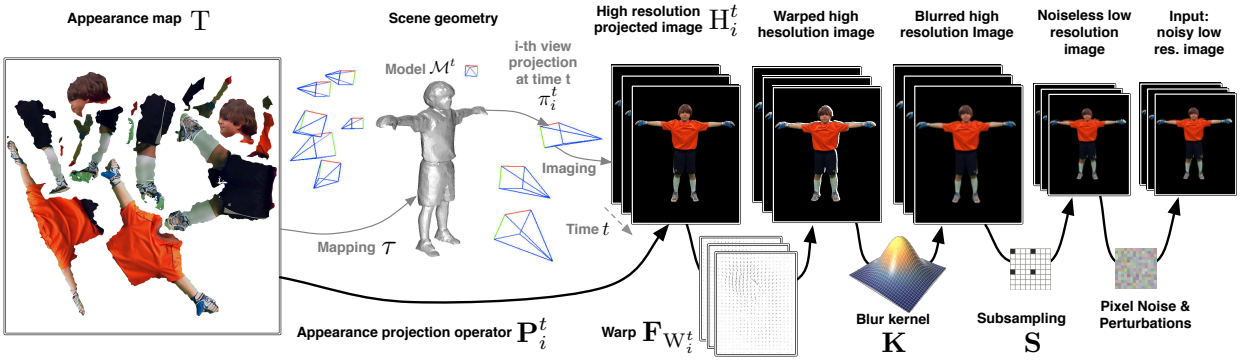



Figure 2. Summary of image formation model and problem notation.

## 3. Image Formation Model

Our goal is to estimate an appearance map $\mathrm{T}$ of an object of interest from a set of input color images $\{\mathrm{I}_i^t\}$, where $i \in \{1, \cdots, n_i\}$ is the camera number and $t \in \{1, \cdots, n_t\}$ the time. We assume a temporally coherent mesh model of the object, *i.e.* whose connectivity is time independent but of varying pose $\{\mathcal{M}^t\}$, obtained by tracking the surface tracking [3].

### 3.1. High Resolution Projection

We project the texture to a high resolution (HR) image $\{\mathrm{H}_i^t\}$ for each viewpoint $\{i, t\}$. Before reaching $\mathrm{H}_i^t$, the texels of $\mathrm{T}$ undergo two geometric transformations.

**Texture Mapping.** For the appearance to be mapped to the mesh, a geometric mapping function must map each texel of $\mathrm{T}$ to the mesh surface. Thanks to fixed connectivity of the mesh across time, only one such function $\tau$ needs to be defined. Conformal mappings are preferred, because preservation of angles ensure low distortion during the transfer, such that the texel density of $\mathrm{T}$ is kept homogeneous on the 3D surface. Note that due to potential cuts and non-zero genus topology of the objects of interest, $\tau$ may not be continuous and may have a support region with several connected components (or charts) in the texture domain. To obtain $\tau$, we use [18], which yields large charts with relatively few components, a useful feature for regularization and to avoid continuity artifacts.

**Projection to High Resolution Image.** We assume projections $\{\pi_i^t\}$ are known for each view $i$ and time $t$. A texel at texture location $x$ is mapped to a geometric point $\tau(x)$ on model $\mathcal{M}^t$, this point is then projected in view $\{i, t\}$ at point $\pi_i^t \circ \tau(x)$. This projection model is intended to provide a high resolution image space to be able to precisely compute a correction warp, which remaps the texture contributions with the matching content of input images $\mathrm{I}_i^t$. In particular we do not model any optical blur here; rather for each HR pixel $q$ we collect all texel contributions projecting within. Because calibration and 3D models are available, we can use GPU z-rendering to filter out non-visible texels [7]. Occasionally the density of projected texels is insufficient (*i.e.* in high curvature regions of the surface) for a pixel to receive any texture samples. In this situation we assume the underlying surface appearance perceived by this pixel is an interpolation of neighboring texels. For a uniform, continuous treatment of both cases, we combine all texel contributions falling in the vicinity of $q$ by a spatial Gaussian weight with small variance $\sigma_p^2$, and normalize to one the sum of texel contribution weights for a pixel $q$. The continuity of this scheme ensures that no artificial discontinuity is created as a result of a discrepancy in the treatment of these cases. This insures that samples present at the pixel contribute overwhelmingly when present at the center of the pixel, and that the pixel is computed as a weighted sum of texels further away otherwise.

Note that, with this formulation, HR pixels are a linear combination of texels of $\mathrm{T}$. Let $\mathbf{P}_i^t$ be the resulting sparse projection operator such that $\mathrm{H}_i^t = \mathbf{P}_i^t \mathrm{T}$, appropriately collecting the weights previously discussed after being mapped and projected in view $\{i, t\}$. Each $\mathbf{P}_i^t$ can be stored as a sparse matrix with $w_{\mathrm{H}_i^t} \times h_{\mathrm{H}_i^t}$ rows and $w_{\mathrm{T}} \times h_{\mathrm{T}}$ columns, respectively HR image resolution and the chosen texture resolution.

### 3.2. Inputs as Warped, Downsampled HR Images

To generate an input image from each HR image $\mathrm{H}_i^t$, the HR image is first warped according to the different apparent sources of variability impacting the input image - calibration error, distortion, model geometry error - using a dense warp field $\mathrm{W}_i^t$. This warp results in a linear operator over the HR image, which we note $\mathbf{F}_{\mathrm{W}_i^t}$. The image then traverses the optical system, where it is blurred and captured

by the CCD which performs light integration at every photosite. Following 2D super-resolution literature [1, 9] this is generally modeled using a Point Spread Function (PSF) with the form of a Gaussian blur kernel, followed by an image subsampling stage. Both operations can be written as linear operators, the image-wide blur operator $\mathbf{K}$ and subsampling operator $\mathbf{S}$, which are applied to the HR image to obtain a view's observed image $\mathrm{I}_i^t = \mathbf{SK}\mathrm{H}_i^t$. Remarkably, in its noiseless form, the full image formation model can thus be noted as a single, sparse linear operator $\mathbf{A}_i^t = \mathbf{SKF}_{\mathrm{W}_i^t}\mathbf{P}_i^t$ for each view $\{i,t\}$, with $w_{\mathrm{I}_i^t} \times h_{\mathrm{I}_i^t}$ rows and $w_{\mathrm{T}} \times h_{\mathrm{T}}$ columns, such that $\mathrm{I}_i^t = \mathbf{A}_i^t\mathrm{T}$ for each view $\{i,t\}$. This elegantly generalizes the linear formation models used in various 2D super-resolution models [17] to the 3D+t case.

## 4. Bayesian Generative Model

The linear model previously discussed describes how input pixels are obtained through warping and blending of texels in noiseless fashion. As in the 2D case, inverting the problem to estimate $\mathrm{T}$ and the warps $\mathrm{W}_i^t$ from $\mathrm{I}_i^t$ is ill-posed, non-convex, and noise ridden [1]. We thus introduce a noise model and priors for better problem conditioning, formulating the solution as a MAP estimation over $\mathrm{T}$, and the warps $\{\mathrm{W}_i^t\}$ for all views and temporal frames $\{i,t\}$:

$$\{\hat{\mathrm{T}}, \{\hat{\mathrm{W}}_i^t\}\} = \arg\max_{\mathrm{T},\{\mathrm{W}_i^t\}} p(\mathrm{T}, \{\mathrm{W}_i^t\}|\{\mathrm{I}_i^t\}), \quad (1)$$

where the posterior is a product of prior and likelihood:

$$p(\mathrm{T}, \{\mathrm{W}_i^t\}|\{\mathrm{I}_i^t\}) = p(\mathrm{T}) \prod_{i,t} p(\mathrm{W}_i^t) \prod_{i,t} p(\mathrm{I}_i^t|\mathrm{W}_i^t, \mathrm{T}). \quad (2)$$

**Prior Terms.** To ensure sparsity of variations of the estimated texture and warp, we impose minimal Total Variation (TV) constraints on appearance image $\mathrm{T}$ and each $\mathrm{W}_i^t$:

$$p(\mathrm{T}) = \frac{1}{Z_{\mathrm{T}}(\lambda)} e^{-\lambda\|\nabla \mathrm{T}\|}, \quad (3)$$

$$p(\mathrm{W}_i^t) = \frac{1}{Z_{\mathrm{W}}(\gamma)} e^{-\nu\left(\|\nabla u_i^t\| + \|\nabla v_i^t\|\right)}, \quad (4)$$

where $\nabla$ is the gradient operator, $\|\nabla \mathrm{T}\| = \sum_q \|\nabla \mathrm{T}(q)\| = \sum_q (\|\mathrm{T}_x(q)\| + \|\mathrm{T}_y(q)\|)$, the sum over pixel index $q$ of the $L_1$-norm over spatial image derivatives of $\mathrm{T}(q)$. The same definition holds for $u_i^t$ and $v_i^t$, the x- and y- components of the warp $\mathrm{W}_i^t$. $Z_{\mathrm{T}}(\lambda), Z_{\mathrm{W}}(\gamma)$ denote the normalization constants of both distributions.

The TV constraint ensures that $\mathrm{T}$ is treated as a natural image to be restored with sparse and preserved edges. However, a discontinuity between some neighboring object surface points can be created due to necessary cuts in the mesh unwrapping algorithm, leading some mapped texels to appear in different charts despite their proximity on the surface [13]. For such texels, we carefully compute gradients by computing the transform of axis directions as reprojected in the chart where surface neighbors were mapped [10, 11]. This minimizes discontinuities in treatment across chart boundaries in the estimation.

**Data Term.** Under the assumption that the noise is independent per pixel given the information about the texture, model and cameras, we impose a Gaussian prior for each frame $\{i,t\}$ :

$$p(\mathrm{I}_i^t|\mathrm{W}_i^t, \mathrm{T}) = \frac{1}{Z(\mathrm{D}_i^t)} e^{-(\mathrm{I}_i^t - \mathbf{A}_i^t\mathrm{T})^\top \mathrm{D}_i^t (\mathrm{I}_i^t - \mathbf{A}_i^t\mathrm{T})}, \quad (5)$$

where $\mathrm{D}_i^t$ is a diagonal covariance matrix introduced to allow different noise characteristics per pixel $q$, and $Z(\mathrm{D}_i^t)$ a normalization function of $\mathrm{D}_i^t$. In 2D super-resolution models, a single variance per input image is usually used, with the i.i.d. noise assumption [9]. However when acquiring appearance in the 3D case it is well known that contributions need to be modulated according to the angle $\theta_q$ between viewing vector and local surface normal [7]. This can be purposely identified in the generative model, where each diagonal element $\mathrm{d}(\theta_q)$ of $\mathrm{D}_i^t$ materializes the breadth of the underlying Gaussian predictive model and thus the confidence in the pixel. We set this value as a robust, conservative function of $\theta_q$ given in §6, which we assume fixed for the purpose of estimation, under small warp perturbations. Note that this is a valid assumption since visibility and grazing angles are generally stable, as we assume given the full poses of the model $\mathcal{M}^t$ for all frames.

## 5. Inference

Directly maximizing all variables in (1) is intrinsically hard and seldom done in the literature. We opt for a coordinate descent scheme, alternating between $\mathrm{T}$ and $\mathrm{W}_i^t$.

**Appearance Map.** We maximize (1) by minimizing its negative log, dropping all terms independent of $\mathrm{T}$:

$$\hat{\mathrm{T}} = \arg\min_{\mathrm{T}} \sum_{i,t} (\mathrm{I}_i^t - \mathbf{A}_i^t\mathrm{T})^\top \mathrm{D}_i^t (\mathrm{I}_i^t - \mathbf{A}_i^t\mathrm{T}) + \lambda\|\nabla \mathrm{T}\|, \quad (6)$$

where the data term develops to a weighted sum of per-pixel $L_2$-norms. Although not specifically using a robust data term norm here as opposed to some works, we nevertheless obtain excellent results enforcing robustness through the constant covariance matrix $\mathrm{D}_i^t$, as will be shown. Optimizing a $L_2$ data term with a TV-regularizer has been specifically studied [4], yielding a family of forward-backward splitting solvers whose implementation are available off-the-shelf [6]. Let us note $f_d(\mathrm{T})$ and $f_{\mathrm{TV}}(\mathrm{T})$ the data and

the TV-term. Forward-backward splitting is an iterative algorithm for estimating $\mathrm{T}$, alternating between computing a gradient update step and projection $\mathrm{prox}_{\gamma,f_{\mathrm{TV}}}$ which computes an implicit subgradient descent step for the TV-norm.

$$\mathrm{T}_{n+1} = \mathrm{prox}_{\gamma,f_{\mathrm{TV}}}\left(\mathrm{T}_n - \gamma\nabla f_d(\mathrm{T}_n)\right), \tag{7}$$

where $\gamma$ is a step-size parameter. Our re-weighted functional (6) only implies a modification of the gradient update with respect to the standard case, with $\nabla f_d(\mathrm{T}_n) = 2\mathbf{A}_i^{t\top}\mathrm{D}_i^t(\mathbf{A}_i^t\,\mathrm{T}_n - \mathrm{I}_i^t)$.

**Warp Estimates.** We independently estimate each $\mathrm{W}_i^t$ for an input view $\{i,t\}$. Minimizing the negative log of (1), and dropping all terms independent of $\mathrm{W}_i^t$ yields:

$$\begin{aligned}\hat{\mathrm{W}}_i^t = \arg\min_{\mathrm{W}_i^t} &\nu\left(\|\nabla u_i^t\| + \|\nabla v_i^t\|\right) \\ &+ (\mathrm{I}_i^t - \mathbf{S}\mathbf{K}\mathbf{F}_{\mathrm{W}_i^t}\mathbf{P}_i^t\mathrm{T})^\top \mathrm{D}_i^t(\mathrm{I}_i^t - \mathbf{S}\mathbf{K}\mathbf{F}_{\mathrm{W}_i^t}\mathbf{P}_i^t\mathrm{T}),\end{aligned} \tag{8}$$

which can be interpreted as a modified optical flow equation with a TV-regularizer, where the data term is re-weighted by $\mathrm{D}_i^t$. The intuition here is that the minimization favors the TV prior of sparse variation over the data term for untrustworthy pixels according to $\mathrm{D}_i^t$, and puts more relative importance on trying to follow data on reliable pixels. We opt for a similar strategy to [17] for solving this equation, and initialize the estimation of $\mathrm{W}_i^t$ with the result of a standard optical flow method [16], applied between $\mathrm{H}_i^t$ and an upsampled $\mathrm{I}_i^t$ at each iteration.

## 6. Experiments

We exhibit results with a MATLAB prototype implementation, and run experiments on a 16-core 2.4 GHz PC with 32GB RAM[1]. Our current implementation is mainly mono-thread, with the exception of the optical flow which we launch in 10 separate threads. To initialize the algorithm, we first use a small C++/OpenGL program to render visibility maps from texture to image space, then initialize the texture map with a simple weighted average of visible inputs. The visibility maps are also used to generate each projection matrix operator $\mathbf{P}_i^t$. We use the Optical Flow package from Liu *et al.* [16] for per-iteration optical flow initialization, and the UNLocBOX package [6] for the texture re-estimation in the loop. We use a threshold on the relative norm of the objective function (6) as stopping criterion, and observe convergence in 30 to 70 iterations for a given $\lambda$. The execution time of the algorithm is in the range of 30 minutes to an hour per iteration depending on the dataset, number of views and number of frames. These are not a good indication of the final achievable performance as many enhancements are possible, including making the flow and image update estimations massively parallel on a GPU, better inter-time flow bootstraps as suggested by [17], more compact data-structures, C++ inner loop.

[1] See video results at `http://hal.inria.fr/hal-00977755`

**Parameter values.** We set the Gaussian variances with $\sigma_p = 0.25$ and $\sigma_k = 0.1$, respectively for the projection weight and PSF kernel $\mathbf{K}$, for all datasets. Although these could be optimized alongside other parameters, we observe low sensitivity to these parameters when set in the $[0.1, 1]$ range. Higher values introduce over-blurring, while lower values tend to reveal the underlying discretization of the texture map ($\sigma_p < 0.1$) or the input image ($\sigma_k < 0.1$). We also fix the convergence parameters to $\gamma = 0.05$ and $\lambda = 5.10^{-4}$ for all experiments, using a second and third round of iterations with $\lambda = 5.10^{-5}$ and $\lambda = 5.10^{-6}$ to down-weigh TV-regularization and thus reveal higher frequency detail. We set $\mathrm{d}(\theta_q) = \frac{1}{C}e^{-s\tan\theta_q}$ as a faster approximation of a normal distribution over the angles of the perceived surface, and use $C$ to normalize these weight contributions to 1 among all pixels that see a common texel $x$ to obtain homogeneous weights among pixels in the data term $\sum_{i,q=\tau\circ\pi_i^t(x)}\mathrm{d}(\theta_q) = 1$. We use $s = 7\pi/16$ over all experiments. This weight is more conservative than the $\cos\theta_q$ weight usually used for blending in multi-view texturing techniques [7], and yields improved results in our experiments, as it downgrades unreliable contributions from surface points at a grazing angle.

### 6.1. Static Multi-View Comparison

We compare our model with the latest state of the art multi-view texture super-resolution technique of Goldlücke *et al.* [10]. As the latter does not deal with temporal sequences, the comparison is performed on the common applicability domain, *i.e.* static images, as shown Fig. 3. The authors provide a public dataset for three objects BEETHOVEN, BUNNY and BIRD, and kindly provided additional data on request, so we could reproduce the experiment in the closest possible setup. This included a high resolution output of their algorithm for the viewpoint originally reported per dataset in [10], to which we compare our high resolution output. We use the same super-resolution ratio of $3\times$ the input resolution for the texture and high resolution image domains. Respectively 108, 52 and 52 calibrated viewpoints were originally used at resolution $768\times576$. We have used identical views, and also use identical 3D models except for the BIRD dataset, for which we observed large reconstruction and silhouette reprojection artifacts on the model provided. In fairness we thus only provide crops in regions where the 3D model geometry is not significantly different.

It can be generally observed that our outputs provide lower noise levels and artifacts. This is particularly visible in the BUNNY dataset, in the ear region and shadow

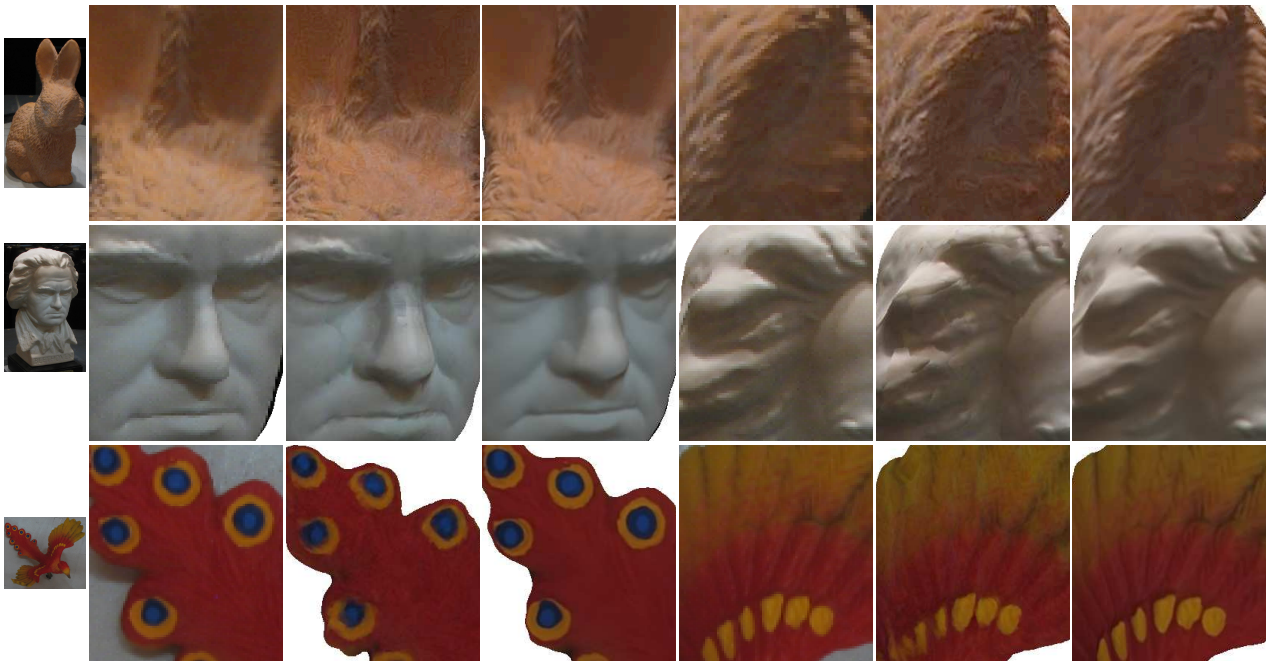


Figure 3. Comparison on BUNNY, BEETHOVEN and BIRD datasets. Left column: input images. Middle: output of [10]. Right: our algorithm. Best viewed magnified and in color.

region around the left eye. The BEETHOVEN exhibits some visibility difficulties due to the face geometry and presence of concave regions around the nose and hair, which generate artifacts for [10]. In contrast, our method is able to deal with these situations efficiently. A single texture domain cut is present on the nose but the discontinuity is barely visible thanks to the inter-chart terms we introduced. More accurate details and sharper pattern borders can be observed on the BIRD wing and tail, notably in the feather textures.

### 6.2. Temporal Superresolution Validation

To evaluate our approach on the temporal aspect, we introduce three synchronised datasets, GOALKEEPER, BACKPACK and ACTOR, in Fig. 4. The datasets were acquired with three different setups and camera models so as to maximize testing variability. All 3D models were obtained using silhouette-based reconstruction techniques and thus yield largely imperfect models. GOALKEEPER consists of 21 calibrated viewpoints at $1024 \times 1024$, which we downsample to $512 \times 512$ for the purpose of evaluation. BACKPACK consists of 15 viewpoints of a person, with resolution $1624 \times 1224$. ACTOR consists of 11 viewpoints in resolution $1920 \times 1080$. The ACTOR dataset is arguably the most difficult one, with lower views and higher noise levels both in the images and the reconstruction. We focus on small motions of the three subjects, and test the method for 2 to 7 frames. Significant improvements can be seen in the figure through temporal accumulation.

There are several difficulties in designing an experiment to quantify this improvement, such as the absence of ground truth data in texture space for real datasets. Synthetic datasets are less than ideal for image restoration and super-resolution problems: a significant conclusion can only be achieved if the different sources of variability are correctly introduced and simulated: sensor noise, calibration error, local reconstruction errors, specularities, temporal misalignments. Instead, we focus here on showing the temporal improvement by running our algorithm on a $2\times$ downsampled version of the GOALKEEPER dataset, and comparing our reprojected result with the higher resolution inputs using the mean squared error metric (MSE). Fig. 5 shows the result of this experiment, with convergence curves from one frame (static case) to three frames, and MSE's evaluated on the 21 input views. Several observations can be made from these curves. First, they illustrate convergence of the iterations toward the high resolution ground truth. Second the temporal improvement leveraged by our algorithm is validated in two forms: acceleration of the rate of convergence using more temporal frames, and improvement of the final result quality over using only one temporal frame.

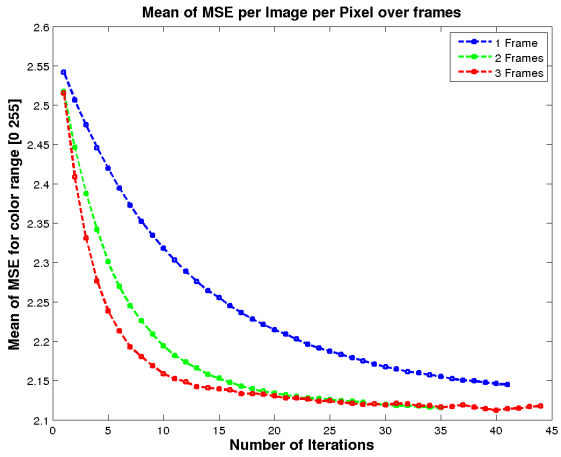



Figure 5. Results from GOALKEEPER dataset. We computed the mean value over frames of the Mean Square Error between our output and high resolution ground truth image. The resolution of input images is $512 \times 512$ and of the super-resolved output images is $1024 \times 1024$. We use a time step of 2 in experiments, corresponding to an acquisition frequency of 15Hz.

## 7. Discussion

We have presented a novel method to retrieve a single, coherent texture from several viewpoints and temporal frames of a deformable subject. The noiseless formation model introduced is linear from texture space to image space, and noise and regularization are achieved using a Bayesian framework. We have demonstrated the usefulness of this approach with respect to state of the art, and quantified the convergence and temporal improvement. The method opens several interesting research possibilities. First, more of the parameters and variability could be automatically learned, such as the projection parameters and regularization weight. The framework proposed enables this, with adapted convergence algorithms. Second, the trade-off between using more views or more temporal frames could be further explored to understand how each modality contributes to the result. Third, longer term resilience could be explored as an extension of this model.

## Acknowledgement

This work was funded by the Seventh Framework Programme EU project RE@CT (grant agreement no. 288369).

## References

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE PAMI*, 24(9):1167–1183, Sept. 2002.

[2] C. Buehler, M. Bosse, L. McMillan, S. J. Gortler, and M. F. Cohen. Unstructured lumigraph rendering. In *SIGGRAPH*, p. 425–432, 2001.

[3] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In *ECCV*, vol. 6314, p. 326–339, 2010.

[4] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20(1-2):89–97, Jan. 2004.

[5] D. Cobzas and M. Jägersand. Tracking and rendering using dynamic textures on geometric structure from motion. In *ECCV*, vol. 2351, p. 415–432, 2002.

[6] P. L. Combettes and J.-C. Pesquet. Proximal Splitting Methods in Signal Processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, p. 185–212. 2011.

[7] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In *SIGGRAPH*, p. 11–20, 1996.

[8] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. de Aguiar, N. Ahmed, C. Theobalt, and A. Sellent. Floating Textures. *Comp. Graph. Forum*, 27(2):409–418, Apr. 2008.

[9] R. Fransens, C. Strecha, and L. V. Gool. Optical flow based super-resolution: A probabilistic approach. *CVIU*, 106(1):106–115, 2007.

[10] B. Goldluecke, M. Aubry, K. Kolev, and D. Cremers. A super-resolution framework for high-accuracy multiview reconstruction. *IJCV*, 2013.

[11] B. Goldluecke and D. Cremers. A superresolution framework for high-accuracy multiview reconstruction. In *Pattern Recognition (Proc. DAGM)*, 2009.

[12] B. Goldluecke and D. Cremers. Superresolution texture maps for multiview reconstruction. *ICCV*, p. 1677–1684, Sept. 2009.

[13] Z. Janko and J.-P. Pons. Spatio-temporal image-based texture atlases for dynamic 3-D models. In *IEEE 3DIM*, p. 1646–1653, Oct. 2009.

[14] V. S. Lempitsky and D. V. Ivanov. Seamless mosaicing of image-based texture maps. In *CVPR*, 2007.

[15] H. P. A. Lensch, W. Heidrich, and H.-P. Seidel. A silhouette-based algorithm for texture registration and stitching. *Graphical Models*, 63(4):245–262, 2001.

[16] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, Massachusetts Institute of Technology, May 2009.

[17] C. Liu and D. Sun. A Bayesian approach to adaptive video super resolution. *CVPR*, p. 209–216, June 2011.

[18] A. Sheffer, B. Lévy, M. Mogilnitsky, and A. Bogom yakov. Abf++ : Fast and robust angle based flattening. *ACM Transactions on Graphics*, Apr 2005.

[19] T. Takai, A. Hilton, and T. Mastuyama. Harmonised texture mapping. In *3DPVT*, 2010.

[20] C. Theobalt, N. Ahmed, H. P. A. Lensch, M. A. Magnor, and H.-P. Seidel. Seeing people in different light-joint shape, motion, and reflectance capture. *IEEE Trans. Vis. Comput. Graph.*, 13(4):663–674, 2007.

[21] T. Tung. Simultaneous super-resolution and 3D video using graph-cuts. *CVPR*, p. 1–8, June 2008.

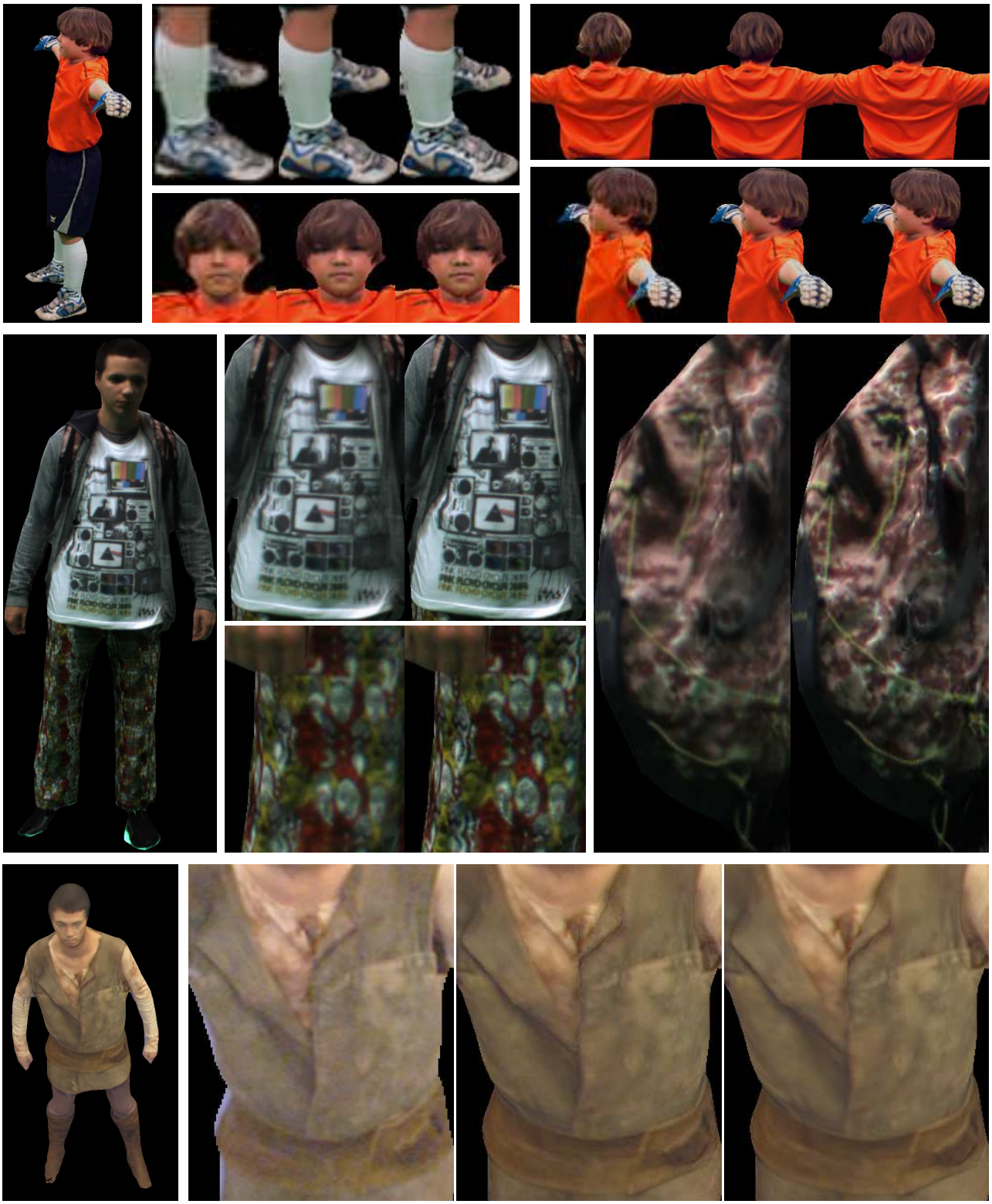


Figure 4. This figure illustrates various temporal improvements and detail enhancements obtained with various acquired datasets, comparing different convergence using one or several temporal frames. Top: GOALKEEPER dataset. Left: output of Frame 3. Input is compared to Frame 1 and Frame 3 for each close-up. Middle: BACKPACK dataset. Input on left, Frame 1 and Frame 2 comparisons for close-ups. Details are revived on the backpack, T-shirt and pants. Bottom: ACTOR; left to right: full result with three frames, close-up comparison between input, against Frame 1 and Frame 3. Best viewed magnified and in color.

$\therefore$

## A.2 On Mean Pose and Variability of 3D Deformable Models

Benjamin Allain, Jean-Sébastien Franco, Edmond Boyer, Tony Tung, ECCV 2014 - European Conference on Computer Vision, Sep 2014, Zurich, Switzerland. Springer

HAL
archives-ouvertes.fr

# On Mean Pose and Variability of 3D Deformable Models

Benjamin Allain, Jean-Sébastien Franco, Edmond Boyer, Tony Tung

▶ **To cite this version:**

Benjamin Allain, Jean-Sébastien Franco, Edmond Boyer, Tony Tung. On Mean Pose and Variability of 3D Deformable Models. ECCV 2014 - European Conference on Computer Vision, Sep 2014, Zurich, Switzerland. pp.284-297, 10.1007/978-3-319-10605-2_19 . hal-01016981

**HAL Id: hal-01016981**
**https://hal.inria.fr/hal-01016981**

Submitted on 2 Jul 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# On Mean Pose and Variability of 3D Deformable Models

Benjamin Allain[1], Jean-Sébastien Franco[1], Edmond Boyer[1], and Tony Tung[2]

[1] LJK, INRIA Grenoble Rhône-Alpes, France
[2] Graduate School of Informatics, Kyoto University, Japan
firstname.lastname@inria.fr, tony2ng@gmail.com

**Abstract.** We present a novel methodology for the analysis of complex object shapes in motion observed by multiple video cameras. In particular, we propose to learn local surface rigidity probabilities (i.e., deformations), and to estimate a mean pose over a temporal sequence. Local deformations can be used for rigidity-based dynamic surface segmentation, while a mean pose can be used as a sequence keyframe or a cluster prototype and has therefore numerous applications, such as motion synthesis or sequential alignment for compression or morphing. We take advantage of recent advances in surface tracking techniques to formulate a generative model of 3D temporal sequences using a probabilistic framework, which conditions shape fitting over all frames to a simple set of intrinsic surface rigidity properties. Surface tracking and rigidity variable estimation can then be formulated as an Expectation-Maximization inference problem and solved by alternatively minimizing two nested fixed point iterations. We show that this framework provides a new fundamental building block for various applications of shape analysis, and achieves comparable tracking performance to state of the art surface tracking techniques on real datasets, even compared to approaches using strong kinematic priors such as rigid skeletons.

**Keywords:** Shape dynamics, Motion analysis, Shape spaces

## 1 Introduction

Recent years have seen the emergence of many solutions for the capture of dynamic scenes, where a scene observed by several calibrated cameras is fully reconstructed from acquired videos using multiview stereo algorithms [24,12,1,20]. These techniques have many applications for media content production, interactive systems [2] and scene analysis [28] since they allow to recover both geometric and photometric information of objects' surface, and also their shape and evolution over time. Since these temporal evolutions were initially reconstructed as a sequence of topologically inconsistent 3D models, significant research work has been done for full 4D modeling and analysis of geometrically time-consistent 3D sequences.

In particular, several techniques propose to deform and match a template to either image data, or to intermediate 3D representations of the surface [25,17,9,26].

These methods allow the recovery of both shape and motion information. However they usually do not consider the intrinsic dynamic properties of a surface. These are either assumed, for instance through a kinematic structure (rigging) or through the surface tension parameters, or are simply ignored. Hence, there is a large interest in better understanding rigidity and motion properties of shapes, with the prospect of improving dynamic models, extracting more useful information, and better automation. In this work, we take the estimation a step further and investigate how to infer dynamics or statistical properties of shapes given temporal sequences.

Recovering this information is yet a largely open research topic with only few exploratory representations proposed for dynamics characteristics of surfaces, e.g. [11,29]. We propose a novel inference framework for the analysis of complex object shapes in motion that learns local surface rigidity probabilities (i.e., deformations), and estimates a mean pose over a temporal sequence. Based on recent advances in surface tracking techniques, we formulate a generative model of 3D temporal sequences using a probabilistic framework, which conditions shape fitting over all frames to a simple set of intrinsic surface rigidity properties. Surface tracking and rigidity variables can then be obtained iteratively using Expectation-Maximization inference by alternatively minimizing two nested fixed point iterations. Thus, our main contribution is a framework that allows the simultaneous tracking and inference of dynamic properties of object surfaces given temporal observations. We show how these properties contribute to a better understanding of surface motion and how they can be used for the dynamic analysis of 3D surface shapes through mean pose estimation and rigidity-based segmentation, while achieving competitive surface tracking.

The remainder of the paper is organized as follows. The next section discusses related work. Details on the mean pose inference model are given in Sect. 3. Section 4 presents various applications and experimental results. Section 5 concludes with a discussion on our contributions.

## 2 Related Work

The analysis of deformable surfaces captured by multi-video systems has gained lot of interest during the last decade due to the rapid progression of computer and image sensing technologies. We focus here on works that relate to dynamic properties of shapes.

*Kinematic structures.* Many popular tracking methods propose to rigidly constrain a model using an articulated structure, for instance a skeleton or a cage, which must be scaled and rigged to a 3D template, and optimally positioned through a sequence of models representing the observed subjects [4,30,17,19]. The template is usually deformed using a skinning technique, according to the optimized structure across the sequence [5]. Such kinematic structures provide intrinsic information on the associated shapes through their parameter evolutions (e.g. their averages can define a mean pose). These approaches require a

priori knowledge on the observed shapes, such as the topology and the rigid parts, and cannot be applied to arbitrary object shapes. Moreover global template deformation across time is subject to loss of local details such as cloth wrinkles and folds.

*Locally rigid structures.* The literature also contains several methods that relax the constraint on the shape structure using looser rigidity priors. A body of works consider deformations that preserve local intrinsic surface properties, e.g. isometric deformations [21,8,22,23]. Such properties relate to local rigidities, for instance in [31,32] local surface distorsions are constrained, however they are usually known priors. While efficient to register or match surfaces, intrinsic surface properties are not necessarily sufficient to track complex shapes such as human bodies. In that case, several approaches introduce local deformation models to drive surface evolutions. For instance, in [9], the observed surface is treated as a piece-wise body with locally rigid motions. We consider a similar model to represent surface deformations which is used to learn local rigidities as well as mean poses along with the tracking. Interestingly, recent approaches also in this category were proposed to characterize local surface deformations. In [11], the authors propose a probabilistic framework for rigid tracking and segmentation of dynamic surfaces where the rigid kinematic structure is learned along time sequences. Our framework does not assume such structure but learns instead local rigidities and mean poses. In [29], the authors model complex local deformation dynamics using linear dynamical systems by observing local curvature variations, using a shape index, and perform rigidity-based surface patch classification. The latter approach assumes surface alignment is given, in contrast to our proposed generative model that simultaneously performs surface tracking and local rigidity estimation.

*Shape Spaces.* Following the work of Kendall [18], a number of works consider shape spaces that characterize the configurations of a given set of points, the vertices of a mesh for instance. This has been used in medical imaging to estimate mean shapes through Procrustes analysis, e.g. [16]. In this case, the shape of the object is the geometrical information that remains when the pose (i.e., similarities) is filtered out. Thus Procrustes distances can be used to measure shape similarities and to estimate shape averages with Fréchet means. We follow here a different strategy where a shape space represents the poses of a single shape and where we estimate a mean pose instead of a mean shape. This relates to other works in this category that also consider shapes spaces to model shape poses with mesh representations. They can either be learned, e.g. [3,15] or defined a priori, e.g. [27] and are used to constrain mesh deformations when creating realistic animations [3,27] or estimating shape and poses from images[15]. While sharing similarities in the deformation model we consider, our objective is not only to recover meaningful shape poses but also to measure pose similarities and intrinsic shape properties. Unlike [3,15], we do not need a pose or shape database and the associated hypothesis of its representativeness. Moreover, our methodology specifically addresses robust temporal window integration.

## 3 Mean Pose Inference Model

**Fig. 1.** Example of patch template used.

We assume given a temporal sequence of 3D reconstructions, incoherent meshes or point clouds, obtained using a multi-view reconstruction approach, e.g. [12,1,20]. We also assume that a template mesh model of the scene is available, e.g. a particular instance within the reconstructed sequence under consideration. The problem of local surface rigidity and mean pose analysis is then tackled through the simultaneous tracking and intrinsic parameter estimation of the template model. We embed intrinsic motion parameters (e.g. rigidities) in the model, which control the motion behavior of the object surface. This implies that the estimation algorithm is necessarily performed over a sub-sequence of frames, as opposed to most existing surface tracking methods which in effect implement tracking through iterated single-frame pose estimation. We first describe in details the geometric model (§3.1) illustrated with Fig. 1, and its associated average deformation parameterization for the observed surface (§3.2). Second, we describe how this surface generates noisy measurements with an appropriate Bayesian generative model (§3.3). We then show how to perform estimation over the sequence through Expectation-Maximization (§3.4).

### 3.1 Shape Space Parameterization

To express non-rigid deformability of shapes, while de-correlating the resolution of deformation parameters from mesh resolution, we opt for a patch-based parameterization of the surface similar to [9]. The reference mesh is partitioned in an overlapping set of patches, pre-computed by geodesic clustering of vertices. Each patch $P_k$ is associated to a rigid transformation $\mathbf{T}_k^t \in SE(3)$ at every time $t$. Each position $\mathbf{x}_{k,v}$ of a mesh vertex $v$ as predicted by the transform of $P_k$ can then be computed from its template position $\mathbf{x}_v^0$ as follows:

$$\mathbf{x}_{k,v} = \mathbf{T}_k(\mathbf{x}_v^0). \tag{1}$$

We thus define a *pose* of the shape space as the set of patch transforms $\mathbf{T} = \{\mathbf{T}_k\}_{k\in\mathcal{K}}$ that express a given mesh deformation. Note here that a pose in the shape space does not necessarily correspond to a proper geometric realization of the reference mesh and, in practice, patch deformations are merged on the template to preserve the mesh consistency.

### 3.2 Mean Pose

To retrieve the mean pose of a given sequence, we provide a definition suitable for the analysis of complex temporal mesh sequences. Following Fréchet's definition of a mean [13], we introduce the *mean pose* $\bar{\mathbf{T}}$ of a given set of poses $\{\mathbf{T}^t\}_{t\in\mathcal{T}}$ over the time sequence $\mathcal{T}$ as the pose minimizing the sum of squared distances to all poses in the set:

$$\bar{\mathbf{T}} = \arg\min_{\mathbf{T}} \sum_{t\in\mathcal{T}} d^2(\mathbf{T}, \mathbf{T}^t), \tag{2}$$

where $d()$ is a distance that measures the similarity of two poses. This distance should evaluate the non-rigidity of the transformation between two poses of a shape and hence should be independent of any global pose. Such a distance is not easily defined in the non-Euclidean shape space spanned by the rigid motion parameters of the patches. However using the Euclidean embedding provided by the mesh representation, we can define a proper metric based on the vertex positions. Inspired by the deformation energy proposed by Botsch *et al.* [7] our distance is expressed as an internal deformation energy between two poses. Let $\mathbf{T}^i$ and $\mathbf{T}^j$ be two poses of the model, the distance can be written as a sum of per patch pair squared distances:

$$d^2(\mathbf{T}^i, \mathbf{T}^j) = \sum_{(P_k,P_l)\in\mathcal{N}} d^2_{kl}(\mathbf{T}^i, \mathbf{T}^j), \tag{3}$$

$$\text{with } d^2_{kl}(\mathbf{T}^i, \mathbf{T}^j) = \sum_{v\in P_k\cup P_l} \|\mathbf{T}^i_{k-l}(\mathbf{x}^0_v) - \mathbf{T}^j_{k-l}(\mathbf{x}^0_v)\|^2, \tag{4}$$

where $\mathbf{T}^i_{k-l} = \mathbf{T}^{i\,-1}_l \circ \mathbf{T}^i_k$ is the relative transformation between patches $P_k$ and $P_l$ for pose $i$, and $\mathcal{N}$ is the set of neighboring patch pairs on the surface. The distance sums, for every pair of patches of the deformable model, its rigid deviation from pose $i$ to $j$. This deviation is given by the sum over each vertex $v$ belonging to the patch pair, of the discrepancy of relative positions of the vertex as displaced by $P_k$ and $P_l$. It can be verified that $d^2$ defines a distance as it inherits this property from the $\mathrm{L}^2$ norm used between vertices.

### 3.3 Generative Model

The expression (2) is useful to characterize the mean over a set of poses *already known*. Our goal however is to estimate this mean in the context where such

poses are indirectly observed through a set of noisy and sparse 3D point clouds of the surface. Thus we cast the problem as the joint estimation of mean pose and fitting of the model to each set of observations. For our purposes, we assume the set of poses $\{\mathbf{T}^t\}_{t\in\mathcal{T}}$ are defined for a set $\mathcal{T}$ corresponding to observations in a temporal sequence. The observed point clouds are noted $\mathbf{Y} = \{\mathbf{Y}^t\}_{t\in\mathcal{T}}$, where $\mathbf{Y}^t = \{\mathbf{y}_o^t\}_{o\in\mathcal{O}_t}$ is the set of point coordinates $\mathbf{y}_o^t$ for an observation $o$ among the set of observations $\mathcal{O}_t$ at time $t$. Note that this set $\mathcal{O}_t$ is different than $\mathcal{V}$ in general as it is obtained from a 3D reconstruction or depth camera, without any direct correspondence to the deformable shape surface model earlier defined.

To express the noisy predictions of observations, we follow the principle of EM-ICP [14] by introducing a set of assignment variables $\mathrm{k}_o^t$ indicating, for each observation $o$, which patch this observation is assigned to. We are also interested in retrieving information about the variations of the rigid deformation with respect to the mean shape. To keep this information in its simplest form, we express in the generative model that each pair of patches $(k,l) \in \mathcal{N}$ is assigned a *binary rigidity variable* $\mathrm{c}_{kl} \in \{0,1\}$, which will condition the patch pair to accordingly be rigid or flexible. This variable is an intrinsic parameter attached to the original deformable model and is thus time-independent. We note the full set of rigidity variables $\mathrm{C} = \{\mathrm{c}_{kl}\}_{(k,l)\in\mathcal{N}}$. This in turn will allow during inference the estimation of a rigid coupling probability for each patch pair $(k,l)$. We express the generative model through the following joint probability distribution:

$$p(\bar{\mathbf{T}}, \mathbf{T}, \mathbf{Y}, \mathrm{C}, \mathrm{K}, \sigma) = p(\bar{\mathbf{T}}) \prod_{t\in\mathcal{T}} \left( p(\mathbf{T}^t \mid \bar{\mathbf{T}}, \mathrm{C}) \prod_{o\in\mathcal{O}_t} p(\mathbf{y}_o^t \mid \mathrm{k}_o^t,\ \mathbf{T}^t, \sigma^t) \right), \quad (5)$$

with $\sigma = \{\sigma^t\}_{t\in\mathcal{T}}$ the set of noise parameters of the observation prediction model, and $\mathrm{K} = \{\mathrm{k}_o^t\}$ the set of all patch selection variables.

**Observation prediction model.** Each observation's point measurement is predicted from the closest vertex $v$ within patch $P_{k=\mathrm{k}_o^t}$. Because the prediction is noisy, this prediction is perturbed by Gaussian noise of variance ${\sigma^t}^2$:

$$p(\mathbf{y}_o^t \mid \mathrm{k}_o^t\ \mathbf{T}^t \sigma^t) \ = \ \mathcal{N}(\mathbf{y}_o^t \mid \mathbf{T}_{\mathrm{k}_o^t}^t(\mathbf{x}_v^0), \sigma^t). \quad (6)$$

**Pose constraining model.** We constrain the fitted poses to be close to the mean pose, using the distance defined earlier (3). We embed the influence of rigidity variables in this term, by computing two versions of the distance, biased by rigidity variables C:

$$p(\mathbf{T}^t \mid \bar{\mathbf{T}}, \mathrm{C}) \ \propto \ \exp\left( - \sum_{(k,l)\in\mathcal{N}} d_{kl}^2(\bar{\mathbf{T}}, \mathbf{T}^t, \mathrm{c}_{kl}) \right), \quad (7)$$

$$\text{where } d_{kl}^2(\bar{\mathbf{T}}, \mathbf{T}^t, \mathrm{c}_{kl}) = \sum_{v\in P_k \cup P_l} \beta_{kl}(v, \mathrm{c}_{kl})\ \|\mathbf{T}_{k-l}^i(\mathbf{x}_v^0) - \mathbf{T}_{k-l}^j(\mathbf{x}_v^0)\|^2, \quad (8)$$

with $\beta_{kl}(v, \mathrm{c}_{kl})$ a uniform function over all vertices of the patch pair if $\mathrm{c}_{kl} = 1$, which encourages common rigid behavior of the two patches, and a non-uniform function encouraging more elasticity when $\mathrm{c}_{kl} = 0$:

$$\beta_{kl}(v, 0) \propto \exp(-\frac{b_{kl}(v)}{\eta \bar{D}}), \tag{9}$$

where $b_{kl}(v)$ is the distance between the vertex $v$ and the border between $P_k$ and $P_l$ on the template, $\bar{D}$ is the average patch diameter and $\eta$ is a global coefficient controlling the flexibility. The $\beta_{kl}(\cdot, 0)$ has larger values on the border between the patches, which allows more flexibility while enforcing continuity between the patches. The coefficients $\beta_{kl}(v, 0)$ are normalized such that $\sum_{P_k \cup P_l} \beta_{kl}(v, 0) = \sum_{P_k \cup P_l} \beta_{kl}(v, 1)$ in order to make both modes as competitive.

**Mean model prior.** In the absence of any prior, the mean pose is unconstrained and could theoretically have completely loose patches unrelated to each other. To avoid this and give the mean pose a plausible deformation, we consider the following a prior which expresses that the intrinsic mean pose should not significantly deviate from the original reference pose (represented by the identity transform **Id**):

$$p(\bar{\mathbf{T}}) \propto \exp(-d^2(\bar{\mathbf{T}}, \mathbf{Id})) \propto \exp\left(\sum_{(P_k, P_l) \in \mathcal{N}} \sum_{v \in P_k \cup P_l} \|\bar{\mathbf{T}}_k(\mathbf{x}_v^0) - \bar{\mathbf{T}}_l(\mathbf{x}_v^0)\|^2\right), \tag{10}$$

### 3.4 Expectation-Maximization Inference

We apply Expectation-Maximization [10] to compute Maximum A Posteriori (MAP) estimates of the tracking and average shape parameters given noisy 3D measurements, using the joint probability described in (5) as described in [6]. The assignment variables K and rigidity coupling variables C are treated as latent variables, which we group by the name $Z = \{\mathrm{K}, \mathrm{C}\}$. For the purpose of clarity let us also rename all parameters to estimate as $\Theta = \{\bar{\mathbf{T}}, \mathbf{T}, \sigma\}$. Expectation-Maximization consists in iteratively maximizing the following auxiliary function $Q$ given the knowledge of the previous parameter estimate $\Theta^m$:

$$\Theta^{m+1} = \arg\max_{\Theta} Q(\Theta|\Theta^m) = \arg\max_{\Theta} \sum_{Z} p(Z|\mathbf{Y}, \Theta^m) \ln p(\mathbf{Y}, Z|\Theta). \tag{11}$$

The **E-Step** consists in computing the posterior distribution $p(Z|\mathbf{Y}, \Theta^m)$ of latent variables given observations and the previous estimate. It can be noted given the form of (5) that all latent variables are individually independent under this posterior according to the D-separation criterion [6], thus following the

factorization of the joint probability distribution:

$$p(\mathbf{Y}, Z|\Theta^m) = \prod_{t\in\mathcal{T}} \left( \prod_{(k,l)\in\mathcal{N}} p(\mathrm{c}_{kl}|\Theta^m) \prod_{o\in\mathcal{O}_t} p(\mathrm{k}_o^t|\mathbf{Y}, \Theta^m) \right), \tag{12}$$

$$\text{where } \; p(\mathrm{c}_{kl}|\Theta^m) = a \cdot \exp\left( - \sum_{v\in P_k \cup P_l} -d_{kl}^2(\mathbf{T}^{t,m}, \bar{\mathbf{T}}^m, \mathrm{c}_{kl}) \right) \tag{13}$$

$$\text{and } \; p(\mathrm{k}_o^t|\mathbf{Y}, \Theta^m) = b \cdot \mathcal{N}(\mathbf{y}_o^t \mid \mathbf{T}_{\mathrm{k}_o^t}^{t,m}(v), \sigma^{t,m}), \tag{14}$$

where $a, b$ are normalization constants ensuring the respective distributions sum to 1, and $v$ is the closest vertex on patch $k$. Equations (13) and (14) are the E-step updates that need to be computed at every iteration for every latent variable. (13) corresponds to a reevaluation of probabilities of rigid coupling between patches, based on the previous $m$-th estimates of temporal and mean poses. (14) corresponds to the probability assignment table of time $t$'s observation $o$ to each patch in the model. This corresponds to the soft matching term commonly found in EM-ICP methods [14].

The **M-Step** maximizes expression (11), which can be shown to factorize similarly to (5) and (12), in a sum of three maximizable independent groups of terms, leading to the following updates:

$$\begin{aligned} \mathbf{T}^{t,m+1} = \arg\min_{\mathbf{T}^t} & \sum_{(k,l)\in\mathcal{N}} \sum_{\mathrm{c}_{kl}} p(\mathrm{c}_{kl}|\Theta^m)\; d_{kl}^2(\bar{\mathbf{T}}^m, \mathbf{T}^t, \mathrm{c}_{kl}) \\ & + \sum_{o\in\mathcal{O}_t} \sum_{\mathrm{k}_o^t} p(\mathrm{k}_o^t|\mathbf{Y}, \Theta^m) \|\mathbf{y}_o^t - \mathbf{T}_{\mathrm{k}_o^t}^t(\mathbf{x}_v^t)\|^2, \end{aligned} \tag{15}$$

$$\sigma^{t,m+1^2} = \frac{1}{3} \frac{\sum_{o\in\mathcal{O}_t} \sum_{\mathrm{k}_o^t} p(\mathrm{k}_o^t|\mathbf{Y}, \Theta^m)\; \|\mathbf{y}_o^t - \mathbf{T}_{\mathrm{k}_o^t}^{t,m+1}(\mathbf{x}_v^t)\|^2}{\sum_{o\in\mathcal{O}_t} \sum_{\mathrm{k}_o^t} p(\mathrm{k}_o^t|\mathbf{Y}, \Theta^m)}, \tag{16}$$

$$\bar{\mathbf{T}}^{m+1} = \arg\min_{\bar{\mathbf{T}}} \; d^2(\bar{\mathbf{T}}, \mathbf{Id}) + \sum_{t\in\mathcal{T}} \sum_{(k,l)\in\mathcal{N}} \sum_{\mathrm{c}_{kl}} p(\mathrm{c}_{kl}|\Theta^m) d_{kl}^2(\bar{\mathbf{T}}, \mathbf{T}^{t,m+1}, \mathrm{c}_{kl}). \tag{17}$$

Expression (15) corresponds to simultaneous updates of all patch transformations for a given time $t$, weighed by E-step probabilities. (16) updates the per-time frame noise parameter with an E-step weighed contribution of each observation. (17) computes the mean pose, accounting for all poses in the time sequence. Note that, for ease of resolution, we decouple the estimation of $\mathbf{T}^{t,m+1}$ and $\bar{\mathbf{T}}^{m+1}$, which is why (17) uses the result $\mathbf{T}^{t,m+1}$. We solve both systems with Gauss-Newton iterations, using a parametrization of the rigid transforms as a rotation matrix and translation.

## 4 Experiments

We evaluate the proposed generative model using 3D sequences reconstructed from real human performances captured by multiple view videos. We propose

two datasets, GoalKeeper and Dancer, which provide two different actions and clothing situations with high resolution inputs. These were processed by extracting visual hull reconstructions, and two neutral topology frames were selected to provide the template model after smoothing and simplifying the obtained mesh down to $5k$ vertices. Additionally, we also validate using two public datasets made available by the community. The Free [25] dataset consists of a photocoherent mesh sequence of a dancer with approximately $135k$ vertices per frame, exhibiting particularly fast and difficult dancing motion. The Marker dataset [19] provides another type of challenging situation with a two-person sequence of reconstructions, with martial art motions. It also provides markers on one of the persons which we will use for quantitative evaluation. For both these public sequences, we use the templates provided downsampled to $5k$ vertices.

In all visualizations, we render mesh poses by computing vertex position $\mathbf{x}_v^t$ at time $t$ as a linear blend of positions $\mathbf{x}_k^t$ of expression (1), weighed by a set of Gaussian weights $\alpha_k(v)$ materializing the region of influence of patch $P_k$ on the mesh. These weights are maximal at the center of mass of $P_k$ and their sum over all non-zero patch influences are normalized to 1 for a given vertex $v$:

$$\mathbf{x}_v^t = \sum_k \alpha_k(v)\, \mathbf{x}_k^t \ . \tag{18}$$

We visualize the rigidity coupling probabilities over the surface with heat-colored probabilities, by diffusing this probability over vertices of influence of patch pairs to obtain a smooth rendering. We provide a supplemental video[3] with the processed results for these datasets.

### 4.1 Tracking Evaluation

We first evaluate the tracking performance of the algorithm. Full sequences may be processed but because of the motion of subjects in the sequence, all poses of the sequence cannot be initialized with a single static pose, as this would surely be susceptible to local minima. We thus process the four datasets using a sliding window strategy for $\mathcal{T}$, where processing starts with a single pose, then additional poses are introduced in the time window after the previous window converges. We provide tracking results with sliding window size 20 which corresponds to approximately one second of video. We show the resulting poses estimated by our algorithm on the four datasets in Fig. 4, Fig. 5a and Fig. 5b. Runtime is approximately 15 seconds per time step on a recent workstation and can be further improved.

We also provide a comparison with state of the art methods Liu *et al.* [19] and with a purely patch-based strategy [9], on the Marker dataset. We reproduce [9] results by neutralizing mean updates and rigid coupling updates from our method, which corresponds to removing these terms from the energy and closely mimics [9]. Note that [19] is a kinematic tracking strategy, where both subjects are rigged to a kinematic skeleton providing a strong, fixed and dataset specific

[3] `http://hal.inria.fr/hal-01016981`

rigidity prior. On the other hand, [9] only use patch rigidity and inter-patch elasticity priors, that are weaker than [19] and our method. The MARKER dataset provides sparse marker positions, at which we estimate geometric positional error with respect to the surface. To this purpose we match the closest vertex on the template model provided, and follow it with the different methods, computing geometric errors in position with respect to the corresponding marker's position in these frames. The average errors are shown in Table 1. We also provide a temporal error graph for our method and [9] in Fig. 2.

**Table 1.** Mean error and standard deviation over the sequence of the MARKER dataset.

| method | mean error (mm) | standard deviation (mm) |
|---|---|---|
| no coupling, no mean pose [9] | 55.11 | 48.02 |
| our method | 43.22 | 29.58 |
| Liu *et al.* [19] | 29.61 | 25.50 |

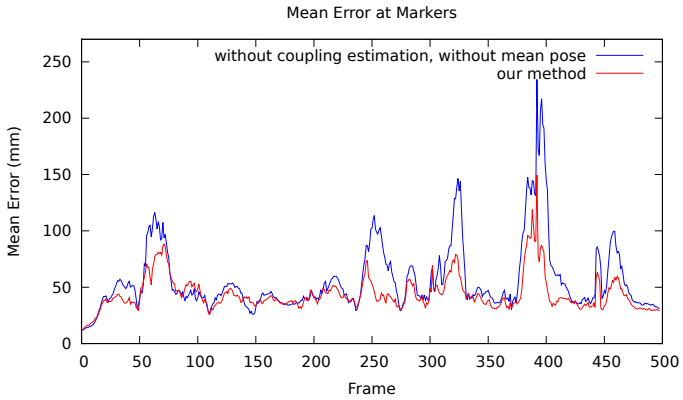
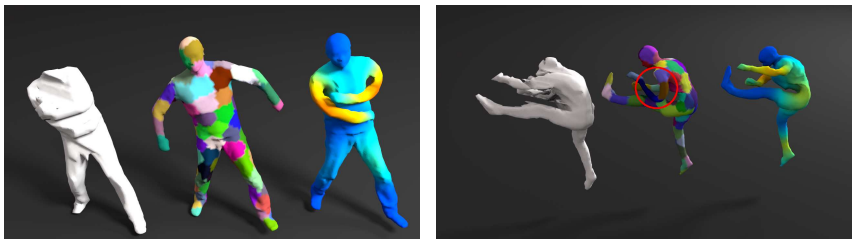



**Fig. 2.** Mean error for temporal evolution over MARKER dataset.

Table 1 shows our method achieves comparable tracking performance to state of the art surface tracking techniques. The slightly higher error with respect to [19] is not unexpected given that they use a stronger kinematic skeleton prior. Regarding [9], the graph and table show a small advantage in error for our method along the sequence, as well as a smaller variance of the error, showing the better constraining provided by our framework. The graph also shows significantly higher error values with [9] than with our method around frames 60, 250, 325, 390 and 460. These error peaks are imputable to difficult segments of the input sequence where [9] loses track of limbs (see Fig. 3a and Fig. 3b) while our method does not. The high error values around frame 390 are due to ambiguous

input meshes where the head of the second character (not seen in Fig. 3b) is out of the field of view. Around this frame, our method still outperforms [9] which misaligns an arm (see Fig. 3b). These results substantiate stronger robustness for our method over [9].

Regarding limitations, the model may fall into local minima when the noise level of inputs is too high similarly to all patch-based methods but this was not a strong limitation on the datasets. As the model favours rigidity and isometric surface deformations, the surface sometimes overfolds in non-rigid sections (as sometimes seen in video), which we will address in future work.



(a) Frame 325 (b) Frame 390

**Fig. 3.** Input mesh (left), tracked mesh with [9] (middle) and with our method (right).

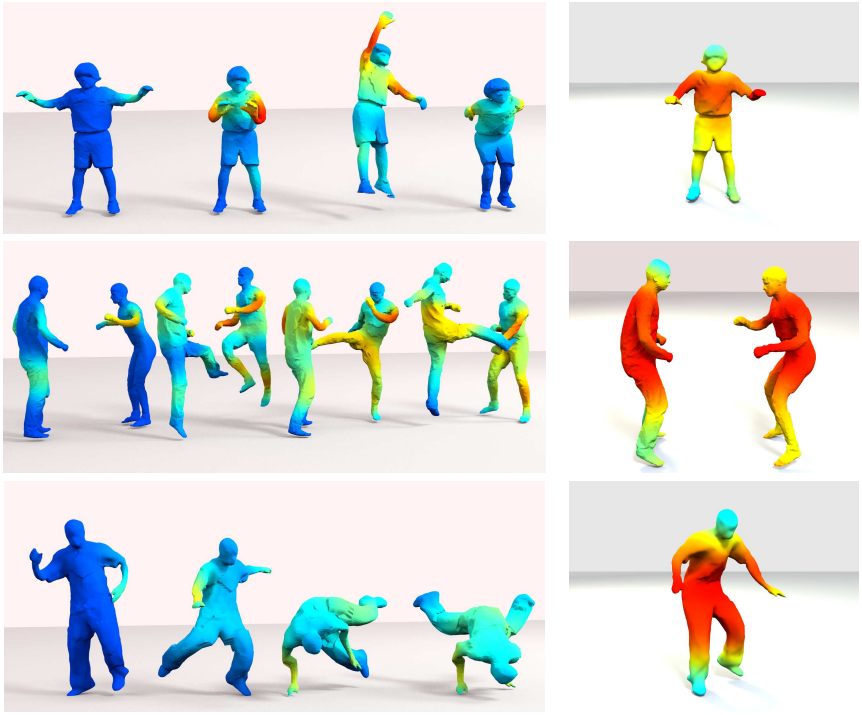
### 4.2 Mean Pose and Rigidity Estimation

Fig. 5a shows tracking results with color coded rigidity coupling probabilities with sliding window size 20. The method accurately reports instantaneous rigidity deviation, such as when the subject folds his elbows or shoulders. Blue regions correspond to regions of the mesh that have no non-rigid distortion with respect to the estimated mean pose. Fig. 5b shows estimates of mean poses for full sequences, colored with the estimated rigidity coupling probabilities over full sequence (no sliding window). It can be noted that the method accurately reports where the most common deviations occur.

The supplemental video shows mean pose sequences for several sliding window sizes. We observe a temporal smoothing of the initial deformation: fast deformation is filtered out. This effect is stronger with wide windows. We interpret this phenomenon as follows: when the temporal window slides along the sequence, it produces a mean pose sequence analogous to the convolution of the estimated pose sequence with a gate function, with the same size as the window size. This process can be seen as a low-pass filtering of the sequence poses.

We also observe that the mean pose is not affected by global rigid motion of the shape (noticeable with the DANCER dataset). This is an expected consequence of using a pose distance that is invariant under global rigid transforms in (2).

**Fig. 4.** Tracking excerpts from the DANCER dataset. Colors code patches.



(a) Tracking Excerpts.

(b) Mean poses computed on full sequences.

**Fig. 5.** Tracking excerpts from GOALKEEPER, MARKER and FREE datasets. Best viewed in color. Please watch supplemental video for more visualizations.

## 5 Conclusions

We present a novel methodology for the analysis of complex object shapes in motion observed by multiple cameras. In particular, we propose a generative model of 3D temporal sequences using a probabilistic framework that simultaneously learns local surface rigidity probabilities and estimates a mean pose over temporal sequence. Hence, rigidity-based surface segmentation can be achieved using local deformation properties, while motion synthesis or surface alignment for compression or morphing applications can be achieved using a mean pose as a sequence keyframe or a cluster prototype.

Our model can also perform surface tracking with state of the art performance, and does not require a priori rigid (kinematic) structure, nor prior model learning from a database. Surface tracking and rigidity variable probabilities are obtained by solving an Expectation-Maximization inference problem which alternatively minimizes two nested fixed point iterations.

To our knowledge, this is the first model that achieves simultaneous estimation of mean pose, local rigidity, and surface tracking. Experimental results on real datasets show the numerous potential applications of the proposed framework for complex shape analysis of 3D sequences.

## Acknowledgements

This work was funded by the Seventh Framework Programme EU project RE@CT (grant agreement no. 288369). It was also supported in part by the INRIA-JSPS Bilateral Program AYAME 146121400001 and the JSPS WAKATE B 26730089.

## References

1. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. ACM Transactions on Graphics 27(3) (2008)
2. Allard, J., Ménier, C., Raffin, B., Boyer, E., Faure, F.: Grimage: Markerless 3d interactions. SIGGRAPH - Emerging Technologies (2007)
3. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: Shape completion and animation of people. ACM Transactions on Graphics 24(3) (2005)
4. Ballan, L., Cortelazzo, G.M.: Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. 3DPVT (2008)
5. Baran, I., Popović, J.: Automatic rigging and animation of 3D characters. ACM Transactions on Graphics 26(3), 72:1–72:8 (2007)
6. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
7. Botsch, M., Pauly, M., Wicke, M., Gross, M.: Adaptive space deformations based on rigid cells. Comput. Graph. Forum 26(3), 339–347 (2007)
8. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Efficient computation of isometry-invariant distances between surfaces. SIAM Journal on Scientific Computing 28 (2006)

9. Cagniart, C., Boyer, E., Ilic, S.: Probabilistic deformable surface tracking from multiple videos. ECCV (2010)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, series B (1977)
11. Franco, J., Boyer, E.: Learning temporally consistent rigidities. CVPR (2011)
12. Franco, J., Menier, C., Boyer, E., Raffin, B.: A distributed approach for real-time 3d modeling. CVPR Workshop (2004)
13. Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. Annales de l'institut Henri Poincaré 10, 215–310 (1948)
14. Granger, S., Pennec, X.: Multi-scale EM-ICP: A fast and robust approach for surface registration. ECCV 4, 6973 (2002)
15. Hasler, N., Ackermann, H., Rosenhahn, B., Thormählen, T., Seidel, H.P.: Multi-linear pose and body shape estimation of dressed subjects from image sets. CVPR (2010)
16. Hufnagel, H., Pennec, X., Ehrhardt, J., Ayache, N., Handel, H.: Generation of a Statistical Shape Model with Probabilistic Point Correspondences and EM-ICP. IJCAR 2(5) (2008)
17. J.Gall, C.Stoll, de Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P.: Motion capture using joint skeleton tracking and surface estimation. CVPR (2009)
18. Kendall, D.: Shape manifolds, procrustean metrics, and complex projective spaces. Bulletin of the London Mathematical Society 16(2), 81–121 (1984)
19. Liu, Y., Gall, J., Stoll, C., Dai, Q., Seidel, H.P., Theobalt, C.: Markerless motion capture of multiple characters using multi-view image segmentation. PAMI (2013)
20. Matsuyama, T., Nobuhara, S., Takai, T., Tung, T.: 3d video and its applications. Springer (2012)
21. Mémoli, F., Sapiro, G.: Comparing Point Clouds. SGP (2004)
22. Ovsjanikov, M., Mrigot, Q., Mmoli, F., Guibas, L.J.: One point isometric matching with the heat kernel. Comput. Graph. Forum 29(5) (2010)
23. Sahilliogu, Y., Yemez, Y.: 3D Shape correspondence by isometry-driven greedy optimization. CVPR (2010)
24. Starck, J., Hilton, A.: Model-based multiple view reconstruction of people. ICCV (2003)
25. Starck, J., Hilton, A.: Spherical matching for temporal correspondence of non-rigid surfaces. ICCV (2005)
26. Straka, M., Hauswiesner, S., Ruether, M., Bischof, H.: Simultaneous shape and pose adaption of articulated models using linear optimization (2012)
27. Sumner, R.W., Popović, J.: Deformation Transfer for Triangle Meshes. ACM Transactions on Graphics 23(3) (2004)
28. Tung, T., T.Matsuyama: Topology dictionary for 3d video understanding. PAMI 34(8), 1645–1647 (2012)
29. Tung, T., T.Matsuyama: Intrinsic characterization of dynamic surfaces. CVPR (2013)
30. Vlasic, D., Baran, I., Matusik, W., Popovic, J.: Articulated mesh animation from multi-view silhouettes. ACM Transactions on Graphics 27(3) (2008)
31. Windheuser, T., Schlickewei, U., Schmidt, F., Cremers, D.: Geometrically consistent elastic matching of 3d shapes: A linear programming solution. ICCV (2011)
32. Zeng, Y., Wang, C., Gu, X., Samaras, D., Paragios, N.: A Generic Deformation Model for Dense Non-Rigid Surface Registration: a Higher-Order MRF-based Approach. ICCV (2013)

$\therefore$

## A.3 AN EFFICIENT VOLUMETRIC FRAMEWORK FOR SHAPE TRACKING

Benjamin Allain, Jean-Sébastien Franco, Edmond Boyer CVPR 2015 - IEEE International Conference on Computer Vision and Pattern Recognition, Jun 2015, Boston, United States. (**Oral Presentation**, acceptance rate **<3%**)

# An Efficient Volumetric Framework for Shape Tracking

Benjamin Allain Jean-Sébastien Franco Edmond Boyer

Inria Grenoble Rhône-Alpes - LJK
Grenoble Universities, France
firstname.lastname@inria.fr

## Abstract

*Recovering 3D shape motion using visual information is an important problem with many applications in computer vision and computer graphics, among other domains. Most existing approaches rely on surface-based strategies, where surface models are fit to visual surface observations. While numerically plausible, this paradigm ignores the fact that the observed surfaces often delimit volumetric shapes, for which deformations are constrained by the volume inside the shape. Consequently, surface-based strategies can fail when the observations define several feasible surfaces, whereas volumetric considerations are more restrictive with respect to the admissible solutions. In this work, we investigate a novel volumetric shape parametrization to track shapes over temporal sequences. In constrast to Eulerian grid discretizations of the observation space, such as voxels, we consider general shape tesselations yielding more convenient cell decompositions, in particular the Centroidal Voronoi Tesselation. With this shape representation, we devise a tracking method that exploits volumetric information, both for the data term evaluating observation conformity, and for expressing deformation constraints that enforce prior assumptions on motion. Experiments on several datasets demonstrate similar or improved precisions over state-of-the-art methods, as well as improved robustness, a critical issue when tracking sequentially over time frames.*

## 1. Introduction

The capture of shapes and their evolutions has been a very active research topic for the last decade, motivated by many applications for which dynamic shape models are useful. This ability is of interest for several fields of activity such as computer-assisted design, virtual reality, entertainment, medical imaging, gesture and sports analysis. Ever since the initial promises of free viewpoint video [9], many models of shape capture have been explored. Initially examined as a per-time reconstruction problem, e.g. [24, 14], temporal integration and tracking of the shape in the time domain have then been considered, e.g. [11, 3]. In any case, however, surface-based models, such as meshes, have been largely dominant to represent and track shapes. This is due to several factors, primarily to the fact that visual observations generally lie on the shape surface, but also to the popularity of surface-based representations in the vision and graphics communities and the availability of efficient tools to manipulate them. Yet it has been observed that certain forms of volume-preserving deformations may be beneficial to model shape deformations in graphics applications such as [1, 5], or to enforce volumetric constraints, nevertheless based on surface tesselations, in dynamic shape capture [10].

While the idea has led to interesting preliminary results, a full volumetric treatment of dynamic shape capture is still to be investigated and its benefits evaluated. Among the expected advantages of this approach are its ability to express volume conservation as well as its ability to enforce local volumetric deformation constraints. In this paper, we address this problem with a twofold contribution: we first propose a dedicated volumetric deformation model based on Centroidal Voronoi Tesselations (CVT) [13], which integrates the latest advances of recent tracking models, and second we propose an evaluation of the method based on a hybrid multi-camera and marker-based capture dataset [21].

### 1.1. Previous Work

A large set of techniques exist to capture moving shapes as a time independent sequence of meshes representing the object's surface [24, 14]. For this process, many volumetric parameterizations have also been devised, based on regular or hierarchical Eulerian grid discretizations [30, 20], although they are mainly dedicated to single time occupancy representation. Some approaches have taken these representations a step further, by examining short term motion characteristics of the shape using regular volume grids [33, 17, 32], yet they do not retrieve long term motion information of the sequence, nor do they embed spe-

cific motion models in the volume.

Various methods attempt leveraging time consistency to retrieve temporally coherent shape models, in the vast majority of cases manipulating a surface model. While in some cases this process is purely data-driven, by aligning surfaces across frames using sparse matching and stereo refinement [29], in most cases a deformation prior is used to drive the method toward the solution within a plausible state space. In its weakest form and without further assumptions, pure spatio-temporal continuity of the observed surface can be used [16]. At the other end of the spectrum a full kinematic rigging of a template model can be assumed, where the surface is expressed from kinematic parameters using e.g. the linear blend skinning deformation framework [23] popularized for 3D animation in the computer graphics community. These parameters can then be estimated for best fitting the model reprojections to image and silhouette data [34, 3, 18, 21]. For tracking more general subjects and situations, more generic surface deformation frameworks have been explored to bypass the rigging stage and allow for more general non-rigid motion components. Central to these methods is the idea of enforcing a cohesive behavior of the surface, such as locally rigid behavior [15], Laplacian deformation [11, 10, 6], inextensibility [25], or elasticity between piecewise-rigid surface patches [7, 6].

Among the existing surface capture methods, only a handful use volumetric representations. Some methods have proposed reparameterizing temporally aligned sequences using a volumetric cage embedding [26, 31] inspired from the animation community [27, 19]. However, no volume deformation model strong enough to solve the full tracking problem has yet emerged from these works. Among the methods successfully using volume preserving constraints, most use a Delaunay tetrahedrization of reconstructed template surface points [11, 10, 6] to enforce as-rigid-as-possible or Laplacian deformation constraints common to 3D animation techniques [1, 28]. It can be noted that the proposed decomposition is not fully volumetric as it only involves tesselating surfaces. In contrast, we propose a fully volumetric treatment of the problem, using an intrinsically volumetric tesselation, deformation model and data terms for rewarding volume alignment.

### 1.2. Approach Overview

We formulate the tracking problem as the MAP estimation of multiple poses of a given geometric template model, non-rigidly adjusted to a set of temporally inconsistent shape measurements. In multi-view camera systems, these measurements typically take the form of time independent 3D mesh reconstructions obtained from a visual hull or multi-view stereo method, which is what we assume here. To efficiently make use of volumetric information, we need to express volume conservation and overlapping constraints from the template to the observed shape volumes. For representational and computational efficiency, we thus need a proper discretization of the interior of the shape. While uniformly located in the volume, regular grids are inherently anisotropic and biased toward the axis of the template basis. Furthermore, their intersection with the object surface yields boundary voxels of irregular shape (Fig. 1(a)). On the other hand, the Constrained Delaunay tetrahedrization of the boundary vertices, previously used in [11, 10, 6], yields a set of highly non-uniform tetrahedra spanning the whole interior of the volume, whose cardinality is not controlled but imposed by the surface discretization (Fig. 1(b)). Taking instead the Voronoi diagram of a uniform set of samples of the interior volume decorrelates the cardinality of the decomposition from the geometry, but still yields cells of irregular shape (Fig. 1(c)). Rejection sampling may statistically impose additional regularity, but this would only come with asymptotic guaranties attainable at high computational cost. We therefore propose to use CVT (Fig. 1(d)), informally a Voronoi tesselation where the samples are iteratively repositioned to coincide with the center of mass of their cell, which achieves the desired properties [13]: isotropy, rotational invariance, uniform cells of compact and regular form factor, regular intersection of boundary cells and surface, independent cardinality and practical computation.

After introducing how to define and compute CVTs in the context of our approach (§2), we show how this discretization can be used to define adequate volumetric deformation (§3) and observation (§4) models in the form of Bayesian prior and likelihoods. The MAP estimation proposed on this basis in §5 is evaluated in §6.

## 2. Centroidal Voronoi Tessellation (CVT)

**Definitions.** To tesselate the shape, we manipulate Voronoi diagrams that are restricted, or clipped, to its inner volume. More formally, let $\mathcal{S}$ be a set of 3D point samples of a volumetric domain $\Omega$, either the template to be fitted or the observed shape for our purposes. The *Clipped Voronoi diagram* of $\mathcal{S}$ in $\Omega$ is defined as the intersection of the Voronoi diagram of $\mathcal{S}$ in $\mathbb{R}^3$ with the domain $\Omega$. Thus the Voronoi cell $\Omega_s$ of a sample $s$ is the set of points from $\Omega$ that are closer to $s$ than to any other sample:

$$\Omega_s = \{\mathbf{x} \in \Omega \mid \forall s' \in \mathcal{S} \backslash \{s\} \quad \|\mathbf{x} - \mathbf{x}_s\| < \|\mathbf{x} - \mathbf{x}_{s'}\|\}\,, \quad (1)$$

where cells $\Omega_s$ are mutually exclusive and define a partition of $\Omega$:

$$\bigcup_{s \in \mathcal{S}} \overline{\Omega_s} = \overline{\Omega}\,, \quad (2)$$

where $\overline{\Omega_s}$ and $\overline{\Omega}$ denote topolgical set closures. If the border $\partial\Omega$ of $\Omega$ is a polyhedral surface, then each cell also has a polyhedral border.

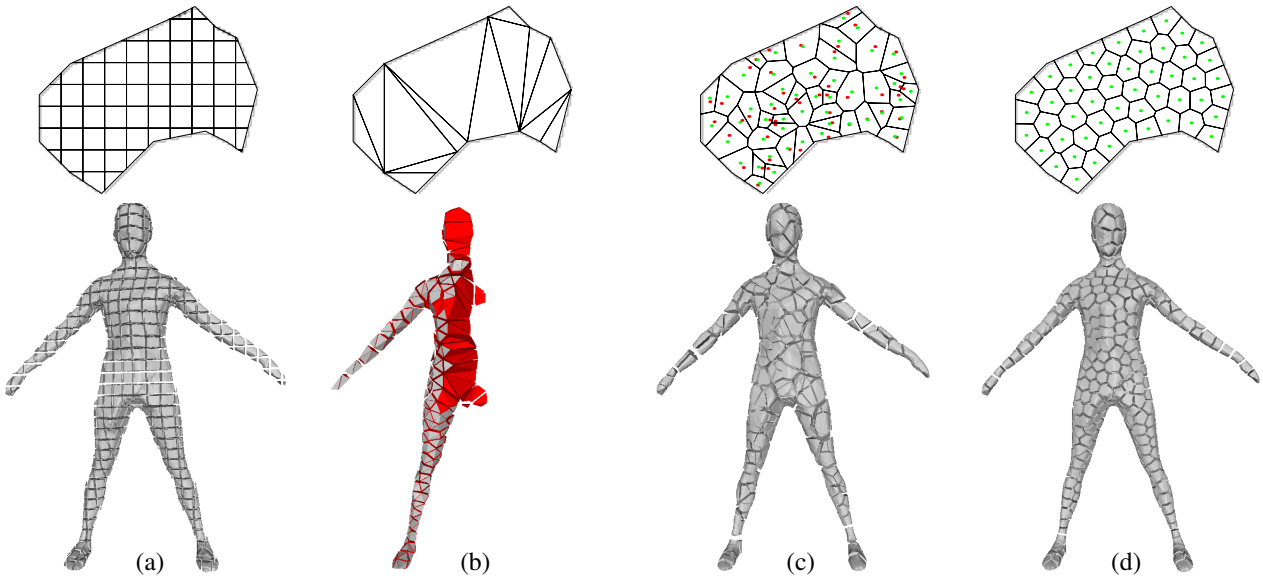



Figure 1. Possible volumetric decompositions of the template and observed shapes. (top) 2D schematic view. (bottom) 3D decomposition example. (a) Voxels on a regular grid. (b) A sliced Constrained Delaunay tetrahedrization showing the elongated inner tetrahedra generated. (c) Voronoi cells with random centroids shown in red, center of mass of each cell in green. (d) Centroidal Voronoi tesselation cells, where the center of mass and cell centroid coincide.

A *Centroidal Voronoi tessellation* is a clipped Voronoi tessellation of $\Omega$ for which each sample $s$ is the center of mass of its (clipped) Voronoi cell $\Omega_s$. CVT cells are of regular size and shapes, and also define a regular connectivity of the samples set, two samples being connected if and only if their respective CVT cells share a face. This connectivity thus encodes the shape volume and topology, a property we will use in the following sections.

**Computing a CVT.** It has been proven [13] that local minima of the energy

$$E(\mathcal{S}) = \sum_{s \in \mathcal{S}} \int_{\mathbf{x} \in \Omega_s} \|\mathbf{x} - \mathbf{x}_s^t\|^2 dV \tag{3}$$

define CVTs on $\Omega$. Thus a CVT can be obtained by iteratively estimating the sample locations that minimize (3), with a quasi-Newton method such as the L-BFGS algorithm [22], for a sample population of desired size and uniform initial position.

## 3. Deformation Model

### 3.1. Principle

Once a regular, anisotropic volumetric decomposition of the template shape is obtained, we can use it as a fundamental building block to build a volumetric deformation model of the shape, which will constrain the estimation. Botsch *et al*. [5] show that a non-linear elastic deformation energy can be devised using small volumetric deformations, typically voxels. While such a reasoning could be advantageously transposed to the CVT discretization proposed, eliminating the grid orientation bias, doing so comes at a high computational cost. Cagniart et al. [7] show that the complexity of the deformation model is best decorrelated from the geometry itself, in their case by using rigid surface patches in lieu of the original surface vertices. Recent works have shown a way to improve the quality and temporal stability using a similar surface decomposition [2], by inferring a mean pose and sequence rigidity behaviour.

We extend the latter ideas to the case of a complete volumetric treatment of the deformation problem. In so doing, we cluster together groups of CVT cells in rigid volume patches $P_k$ using a k-medoids algorithm. Note that such patches can lie either on the surface or completely inside the template shape's volume, which is of particular interest to express non-rigid deformation of the model while preserving the local volume and averting over-compression or dilation. We associate to each patch a rigid transform $\mathbf{T}_k^t \in SE(3)$ at every time $t$. Each position $\mathbf{x}_{k,q}$ of a mesh vertex or inner sample is indiscriminately labeled as a point $q$. Its position can be written as a transformed version of its template position $\mathbf{x}_q^0$ as follows, once the rigid transform of its patch is applied:

$$\mathbf{x}_{k,q} = \mathbf{T}_k(\mathbf{x}_q^0). \tag{4}$$

This makes it possible to define a *pose* of the shape as the the set of patch transforms $\mathbf{T} = \{\mathbf{T}_k\}_{k \in \mathcal{K}}$, which expresses a given volumetric shape deformation.

### 3.2. Formulation

To prevent patch poses of the shape from being arbitrary, we constrain the shape to be close to a sequence rest pose $\bar{\mathbf{T}}$ and to follow constant rigidity characteristics C over the sequence. These rigid characteristics are defined for neighboring patch pairs in the volume $(P_k, P_l)$, as a binary valued property $\mathrm{c}_{kl}$, whose value in $\{0, 1\}$ reflects wether the relative motion between patches $P_k$ and $P_l$ is respectively articulated or rigid. To define the rest pose $\bar{\mathbf{T}}$ we rely on the following measure [5, 2] of the relative deformation energy between two arbitrary poses $\mathbf{T}^i$ and $\mathbf{T}^j$ of the template shape, given a rigidity configuration C:

$$\mathcal{E}(\mathbf{T}^i, \mathbf{T}^j|\mathrm{C}) = \sum_{(P_k,P_l)\in\mathcal{N}} \mathcal{E}_{kl}(\mathbf{T}^i, \mathbf{T}^j|\mathrm{c}_{kl}), \quad \text{with} \tag{5}$$

$$\mathcal{E}_{kl}(\mathbf{T}^i,\mathbf{T}^j|\mathrm{c}_{kl}) = \sum_{q\in P_k\cup P_l} \beta_{kl}(q,\mathrm{c}_{kl}) \|\mathbf{T}^i_{k-l}(\mathbf{x}^0_q) - \mathbf{T}^j_{k-l}(\mathbf{x}^0_q)\|^2,$$

where $\mathbf{T}^i_{k-l} = \mathbf{T}^{i\,-1}_l \circ \mathbf{T}^i_k$ is the relative transformation between patches $P_k$ and $P_l$ for pose $i$, and $\mathcal{N}$ is the set of neighboring patch pairs on the surface. The energy measures the rigid deviation from pose $i$ to $j$ of every neighboring patch pair, as the sum over each of the samples $s$ of the pair, of the discrepancy in relative positions of the vertex as displaced by $P_k$ on one hand, and $P_l$ on the other. If the two patches are rigidly linked ($\mathrm{c}_{kl} = 1$), then the discrepancy of all samples of the pair should be equally penalized, therefore $\beta_{kl}(s, 1)$ is chosen to be constant over all samples $s$ of the pair. On the other hand, if the patch pair is articulated ($\mathrm{c}_{kl} = 0$), only samples that lie near the boundary between the two patch volumes should be penalized for deviating relative positions: those samples materialize the locus of the inter-patch articulation, whereas samples that aren't close the inter-patch boundary can move more freely. We express this using $\beta_{kl}(q, 0) \propto \exp(-\frac{b_{kl}(s)}{\eta \bar{D}})$ where $b_{kl}(s)$ is the distance between the sample $s$ and the boundary between $P_k$ and $P_l$ on the template, $\bar{D}$ is the average patch diameter and $\eta$ is a global coefficient controlling the flexibility.

### 3.3. Resulting Pose Likelihoods

The relative pose energy described in (5) makes it possible to express the expected behavior of the estimated models as a prior and likelihood over the poses:

$$p(\bar{\mathbf{T}}) \;\propto\; \exp(-\mathcal{E}(\bar{\mathbf{T}}, \mathbf{Id})), \tag{6}$$

$$p(\mathbf{T}^t \mid \bar{\mathbf{T}}, \mathrm{C}) \;\propto\; \exp\left(-\mathcal{E}(\mathbf{T}^t, \bar{\mathbf{T}}|\mathrm{C})\right). \tag{7}$$

$p(\bar{\mathbf{T}})$ is the prior over the rest pose, which should minimize the relative displacement energy to the default template pose (transformed by identity $\mathbf{Id}$). This terms ensures minimal cohesion of the volume patches of the rest pose model, as it enforces mutual patch elasticity.

$p(\mathbf{T}^t|\bar{\mathbf{T}}, \mathrm{C})$ is the likelihood of a given tracked pose at time $t$, which should minimize the relative displacement energy with respect to the sequence rest pose $\bar{\mathbf{T}}$ given a current rigidity state C. This ensures the inter-patch cohesion of pose $\mathbf{T}^t$ as well as a general proximity to the rest pose, which stabilizes the resulting pose estimates. In turn the rest pose will be simultaneously estimated as the pose which minimizes the relative deformation energy to all poses in the sequence.

## 4. Observation Model

### 4.1. Probabilistic Shape Fitting

The observed shape $\Omega^t$ at time $t$ is described by the point cloud $\mathbf{Y}^t = \{\mathbf{y}^t_o\}_{o\in\mathcal{O}_t}$. To describe how a deformed template can explain the observed shape, we propose a generative data term following EM-ICP, expressing how a given deformed template point predicts the position of an observed shape point $o$. A set of cluster association variables $\mathrm{k}^t_o$ is therefore instantiated for every observed point in time, indicating which cluster generates this point. For simplicity, each observation $o$ is associated to its cluster $\mathrm{k}^t_o$ via the best candidate $q$ of cluster $\mathrm{k}^t_o$. The best candidate is chosen as the closest compatible sample in the cluster during iterative resolution. We consider that each cluster $P_k$ generates observations perturbed by a Gaussian noise with isotropic variance $\sigma^2$:

$$p(\mathbf{y}^t_o \mid \mathrm{k}^t_o, \mathbf{T}^t_k, \sigma) \;=\; \mathcal{N}(\mathbf{y}^t_o \mid \mathbf{T}^t_{\mathrm{k}}(\mathbf{x}^0_q), \sigma). \tag{8}$$

Note that $o$ indiscriminately refers to surface or volume sample points of the observed shape, as the principles we describe here apply to both, with the restriction that observed surface points only associate to surface template points, and volume samples are associated to volume samples of the template. As often proposed in ICP methods, we additionally filter associations using a compatibility test, described in the following sections. The compatibility test is specific to the nature (surface or volume) of the observed point and is detailed in the next paragraphs. If there is no compatible candidate in the cluster, then we set conditional probability density (8) to zero. We deal with outliers by introducing an outlier class among values of k, which generate observations with a uniform probability density over the scene.

### 4.2. Compatibility Tests

Compatibility tests are useful for pruning the association graph for obvious mismatches that would perturb and otherwise slow down convergence. We use two compatibilty tests respectively designed for surface fitting and volumetric fitting.

**Surface Observations.** While surface points may be matched based on position only, obvious orientation incompatibilites can be filtered by detecting large discrepancies between the normal of the deformed template candidate point $v$, and the normal of surface observation vertex $o$:

$$\vec{\mathbf{n}}_o^t \,.\, \mathbf{R}_k^t(\vec{\mathbf{n}}_v^0) \geq \cos(\theta_{\max}), \tag{9}$$

where $\vec{\mathbf{n}}_o^t$ is the surface normal of observation $o$, $\vec{\mathbf{n}}_v^0$ is the surface normal of the template at vertex $v$, $\mathbf{R}_k^t$ is the rotation component of $\mathbf{T}_k^t$, and $\theta_{\max}$ is an arbitrary threshold.

**Volume Observations.** We introduce a compatibility test specific to volumetric fitting, by assuming that the distance of inner surface points to the shape's surface remains approximately constant under deformation. Let us define the distance between an inner shape point $x$ and the shape's surface by:

$$d(x, \partial\Omega) = \min_{p \in \partial\Omega} d(x, p). \tag{10}$$

In our observation model, this hypothesis can be leveraged by the following compatibility test: a volumetric observation $o$ can be associated to a template point $s$ only if

$$d(\mathbf{x}_s^0, \partial\Omega^0) = d(\mathbf{y}_o^t, \partial\Omega^t). \tag{11}$$

To account for small deviations to this assumption, which might occur under e.g. slight compression or dilation of the perceived shape, we relax the equality constraint up to a precision $\epsilon$, where $\epsilon$ accounts for the distance-to-surface inconsistency caused by the discrete sampling of the template. Using the triangular inequality, it can be shown that this error is bounded by the maximum cell radius over the set of the template's CVT cells. This leads to the following compatibility test:

$$d(\mathbf{y}_o^t, \partial\Omega^t) - \epsilon \leq d(\mathbf{x}_s^0, \partial\Omega^0) \leq d(\mathbf{y}_o^t, \partial\Omega^t) + \epsilon \tag{12}$$

For the particular case of silhouette-base observed shapes, it can be noted that reconstruction algorithms based on the visual hull inherently provides inflated estimates of the true shape. This phenomenon results in an overestimation of the distance to the surface when computed on the reconstructed shape. Hence, we only impose a volumetric inclusion constraint instead of complete depth correspondance, i.e. we only keep the right inequality from expression (12) in this case:

$$d(\mathbf{x}_s^0, \partial\Omega^0) \leq d(\mathbf{y}_o^t, \partial\Omega^t) + \epsilon. \tag{13}$$

Contrary to the surface compatibility test, this test does not depend on pose parameters $\mathbf{T}$, consequently it is robust to convergence failures of inference algorithms.

## 5. Inference

The model proposed with (6), (7) and (8), defines a joint likelihood over the rest pose, the rigidity configuration, all temporal poses, the observed points and their selection variables, and prediction noise $\sigma$:

$$p(\bar{\mathbf{T}}) \prod_{t \in \mathcal{T}} \left( p(\mathbf{T}^t | \bar{\mathbf{T}}, \mathrm{C}) \prod_{o \in \mathcal{O}_t} p(\mathbf{y}_o^t \mid \mathrm{k}_o^t, \mathbf{T}^t, \sigma^t) \right), \tag{14}$$

It can be shown that this likelihood can be maximized using an Expectation Maximization algorithm [2, 12, 4], yielding maximum a posteriori estimates of the pose parameters $\bar{\mathbf{T}}$, $\mathbf{T}$ and prediction noise $\sigma$. This results in an algorithm iterating between two steps.

Intuitively, The **E-Step** computes all observation cluster assignment probabilities over $\mathrm{K}$, based on the distance to the predicted template positions under the currently estimated poses. Compatibility rules are applied at this stage. Probabilities over inter-cluster rigid links $\mathrm{C}$ are also estimated based on the current deformation energy of the poses. The **M-Step** updates the rest pose $\bar{\mathbf{T}}$, all poses $\mathbf{T}$, and prediction noise $\sigma$, using the assignment and rigid link probabilities to weigh individual observation contributions to each cluster transform estimate.

## 6. Experiments

### 6.1. Datasets

We validate our framework using four synchronized and calibrated multiple-camera datasets, labeled GOALKEEPER-13, DANCER [2], MARKER [21], and the newly proposed BALLET, whose content reflect a wide variety of shape tracking situations. DANCER is a long sequence (1362 frames, $2048 \times 2048$ resolution, 48 viewpoints) showing slow and medium speed dance moves, and thus offers good opportunity to verify the tracking stability. GOALKEEPER-13 ($2048 \times 2048$, 150 frames, 48 viewpoints) illustrates a specific soccer goalie plunging move, of particular interest when the goalie is on the floor, where the reconstruction data is of poor quality due to grazing camera angle and challenges the tracking performance. Both previous sequences otherwise have very high quality and detailed reconstructions. BALLET is a more challenging full HD ($1920 \times 1080$) sequence we have acquired with fewer cameras (9 viewpoints and 500 frames) and thus coarser reconstructions, consisting in a number of ballet moves with various levels of difficulty, including fast moves, spinning and crossing legs. MARKER ($1296 \times 972$, 500 frames, 12 viewpoints) is a sequence with two actors performing karate moves, illustrating the robustness to several subjects, and which was captured simultaneously with a set of sparse markers offering a reference and comparison basis

| method | std. dev. (L) | |
|---|---|---|
| | MARKER | BALLET |
| Cagniart *et al.* 2010 [8] | 3.85 | 1.22 |
| Allain *et al.* 2014 [2] | 4.32 | 1.20 |
| our method | **2.24** | **0.95** |

Table 1. Variation of the estimated volume over the sequence for MARKER and BALLET datasets.

with [21]. The reconstructions are of coarser quality due to relatively noisy inputs and occasional jumps where actor heads get clipped.

### 6.2. Experimental Protocol

We first select a template among the better frames with correct reconstruction topology, then compute a CVT using §2 with 5'000 samples per person (10'000 for the two-person MARKER sequence) and 250 clusters per person (500 for MARKER), as illustrated in Fig. 3. Each shape reconstruction is obtained from a silhouette-based reconstruction algorithm [14] and CVTs are also extracted (1 minute/frame). We then apply the algorithm described using a sliding window strategy over a 10 frame window, where the rest position is computed for each time slice to locally stabilize the estimation. The sequences are initially solved for frames in the vicinity of the template pose, then the solution is classically propagated to future windows as initialization. Convergence has been achieved for all frames, typically in a few hundred iterations, with a convergence time per frame of the order of a minute to a few minutes. The provided supplemental video[1] illustrates the results obtained with these sequences.

### 6.3. Quantitive Analysis

**Volume Stability.** We verify here the assertion that the volumetric parameterization of the tracking produces poses with stable volumes. As we use silhouette based reconstructions, it is not relevant to compare the estimated volumes with the observed shape volumes. Instead, we compute the standard deviation to this volume, in Table 1 and provide a comparison of these results with best runs of two state of the art methods [8, 2]. This comparison supports the initial intuition of volumetric stability in the sequence, as the standard deviation of the estimated volumes is significantly lower for our method.

**Silhouette reprojection error.** We evaluate the silhouette reprojection error as the symmetric pixel difference between the reprojected and the silhouette projection of the reconstructions used as observed input shapes. We then express this value as a percentage of the area of the silhouette region in each view. Table 2 shows favorable comparisons to state of the art methods [8, 2]. In particular, the mean error and maximum error achieved by our method over the sequences is significantly lower, and exhibits lower variance. Additionally we test the influence of the volumetric data term by comparing the results with a run where it is disabled (surface data-term only), all other parameters being equal. Interestingly, the method still achieves better mean error than state of the art, but with less stable behavior.

[1] Video available at https://hal.inria.fr/hal-01141207

| method | mean | stddev. | median | max |
|---|---|---|---|---|
| Cagniart *et al.* [8] | 5.74 | 1.88 | 5.48 | 15.20 |
| Allain *et al.* [2] | 5.81 | 1.70 | 5.61 | 13.77 |
| Ours, no vol. fitting | 4.62 | 1.94 | **4.28** | 17.20 |
| Ours | **4.56** | **1.21** | 4.43 | **11.00** |

Table 2. Mean and statistics of silhouette reprojection error over BALLET dataset, expressed in percentage of silhouette area.

| method | mean | std. dev. |
|---|---|---|
| Cagniart *et al.* [8] | 55.11 | 48.02 |
| Allain *et al.* [2] | 43.22 | 29.58 |
| Proposed, no surface fitting | 42.60 | 29.32 |
| Proposed | **38.41** | **26.70** |
| Liu *et al.* [21] | **29.61** | **25.50** |

Table 3. Mean marker error and std.dev. (mm), MARKER dataset. Note that Liu *et al.* assume a rigged skeleton is associated to the template, a stronger and more restrictive assumption.

**Marker reference error.** We use the MARKER sequence provided by Liu *et al.* [21] to sparsely compare the output quality of our method against state of the art methods. This comparison is illustrated in Table 3 and plotted against time in the sequence in Fig. 2. Again we illustrate turning off the surface data term, in which case the estimation is slightly worse. The method proposed performs consistently better than comparable surface-based state of the art. Note that Liu *et al.* fully rig a skeleton to the template, which provides slightly better mean results than ours thanks to the stronger assumption. On the other hand, our method is generic and can be applied to arbitrary objects.

### 6.4. Qualitative Assessment

To illustrate the benefits of the approach, in particular where the improvements can be seen, we provide excerpts of the datasets (see supplemental video for further examples). Fig. 4 shows the improvements of the method over surface-based methods [8, 2] in poses of strong contortion, such as an elbow or knee bending gesture. Because of their use of the elastic energy on the surface, these methods tend to dilute error compensation over a smooth and extended location of the folds, yielding curvy elbow and knee shapes in the tracking. A usual side effect here is the local decrease of

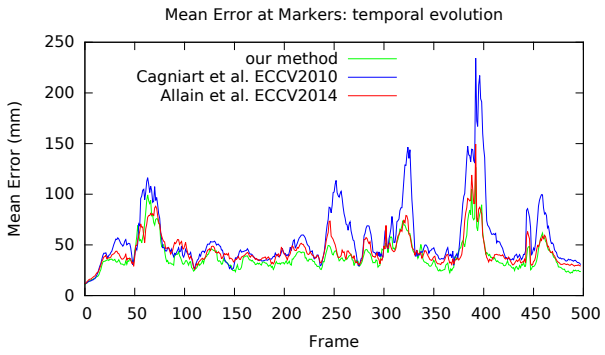



Figure 2. Mean error for temporal evolution over MARKER dataset.

the shape volume in the vicinity of the fold. In contrast, our method being volumetrically constrained, it penalizes such local volume changes and prefers to focus the bending energy on fewer volumetric patch articulations, yielding more consistent and accurate pose estimates. The GOALKEEPER-13 dataset illustrates the increased robustness in the presence of highly perturbed reconstructions thanks to the volumetric constraints, where other methods yield random results. The reconstructed visual hull input is very ambiguous on the shown frame because of the presence of concavities and the strong topology mismatch creates errors for surface-based methods.

## 7. Discussion

We have presented a novel volumetric approach to shape tracking based on CVT volume decomposition. The approach leverages CVT desirable properties to build suitable volumetric deformable constraints, while formulating a discrete volume assignment scheme as data term through the uniform cell centroid coverage of the volume. Currently, the volumetric clustering proposed for volumes yields uniform sizes over the entire template shape, which can be a limitation for parts that are thinner than the cluster size, such as arms. We will address this in future work with adaptive cluster densities, ensuring the volumetric prior is equally efficient regardless of thickness. Numerical analysis nevertheless shows significant improvement over state of the art tracking methods, both in terms of tracking error over the surface and silhouette reprojection. The framework is also shown to conform to initial intuition in being more stable in terms of the errors and volume measures of the fitted template shapes. We believe the approach paves the way for proper use of volumetric priors in any shape tracking framework.

## Acknowledgements

This work was funded by the Seventh Framework Programme EU project RE@CT (grant agreement no. 288369). We thank Li Wang and Franck Hetroy for providing the CVT computation code.

## References

[1] M. Alexa, D. Cohen-Or, and D. Levin. As-rigid-as-possible shape interpolation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 157–164, 2000. 1, 2

[2] B. Allain, J.-S. Franco, E. Boyer, and T. Tung. On mean pose and variability of 3d deformable models. In *ECCV*, 2014. 3, 4, 5, 6, 8

[3] L. Ballan and G. M. Cortelazzo. Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. In *3DPVT*, 2008. 1, 2

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 5

[5] M. Botsch, M. Pauly, M. Wicke, and M. Gross. Adaptive space deformations based on rigid cells. *Comput. Graph. Forum*, 26(3):339–347, 2007. 1, 3, 4

[6] C. Budd and A. Hilton. Temporal alignment of 3d video sequences using shape and appearance. *Conference on Visual Media Production*, pages 114–122, 2010. 2

[7] C. Cagniart, E. Boyer, and S. Ilic. Free-from mesh tracking: a patch-based approach. In *CVPR*, 2010. 2, 3

[8] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In *ECCV*, 2010. 6, 8

[9] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. In *ACM SIGGRAPH 2003 Papers*, pages 569–577, 2003. 1

[10] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. *ACM Transactions on Graphics*, 27(3), 2008. 1, 2

[11] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less deformable mesh tracking for human shape and motion capture. In *CVPR*, 2007. 1, 2

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, series B*, 1977. 5

[13] Q. Du, V. Faber, and M. Gunzburger. Centroidal voronoi tessellations: Applications and algorithms. *SIAM review*, 41:637–676, 1999. 1, 2, 3

[14] J.-S. Franco and E. Boyer. Exact polyhedral visual hulls. In *British Machine Vision Conference (BMVC'03)*, volume 1, pages 329–338, Norwich, United Kingdom, Sept. 2003. 1, 6

[15] Y. Furukawa and J. Ponce. Dense 3d motion capture from synchronized video streams. *CVPR*, 2008. 2

[16] B. Goldlücke and M. Magnor. Space-time isosurface evolution for temporally coherent 3D reconstruction. In *CVPR*, 2004. 2

[17] L. Guan, J.-S. Franco, E. Boyer, and M. Pollefeys. Probabilistic 3d occupancy flow with latent silhouette cues. In *CVPR*, 2010. 1

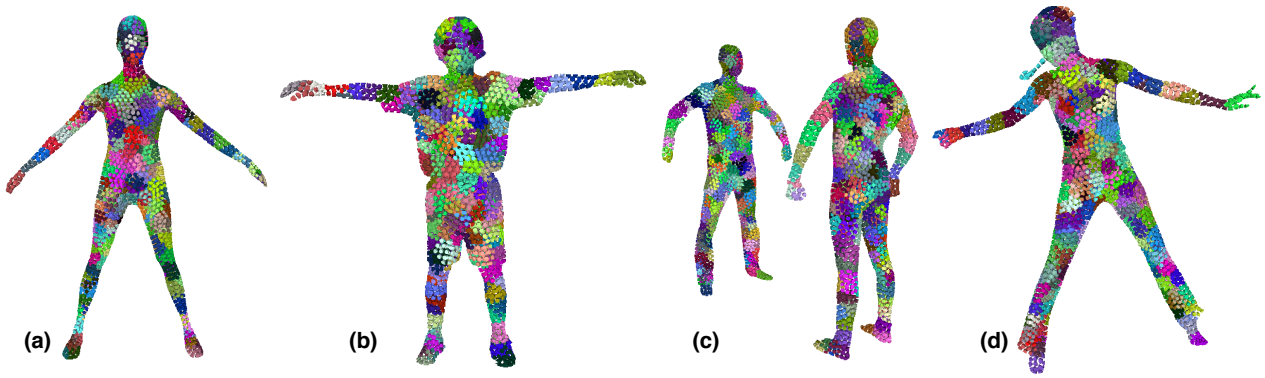



Figure 3. Patched volumetric decompositions of the template for sequences (a) BALLET, (b) GOALKEEPER-13, (c) MARKER and (d) DANCER. A color has been assigned to each patch for a visualization purpose.

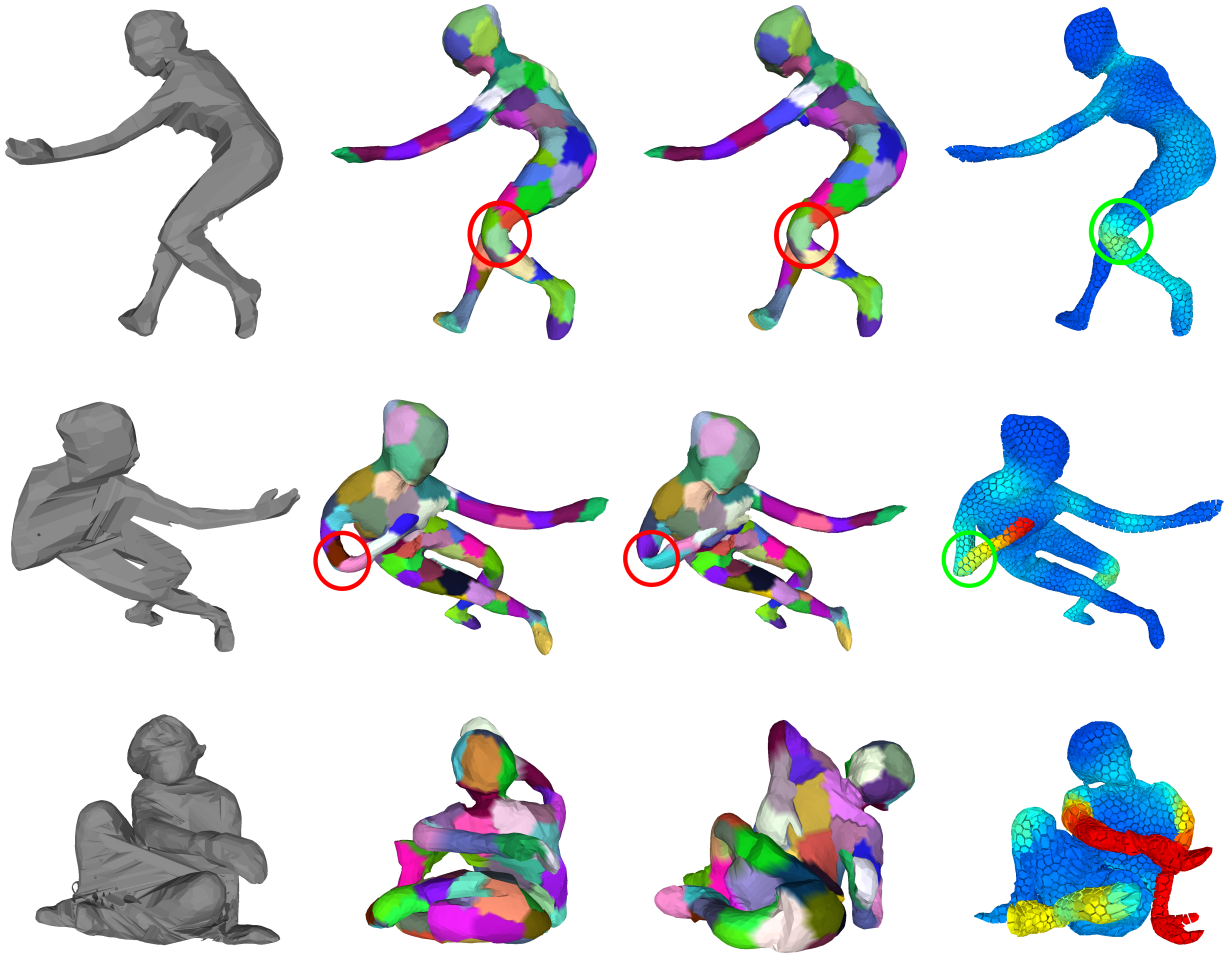



Figure 4. Frames of the BALLET (top and middle) and GOALKEEPER-13 datasets (bottom). (a) Visual hull input. (b) Tracking result of Cagniart *et al*. [8]. (c) Allain *et al*. [2]. (d) Our method. Note the improved angular shapes on the dancer's knee (top) and elbow (middle), and the improved robustness (bottom).

[18] J.Gall, C.Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, 2009. 2

[19] P. Joshi, M. Meyer, T. DeRose, B. Green, and T. Sanocki. Harmonic coordinates for character articulation. In *ACM SIGGRAPH 2007 Papers*, 2007. 2

[20] K. Kolev, T. Brox, and D. Cremers. Fast joint estimation of silhouettes and dense 3d geometry from multiple images. *PAMI*, 2012. 1

[21] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt. Markerless motion capture of multiple characters using multi-view image segmentation. *PAMI*, 2013. 1, 2, 5, 6

[22] Y. Liu, W. Wang, B. Lévy, F. Sun, D.-M. Yan, L. Liu, and C. Yang. On centroidal voronoi tessellation - energy smoothness and fast computation. *ACM Transactions on Graphics*, 28(101), 2009. 3

[23] N. Magnenat-Thalmann, R. Laperrière, and D. Thalmann. Joint-dependent local deformations for hand animation and object grasping. In *Proceedings on Graphics Interface '88*, pages 26–33, Toronto, Ont., Canada, Canada, 1988. Canadian Information Processing Society. 2

[24] W. Matusik, C. Buehler, and L. McMillan. Polyhedral visual hulls for real-time rendering. In S. Gortler and K. Myszkowski, editors, *Rendering Techniques 2001*, Eurographics, pages 115–125. Springer Vienna, 2001. 1

[25] M. Salzmann, F. Moreno-Noguer, V. Lepetit, and P. Fua. Closed-form solution to non-rigid 3d surface registration. In *ECCV*, 2008. 2

[26] Y. Savoye and J.-S. Franco. Cage-based tracking for performance animation. In *ACCV*, 2010. 2

[27] T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '86, pages 151–160, 1986. 2

[28] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, SGP '04, pages 175–184, 2004. 2

[29] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE CGA*, 2007. 2

[30] R. Szeliski. Rapid octree construction from image sequences. *CVGIP: Image Underst.*, 58(1):23–32, 1993. 1

[31] J.-M. Thiery, J. Tierny, and T. Boubekeur. Cager: Cage-based reverse engineering of animated 3d shapes. *Computer Graphics Forum*, 31(8):2303–2316, 2012. 2

[32] A. O. Ulusoy, O. Biris, and J. L. Mundy. Dynamic probabilistic volumetric models. In *ICCV*, 2013. 1

[33] S. Vedula, S. Baker, S. M. Seitz, and T. Kanade. Shape and motion carving in 6d. In *CVPR*, 2000. 1

[34] D. Vlasic, I. Baran, W. Matusik, and J. Popovic. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3), 2008. 2

$\therefore$

## A.4 Estimation of Human Body Shape in Motion with Wide Clothing

Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, Stefanie Wuhrer
European Conference on Computer Vision 2016, Oct 2016, Amsterdam, Netherlands

# Estimation of Human Body Shape in Motion with Wide Clothing

Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer

Inria Grenoble Rhône-Alpes, France
Laboratoire Jean Kuntzmann, Université Grenoble Alpes, France
{jinlong.yang,jean-sebastien.franco,franck.hetroy,stefanie.wuhrer}inria.fr

**Abstract.** Estimating 3D human body shape in motion from a sequence of unstructured oriented 3D point clouds is important for many applications. We propose the first automatic method to solve this problem that works in the presence of loose clothing. The problem is formulated as an optimization problem that solves for identity and posture parameters in a shape space capturing likely body shape variations. The automation is achieved by leveraging a recent robust pose detection method [1]. To account for clothing, we take advantage of motion cues by encouraging the estimated body shape to be inside the observations. The method is evaluated on a new benchmark containing different subjects, motions, and clothing styles that allows to quantitatively measure the accuracy of body shape estimates. Furthermore, we compare our results to existing methods that require manual input and demonstrate that results of similar visual quality can be obtained.

**Keywords:** human body modeling, shape and motion estimation, statistical shape space

## 1 Introduction

Estimating 3D human body shape in motion is important for applications ranging from virtual change rooms to security. While it is currently possible to effectively track the surface of the clothing of dressed humans in motion [2] or to accurately track body shape and posture of humans dressed in tight clothing [3], it remains impossible to automatically estimate the 3D body shape in motion for humans captured in loose clothing.

Given an input motion sequence of raw 3D meshes or oriented point clouds (with unknown correspondence information) showing a dressed person, the goal of this work is to estimate the body shape and motion of this person. Existing techniques to solve this problem are either not designed to work in the presence of loose clothing [4,5] or require manual initialization for the pose [6,7], which limits their use in general scenarios. The reason is that wide clothing leads to strong variations of the acquired surface that is challenging to handle automatically. We propose an *automatic* framework that allows to estimate the human body shape and motion that is robust to the presence of *loose clothing*.

Existing methods that estimate human body shape based on an input motion sequence of 3D meshes or oriented point clouds use a shape space that models human body shape variations caused by different identities and postures as prior. Such a prior allows to reduce the search space to likely body shapes and postures. Prior works fall into two lines of work. On the one hand, there are human body shape estimation methods specifically designed to work in the presence of loose clothing [6,7]. These techniques take advantage of the fact that observations of a dressed human in motion provides important cues about the underlying body shape as different parts of the clothing are close to the body shape in different frames. However, these methods require manually placed markers to initialize the posture. On the other hand, there are human body shape estimation methods designed to robustly and automatically compute the shape and posture estimate over time [4,5]. However, these methods use strong priors of the true human body shape to track the posture over time and to fit the shape to the input point cloud, and may therefore fail in the presence of loose clothing.

In this work, we combine the advantages of these two lines of work by proposing an automatic framework that is designed for body shape estimation under loose clothing. Like previous works, our method restricts the shape estimate to likely body shapes and postures, as defined by a shape space. We use a shape space that models variations caused by different identities and variations caused by different postures as linear factors [8]. This simple model allows for the development of an efficient fitting approach. To develop an automatic method, we employ a robust pose detection method that accounts for different identities [1] and use the detected pose to guide our model fitting. To account for clothing, we take advantage of motion cues by encouraging the estimated body shape to be located inside the acquired observation at each frame. This constraint, which is expressed as a simple energy that is optimized over all input frames jointly, allows to account for clothing without the need to explicitly detect skin regions on all frames as is the case for previous methods [7,9].

To the best of our knowledge, existing datasets in this research area do not provide 3D sequences of both body shape as ground truth and dressed scans for estimation. Therefore, visual quality is the only evaluation choice. To quantitatively evaluate our framework and allow for future comparisons, we propose the first dataset consisting of synchronized acquisitions of dense unstructured geometric motion data and sparse motion capture data of 6 subjects with 3 clothing styles (tight, layered, wide) under 3 representative motions, where the capture in tight clothing serves as ground truth body shape.

The main contributions of this work are the following.

- An automatic approach to estimate 3D human body shape in motion in the presence of loose clothing.
- A new benchmark consisting of 6 subjects captured in 3 motions and 3 clothing styles each that allows to quantitatively compare human body shape estimates.

## 2 Related work

Many works estimate human posture without aiming to estimate body shape, or track a known body shape over time. As our goal is to simultaneously estimate body shape and motion automatically and in the presence of loose clothing, we will focus our discussion on this scenario.

**Statistical shape spaces.** To model human body shape variations caused by different identities, postures, and motions, statistical shape spaces are commonly used. These shape spaces represent a single frame of a motion sequence using a low-dimensional parameter space that typically models shape variations caused by different identities and caused by different postures using separate sets of parameters. Such shape spaces can be used as prior when the goal is to predict a likely body shape under loose clothing.

Anguelov et al. [10] proposed a statistical shape space called SCAPE that combines an identity model computed by performing principal component analysis (PCA) on a population of 3D models in standard posture with a posture model computed by analyzing near-rigid body parts corresponding to bones. This model performs statistics on triangle transformations, which allows to model non-rigid deformations caused by posture changes. Achieving this accuracy requires solving an optimization problem to reconstruct a 3D mesh from its representation in shape space. To improve the accuracy of the SCAPE space, Chen et al. [11] propose to combine the SCAPE model with localized multilinear models for each body part. To model the correlation of the shape changes caused by identity and posture changes, Hasler et al. [12] perform PCA on a rotation-invariant encoding of the model's triangles. These models may be used as priors when estimating human body shape in motion, but none of them allow to efficiently reconstruct a 3D human model from the shape space.

To speed up the reconstruction time from the SCAPE representation, Jain et al. [13] propose a simplified SCAPE model, denoted by S-SCAPE in the following, that computes the body shape by performing PCA on the vertex coordinates of a training set in standard posture and combines this with a linear blend skinning (LBS) to model posture changes. Any posture variations present in the training data cause posture variation to be modeled in identity space, which is known to cause counter-intuitive deformations [8]. To remedy this, recently proposed shape spaces start by normalizing the posture of the training data before performing statistics and model shape changes caused by different factors such as identity and posture as multilinear factors [14,8,15]. We use the normalized S-SCAPE model [8] in this work; however, any of these shape spaces could be used within our framework.

Recently, Pons-Moll et al. [16] proposed a statistical model that captures fine-scale dynamic shape variation of the naked body shape. We do not model dynamic geometry in this work, as detailed shape changes are typically not observable under loose clothing.

**Estimation of static body shape under clothing.** To estimate human body shape based on a static acquisition in loose clothing and in arbitrary posture, the following two approaches have been proposed. Balan et al. [9] use a SCAPE model to estimate the body shape under clothing based on a set of calibrated multi-view images. This work is evaluated on a static dataset of different subjects captured in different postures and clothing styles. Our evaluation on 3D motion sequences of different subjects captured in different motions and clothing styles is inspired by this work. Hasler et al. [17] use a rotation-invariant encoding to estimate the body shape under clothing based on a 3D input scan. While this method leads to accurate results, it cannot easily be extended to motion sequences, as identity and posture parameters are not separated in this encoding.

Both of these methods require manual input for posture initialization. In this work, we propose an automatic method to estimate body shape in motion.

**Estimation of body shape in motion.** The static techniques have been extended to motion sequences with the help of shape spaces that separate shape changes caused by identity and posture. Several methods have been proposed to fit a SCAPE or S-SCAPE model to Kinect data by fixing the parameters controlling identity over the sequence [4,5]. These methods are not designed to work with clothing, and it is assumed that only tight clothing is present.

Two more recent methods are designed to account for the presence of clothing. The key idea of these methods is to take advantage of temporal motion cues to obtain a better identity estimate than would be possible based on a single frame. Our method also takes advantage of motion cues.

Wuhrer et al. [6] use a shape space that learns local information around each vertex to estimate human body shape for a 3D motion sequence. The final identity estimate is obtained by averaging the identity estimates over all frames. While this shape space leads to results of high quality, the fitting is computationally expensive, as the reconstruction of a 3D model from shape space requires solving an optimization problem. Our method uses a simpler shape space while preserving a similar level of accuracy by using an S-SCAPE model that pre-normalizes the training shapes with the help of localized information.

Neophytou and Hilton [7] propose a faster method based on a shape space that models identity and posture as linear factors and learns shape variations on a posture-normalized training database. To constrain the estimate to reliable regions, the method detects areas that are close to the body surface. In contrast, our method constrains the estimate to be located inside the observed clothing at every input frame, which results in an optimization problem that does not require a detection.

Both of these methods require manual input for posture initialization on the first frame. Additionally, a temporal alignment is required by Neophytou and Hilton. Computing temporal alignments is a difficult problem, and manual annotation is tedious when considering larger sets of motion sequences. In contrast, our method is fully automatic and addresses both aspects.

## 3 S-SCAPE model

In this work, we use the S-SCAPE model as prior for human body shape changes caused by different identities and postures. While we choose this shape space, any shape space that models identity and posture as multilinear factors could be used [14,15]. Although such a simple shape space does not accurately model correlated shape changes, such as muscle bulging, it allows to effectively separate the different variations and can be fitted efficiently to input scans.

This section briefly reviews the S-SCAPE model introduced by Jain et al. [13] that allows to separate the influence of parameters controlling identity and parameters controlling posture of a human body shape. In the following, we denote by $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ the parameter vectors that influence shape changes caused by identity and posture changes, respectively. In this work, we use the publicly available posture-normalized S-SCAPE model [8], where each training shape was normalized with the help of localized coordinates [18].

In the following, let $N_v$ denote the number of vertices on the S-SCAPE model, let $\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{\Theta}) \in \mathbb{R}^{3N_v}$ denote the vector containing the vertex coordinates of identity $\boldsymbol{\beta}$ in posture $\boldsymbol{\Theta}$, and let $\widetilde{\boldsymbol{s}}(\boldsymbol{\beta}, \boldsymbol{\Theta}) \in \mathbb{R}^{4N_v}$ denote the vector containing the corresponding homogeneous vertex coordinates. For the fixed posture $\boldsymbol{\Theta}_0$ that was used to train the identity space, S-SCAPE models the shape change caused by identity using a PCA model as

$$\widetilde{\boldsymbol{s}}(\boldsymbol{\beta}, \boldsymbol{\Theta}_0) = \widetilde{\boldsymbol{A}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\mu}}, \tag{1}$$

where $\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^{4N_v}$ contains the homogeneous coordinates of the mean body shape, $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{4N_v \times d_{id}}$ is the matrix found by PCA, and $d_{id}$ is the dimensionality of the identity shape space. For a fixed identity $\boldsymbol{\beta}_0$, S-SCAPE models the shape change caused by posture using LBS as

$$\boldsymbol{s}_i(\boldsymbol{\beta}_0, \boldsymbol{\Theta}) = \sum_{j=1}^{N_b} \omega_{ij} \boldsymbol{T}_j(\boldsymbol{\Theta}) \widetilde{\boldsymbol{s}}_i(\boldsymbol{\beta}_0, \boldsymbol{\Theta}_0), \tag{2}$$

where $\boldsymbol{s}_i$ and $\widetilde{\boldsymbol{s}}_i$ denote the standard and homogenous coordinate vector of the $i$-th vertex of $\boldsymbol{s}$, $N_b$ denotes the number of bones used for LBS, $\boldsymbol{T}_j(\boldsymbol{\Theta}) \in \mathbb{R}^{3\times 4}$ denotes the transformation matrix applied to the $j$-th bone, and $\omega_{ij}$ denotes the rigging weight binding the $i$-th vertex to the $j$-th bone.

Combining Eq. 1 and 2 in matrix notation leads to

$$\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \boldsymbol{T}(\boldsymbol{\Theta}) \widetilde{\boldsymbol{A}}\boldsymbol{\beta} + \boldsymbol{T}(\boldsymbol{\Theta}) \widetilde{\boldsymbol{\mu}}, \tag{3}$$

where $\boldsymbol{T}(\boldsymbol{\Theta}) \in \mathbb{R}^{3N_v \times 4N_v}$ is a sparse matrix containing the per-vertex transformations. Using this notation, it is easy to see that S-SCAPE is linear in both $\boldsymbol{\beta}$ and $\boldsymbol{T}(\boldsymbol{\Theta})$, which allows for a simple optimization w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$.

## 4 Estimating model parameters for a motion sequence

We start by providing an overview of the proposed method. Fig. 1 shows the different parts of the algorithm visually. Given as input a trained S-SCAPE

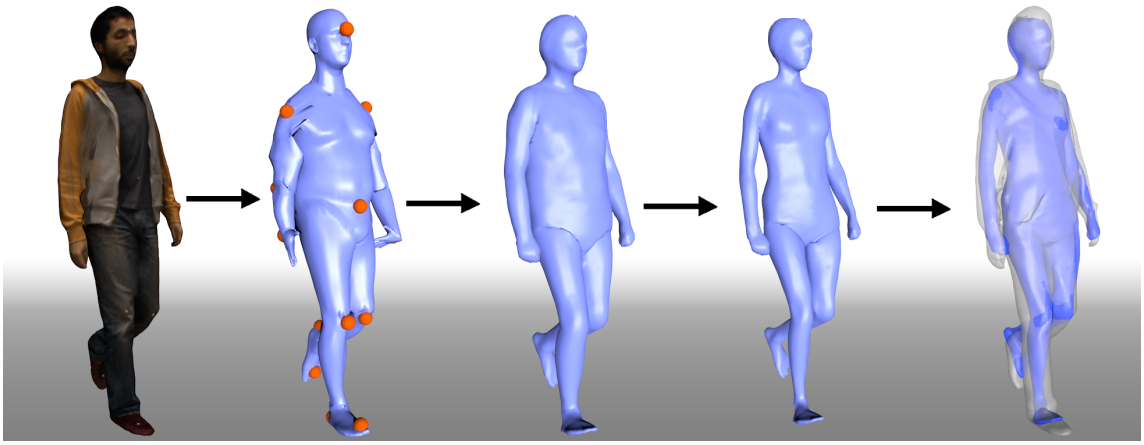


**Fig. 1.** Overview of the proposed pipeline. From left to right: input frame, result of Stitched Puppet [1] with annotated landmarks, result after estimation of initial identity and posture, final result, and overlay of input and final result.

model and a motion sequence consisting of $N_f$ frames $\boldsymbol{F}_i$ represented by triangle meshes with unknown correspondence, we aim to compute a single parameter vector $\boldsymbol{\beta}$ controlling the shape of the identity (as the identity of the person is fixed during motion) along with $N_f$ parameter vectors $\boldsymbol{\Theta}_i$ controlling the postures in each frame, such that $\boldsymbol{s}_i(\boldsymbol{\beta}, \boldsymbol{\Theta}_i)$ is close to $\boldsymbol{F}_i$.

To fit the S-SCAPE model to a single frame $\boldsymbol{F}$, we aim to minimize

$$E(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta}) = \omega_{lnd} E_{lnd}(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta}) + \omega_{data} E_{data}(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta}) + \omega_{cloth} E_{cloth}(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta}) \tag{4}$$

w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ subject to constraints that keep $\boldsymbol{\beta}$ in the learned probability distribution of parameter values. Here, $\omega_{lnd}$, $\omega_{data}$, and $\omega_{cloth}$ are weights that trade off the influence of the different energy terms. The energy $E_{lnd}$ measures the distance between a sparse set of provided landmarks, which correspond to distinctive positions on the human body, to their corresponding locations on $\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{\Theta})$. The provided landmarks are computed automatically in the following. The energy $E_{data}$ measures the distance between $\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ and $\boldsymbol{F}$ using a nearest neighbor cost. The energy $E_{cloth}$ is designed to account for loose clothing by encouraging $\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ to be located inside the observation $\boldsymbol{F}$.

For a motion sequence of $N_f$ frames, our goal is then to minimize

$$E\left(\boldsymbol{F}_{1:N_f},\boldsymbol{\beta},\boldsymbol{\Theta}_{1:N_f}\right) = \sum_{i=1}^{N_f} E\left(\boldsymbol{F}_i,\boldsymbol{\beta},\boldsymbol{\Theta}_i\right) \tag{5}$$

w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}_{1:N_f}$ subject to constraints that keep $\boldsymbol{\beta}$ in the learned probability distribution of parameter values. Here, $\boldsymbol{F}_{1:N_f} = \{\boldsymbol{F}_1, \ldots, \boldsymbol{F}_{N_f}\}$ is the set of frames and $\boldsymbol{\Theta}_{1:N_f} = \{\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{N_f}\}$ is the set of posture parameters. The energy $E_{cloth}$ allows to take advantage of motion cues in this formulation as it encourages the body shape to lie inside all observed frames.

In the following sections, we detail the prior that is used to constrain $\boldsymbol{\beta}$ as well as the different energy terms. Optimizing Eq. 5 w.r.t. all parameters jointly results in a high-dimensional optimization problem that is inefficient to solve

and prone to get stuck in undesirable local minima. After introducing all energy terms, we discuss how this problem can be divided into smaller problems that can be solved in order, thereby allowing to find a good minimum in practice.

### 4.1 Prior model for $\boldsymbol{\beta}$

A prior model is used to ensure that the body shape stays within the learned shape space that represents plausible human shapes. The identity shape space is learned using PCA, and has zero mean and standard deviation $\sigma_i$ along the $i$-th principal component. Similarly to previous work [9], we do not penalize values of $\boldsymbol{\beta}$ that stay within $3\sigma_i$ of the mean to avoid introducing a bias towards the mean shape. However, rather than penalizing a larger distance from the mean, we constrain the solution to lie inside the hyperbox $\pm 3\sigma_i$ using a constrained optimization framework. This constraint can be handled by standard constrained optimizers since the hyperbox is axis-aligned, and using this hard constraint removes the need to appropriately weigh a prior energy w.r.t. other energy terms.

### 4.2 Landmark energy

The landmark energy helps to guide the solution towards the desired local minimum with the help of distinctive anatomical landmarks. This energy is especially important during the early stages of the optimization as it allows to find a good initialization for the identity and posture parameters. In the following, we consider the use of $N_{lnd}$ landmarks and assume without loss of generality that the vertices corresponding to landmarks are the first $N_{lnd}$ vertices of $\boldsymbol{s}$. The landmark term is defined as

$$E_{lnd}\left(\boldsymbol{F}, \boldsymbol{\beta}, \boldsymbol{\Theta}\right) = \sum_{i=1}^{N_{lnd}} \left\| \boldsymbol{s}_i\left(\boldsymbol{\beta}, \boldsymbol{\Theta}\right) - \boldsymbol{l}_i(\boldsymbol{F}) \right\|^2, \tag{6}$$

where $\boldsymbol{l}_i(\boldsymbol{F})$ denotes the $i$-th landmark of frame $\boldsymbol{F}$, $\boldsymbol{s}_i\left(\boldsymbol{\beta}, \boldsymbol{\Theta}\right)$ denotes the vertex corresponding to the $i$-th landmark of $\boldsymbol{s}\left(\boldsymbol{\beta}, \boldsymbol{\Theta}\right)$, and $\|\cdot\|$ denotes the $\ell^2$ norm.

The landmarks $\boldsymbol{l}_i(\boldsymbol{F})$ are computed automatically with the help of the state of the art Stitched Puppet [1], which allows to robustly fit a human body model to a single scan using a particle-based optimization. Specifically, we once manually select a set of vertex indices to be used as landmarks on the Stitched Puppet model, which is then fixed for all experiments. To fit the Stitched Puppet to a single frame, randomly distributed particles are used to avoid getting stuck in undesirable local minima. We fit the Stitched Puppet model to frame $\boldsymbol{F}$, and report the 3D positions of the pre-selected indices after fitting as landmarks $\boldsymbol{l}_i(\boldsymbol{F})$. While the Stitched Puppet aims to fit the body shape and posture of $\boldsymbol{F}$, only the coordinates $\boldsymbol{l}_i(\boldsymbol{F})$ are used by our framework. Note that our method does not require accurate $\boldsymbol{l}_i(\boldsymbol{F})$, since $\boldsymbol{l}_i(\boldsymbol{F})$ are only used to initialize the optimization.

Using many particles on each frame of a motion sequence is inefficient. Furthermore, since the Stitched Puppet is trained on a database of minimally

dressed subjects, using many particles to fit to a frame in wide clothing may lead to overfitting problems. This is illustrated in Fig. 2. To remedy this, we choose to use a relatively small number of particles which is set to 30. Starting at the second frame, we initialize the particle optimization to the result of the previous frame to guide the optimization towards the desired optimum.

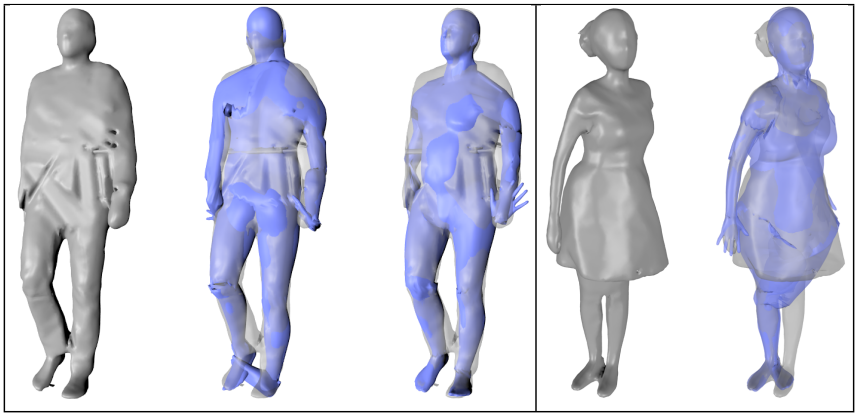


**Fig. 2.** Left: overfitting problem of Stitched Puppet in the presence of clothing. Input frame, Stitched Puppet result with 160 particles, and Stitched Puppet result with 30 particles are shown in order. Right: the failure case from our database caused by mismatching of Stitched Puppet.

### 4.3 Data energy

The data energy pulls the S-SCAPE model towards the observation $\boldsymbol{F}$ using a nearest neighbor term. This energy, which unlike the landmark energy considers all vertices of $\boldsymbol{s}$, is crucial to fit the identity and posture of $\boldsymbol{s}$ to the input $\boldsymbol{F}$ as

$$E_{data}\left(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta}\right)=\sum_{i=1}^{N_v}\delta_{\boldsymbol{NN}}\left\|\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right)-\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right),\boldsymbol{F}\right)\right\|^2, \tag{7}$$

where $N_v$ denotes the number of vertices of $\boldsymbol{s}$ and $\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right),\boldsymbol{F}\right)$ denotes the nearest neighbour of vertex $\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right)$ on $\boldsymbol{F}$. To remove the influence of outliers and reduce the possibility of nearest neighbour mismatching, we use a binary weight $\delta_{\boldsymbol{NN}}$ that is set to one if the distance between $\boldsymbol{s}_i$ and its nearest neighbor on $\boldsymbol{F}$ is below $200mm$ and the angle between their outer normal vectors is below 60°, and to zero otherwise.

### 4.4 Clothing energy

The clothing energy is designed to encourage the predicted body shape $\boldsymbol{s}$ to be located entirely inside the observation $\boldsymbol{F}$. This energy is particularly important when considering motion sequences acquired with loose clothing. In such cases, merely using $E_{lnd}$ and $E_{data}$ leads to results that overestimate the circumferences

of the body shape because $\boldsymbol{\beta}$ is estimated to fit to $\boldsymbol{F}$ rather than to fit inside of $\boldsymbol{F}$, see Fig. 3. To remedy this, we define the clothing energy as

$$E_{cloth}(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta})=\sum_{i=1}^{N_v}\delta_{out}\delta_{\boldsymbol{NN}}\left\|\boldsymbol{s}_i(\boldsymbol{\beta},\boldsymbol{\Theta})-\boldsymbol{NN}\left(\boldsymbol{s}_i(\boldsymbol{\beta},\boldsymbol{\Theta}),\boldsymbol{F}\right)\right\|^2+\omega_r\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\right\|^2, \quad (8)$$

where $\delta_{out}$ is used to identify vertices of $\boldsymbol{s}$ located outside of $\boldsymbol{F}$. This is achieved by setting $\delta_{out}$ to one if the angle between the outer normal of $\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right),\boldsymbol{F}\right)$ and the vector $\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right)-\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right),\boldsymbol{F}\right)$ is below 90°, and to zero otherwise. Furthermore, $\omega_r$ is a weight used for the regularization term, and $\boldsymbol{\beta}_0$ is an initialization of the identity parameters used to constrain $\boldsymbol{\beta}$.

When observing a human body dressed in loose clothing in motion, different frames can provide valuable cues about the true body shape. The energy $E_{cloth}$ is designed to exploit motion cues when optimizing $E_{cloth}$ w.r.t. all available observations $\boldsymbol{F}_i$. This allows to account for clothing using a simple optimization without the need to find skin and non-skin regions as in previous work [9,19,7]. The regularization $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2$ used in Eq. 8 is required to avoid excessive thinning of limbs due to small misalignments in posture.

Fig. 3 shows the influence of $E_{cloth}$ on the result of a walking sequence in layered clothing. The left side shows overlays of the input and the result for $\omega_{cloth}=0$ and $\omega_{cloth}=1$. Note that while circumferences are overestimated when $\omega_{cloth}=0$, a body shape located inside the input frame is found for $\omega_{cloth}=1$. The comparison to the ground truth body shape computed as discussed in Sec. 6 is visualized in the middle and the right of Fig. 3, and shows that $E_{cloth}$ leads to a significant improvement of the accuracy of $\boldsymbol{\beta}$.

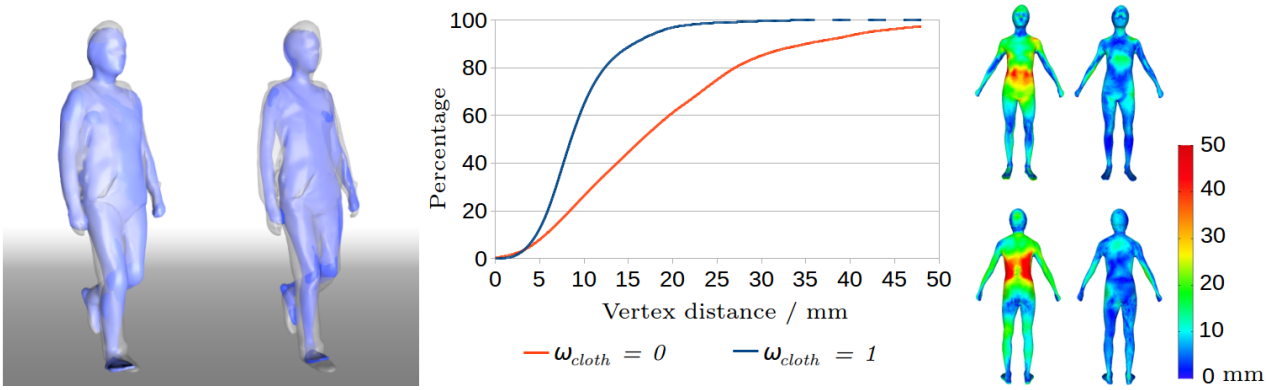



**Fig. 3.** Influence of $E_{cloth}$ on walking sequence. Left: input data overlayed with result with $\omega_{cloth}=0$ (left) and $\omega_{cloth}=1$ (right). Middle: cumulative per-vertex error of estimated body shape with $\omega_{cloth}=0$ and $\omega_{cloth}=1$. Right: color-coded per-vertex error with $\omega_{cloth}=0$ (left) and $\omega_{cloth}=1$ (right).

### 4.5 Optimization schedule

Minimizing $E\left(\boldsymbol{F}_{1:N_f},\boldsymbol{\beta},\boldsymbol{\Theta}_{1:N_f}\right)$ defined in Eq. 5 over all $N_f$ frames w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}_i$ jointly is not feasible when considering motion sequences containing

hundreds of frames as this is a high-dimensional optimization problem. To solve this problem without getting stuck in undesirable local minima, we optimize three smaller problems in order.

**Initial identity estimation.** We start by computing an initial estimate $\boldsymbol{\beta}_0$ based on the first $N_k$ frames of the sequence by optimizing $E\left(\boldsymbol{F}_{1:N_k}, \boldsymbol{\beta}, \boldsymbol{\Theta}_{1:N_k}\right)$ w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}_i$. For increased efficiency, we start by computing optimal $\boldsymbol{\beta}_i$ and $\boldsymbol{\Theta}_i$ for each frame using Eq. 4 by alternating the optimization of $\boldsymbol{\Theta}_i$ for fixed $\boldsymbol{\beta}_i$ with the optimization of $\boldsymbol{\beta}_i$ for fixed $\boldsymbol{\Theta}_i$. This is repeated for $N_{it}$ iterations. Temporal consistency is achieved by initializing $\boldsymbol{\Theta}_{i+1}$ as $\boldsymbol{\Theta}_i$ and $\boldsymbol{\beta}_{i+1}$ as $\boldsymbol{\beta}_i$ starting at the second frame. As it suffices for the identity parameters to roughly estimate the true body shape at this stage, we set $\omega_{cloth} = 0$. In the first iterations, $E_{lnd}$ is essential to guide the fitting towards the correct local optimum, while in later iterations $E_{data}$ gains in importance. We therefore set $\omega_{data} = 1 - \omega_{lnd}$ and initialize $\omega_{lnd}$ to one. We linearly reduce $\omega_{lnd}$ to zero in the last two iterations. We then initialize the posture parameters to the computed $\boldsymbol{\Theta}_i$, and the identity parameters to the mean of the computed $\boldsymbol{\beta}_i$ and iteratively minimize $E\left(\boldsymbol{F}_{1:N_k}, \boldsymbol{\beta}, \boldsymbol{\Theta}_{1:N_k}\right)$ w.r.t. $\boldsymbol{\Theta}_{1:N_k}$ and $\boldsymbol{\beta}$. This leads to stable estimates for $\boldsymbol{\Theta}_{1:N_k}$ and an initial estimate of the identity parameter, which we denote by $\boldsymbol{\beta}_0$ in the following.

**Posture estimation.** During the next stage of our framework, we compute the posture parameters $\boldsymbol{\Theta}_{N_k+1:N_f}$ for all remaining frames by sequentially minimizing Eq. 4 w.r.t. $\boldsymbol{\Theta}_i$. As before, $\boldsymbol{\Theta}_{i+1}$ is initialized to the result of $\boldsymbol{\Theta}_i$. As the identity parameters are not accurate at this stage, we set $\omega_{cloth} = 0$. For each frame, the energy is optimized $N_{it}$ times while reducing the influence of $\omega_{lnd}$ in each iteration, using the same weight schedule as before. This results in posture parameters $\boldsymbol{\Theta}_i$ for each frame.

**Identity refinement.** In a final step, we refine the identity parameters to be located inside all observed frames $\boldsymbol{F}_{1:N_f}$. To this end, we initialize the identity parameters to $\boldsymbol{\beta}_0$, fix all posture parameters to the computed $\boldsymbol{\Theta}_i$, and minimize $E\left(\boldsymbol{F}_{1:N_f}, \boldsymbol{\beta}, \boldsymbol{\Theta}_{1:N_f}\right)$ w.r.t. $\boldsymbol{\beta}$. As the landmarks and observations are already fitted adequately, we set $\omega_{lnd} = \omega_{data} = 0$ at this stage of the optimization.

## 5 Implementation details

The S-SCAPE model used in this work consists of $N_v = 6449$ vertices, and uses $d_{id} = 100$ parameters to control identity and $d_{pose} = 30$ parameters to control posture by rotating the $N_b = 15$ bones. The bones, posture parameters, and rigging weights are set as in the published model [8].

For the Stitched Puppet, we use 60 particles for the first frame, and 30 particles for subsequent frames. We use a total of $N_{lnd} = 14$ landmarks that have been shown sufficient for the initialization of posture fitting [6], and are located at forehead, shoulders, elbows, wrists, knees, toes, heels, and abdomen. Fig. 1 shows the chosen landmarks on the Stitched Puppet model. During the optimization, we set $N_{it} = 6$ and $N_k = 25$ . The optimization w.r.t. $\boldsymbol{\beta}$ uses analytic gradients, and we use Matlab L-BFGS-B to optimize the energy. The setting of the regularization weight $\omega_r$ depends on the clothing style. The looser the clothing, the smaller $\omega_r$, as this allows for more corrections of the identity parameters. In our experiments, we use $\omega_r = 1$ for all the sequences with layered and wide clothing in our dataset.

## 6 Evaluation

### 6.1 Dataset

This section introduces the new dataset we acquired to allow quantitative evaluation of human body shape estimation from dynamic data. The dataset consists of synchronized acquisitions of dense unstructured geometric motion data and sparse motion capture (MoCap) data of 6 subjects (3 female and 3 male) captured in 3 different motions and 3 clothing styles each. The geometric motion data are sequences of meshes obtained by applying a visual hull reconstruction to a 68-color-camera (4M pixels) system at 30FPS. The basic motions that were captured are walk, rotating the body, and pulling the knees up. The captured clothing styles are very tight, layered (long-sleeved layered clothing on upper body), and wide (wide pants for men and dress for women). The body shapes of 6 subjects vary significantly. Fig. 4 shows some frames of the database.

To evaluate algorithms using this dataset, we can compare the body shapes estimated under loose clothing with the tight clothing baseline. The comparison is done per vertex on the two body shapes under the same normalized posture. Cumulative plots are used to show the results.

**Fig. 4.** Six representative examples of frames of our motion database. From left to right, a female and male subject is shown for tight, layered, and wide clothing each.

### 6.2 Evaluation of posture and shape fitting

We applied our method to all sequences in the database. For one sequence of a female subject captured while rotating the body in wide clothing, Stitched Puppet fails to find the correct posture, which leads to a failure case of our method (see Fig. 2). We exclude this sequence from the following evaluation.

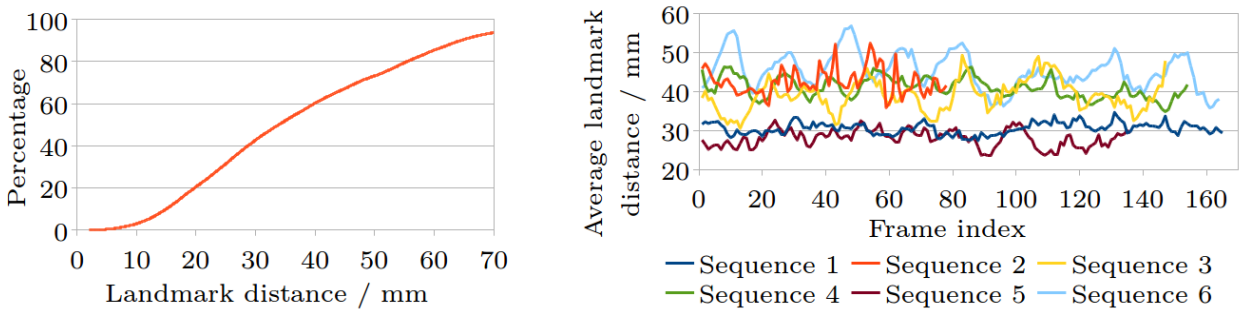



**Fig. 5.** Accuracy of posture estimation over the walking sequences in tight clothing. Left: cumulative landmark errors. Right: average landmark error for each sequence.

To evaluate the accuracy of the posture parameters $\boldsymbol{\Theta}$, we compare the 3D locations of a sparse set of landmarks captured using a MoCap system with the corresponding model vertices of our estimate. This evaluation is performed in very tight clothing, as no accurate MoCap markers are available for the remaining clothing styles. Fig. 5 summarizes the per-marker errors over the walking sequences of all subjects. The results show that most of the estimated landmarks are within $35mm$ of the ground truth and that our method does not suffer from drift for long sequences. As the markers on the Stitched Puppet and the MoCap markers were placed by non-experts, the landmark placement is not fully repeatable, and errors of up to $35mm$ are considered fairly accurate.

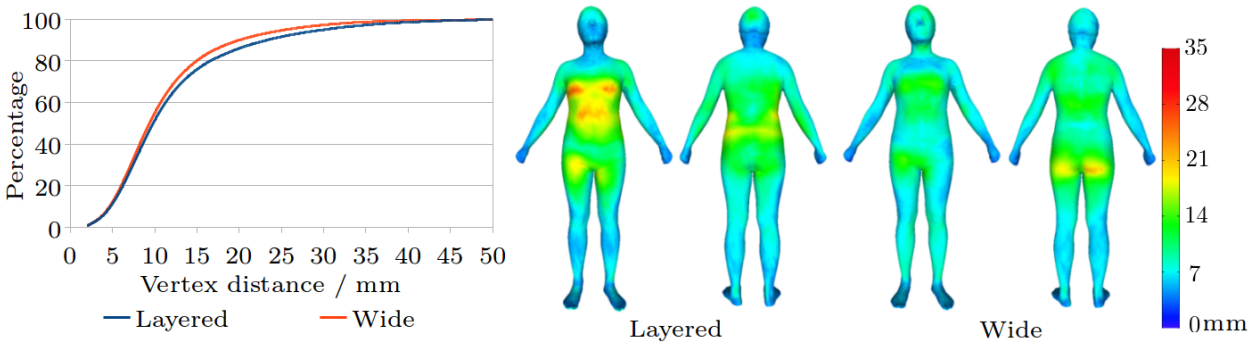



**Fig. 6.** Summary of shape accuracy computed over the frames of all motion sequences of all subjects captured in layered and wide clothing. Left: cumulative plots showing the per-vertex error. Right: mean per-vertex error color-coded from blue to red.

To evaluate the accuracy of the identity parameters $\boldsymbol{\beta}$, we use for each subject the walking sequence captured in very tight clothing to establish a ground truth

identity $\boldsymbol{\beta}_0$ by applying our shape estimation method. Applying our method to sequences in looser clothing styles of the same subject leads identity parameters $\boldsymbol{\beta}$, whose accuracy can be evaluated by comparing the 3D geometry of $\boldsymbol{s}\left(\boldsymbol{\beta}_0, \boldsymbol{\Theta}_0\right)$ and $\boldsymbol{s}\left(\boldsymbol{\beta}, \boldsymbol{\Theta}_0\right)$ for a standard posture $\boldsymbol{\Theta}_0$.

Fig. 6 summarizes the per-vertex errors over all motion sequences captured in layered and wide clothing, respectively. The left side shows the cumulative errors, and the right side shows the color-coded mean per-vertex error. The color coding is visualized on the mean identity of the training data. The result shows that our method is robust to loose clothing with more than 50% of all the vertices having less than $10mm$ error for both layered and wide clothing. The right side shows that as expected, larger errors occur in areas where the shape variability across different identities is high.

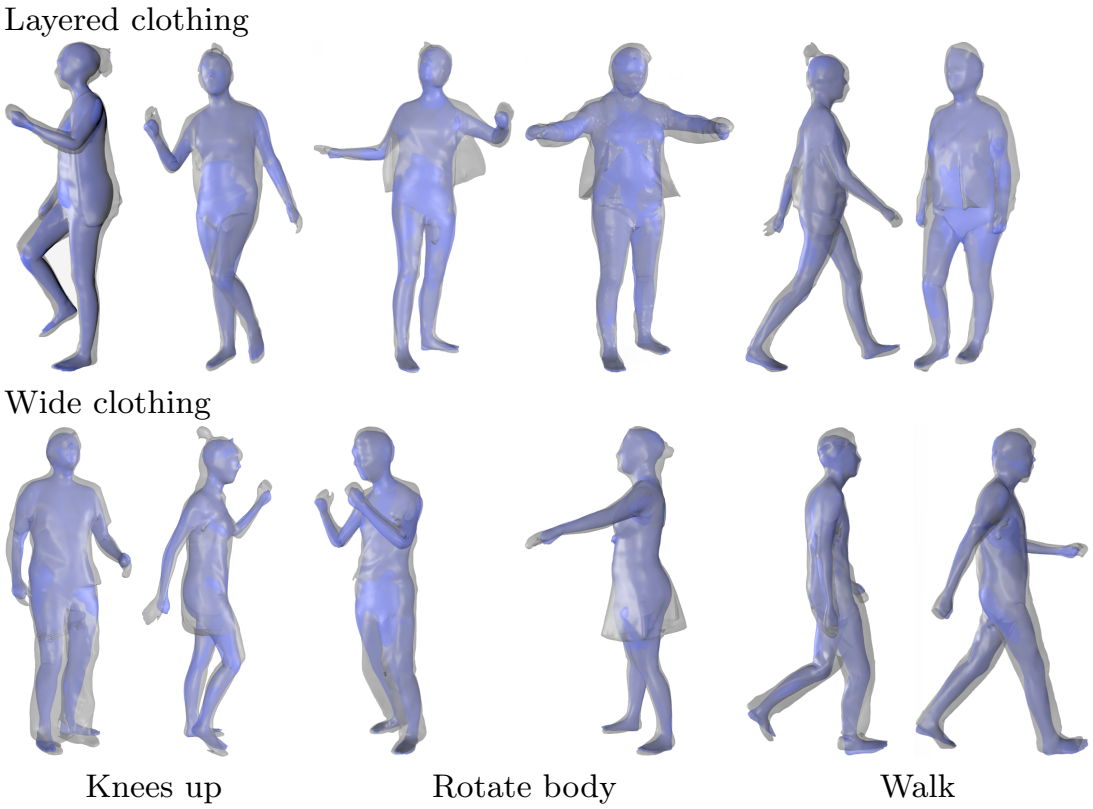



**Fig. 7.** Overlay of input data and our result.

Fig. 7 shows some qualitative results for all three types of motions and two clothing styles. Note that accurate body shape estimates are obtained for all frames. Consider the frame that shows a female subject performing a rotating motion in layered clothing. Computing a posture or shape estimate based on this frame is extremely challenging as the geometry of the layered cloth locally resembles the geometry of an arm, and as large portions of the body shape are occluded. Our method successfully leverages temporal consistency and motion cues to find reliable posture and body shape estimates.

### 6.3 Comparative evaluation

As we do not have results on motion sequences with ground truth for existing methods, this section presents visual comparisons, shown in Fig. 8. We compare to Wuhrer et al. [6] on the dancer sequence [20] presented in their work. Note that unlike the results of Wuhrer et al., our shape estimate does not suffer from unrealistic bending at the legs even in the presence of wide clothing. Furthermore, we compare to Neophytou and Hilton [7] on the swing sequence [21] presented in their work. Note that we obtain results of similar visual quality without the need for manual initializations and pre-aligned motion sequences. In summary, we present the first fully automatic method for body shape and motion estimation, and show that this method achieves state of the art results.

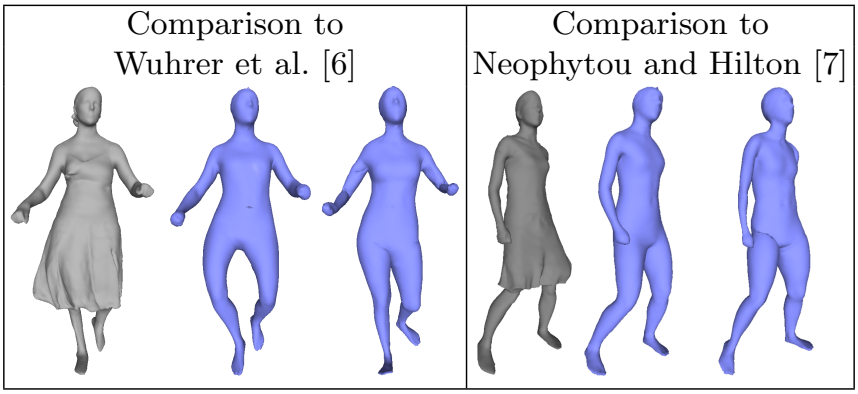



**Fig. 8.** Per comparison from left to right: input, result of prior works, our result.

## 7 Conclusion

We presented an approach to automatically estimate the human body shape under motion based on a 3D input sequence showing a dressed person in possibly loose clothing. The accuracy of our method was evaluated on a newly developed benchmark[1] containing 6 different subjects performing 3 motions in 3 different styles each. We have shown that, although being fully automatic, our posture and shape estimation achieves state of the art performance. In the future, the body shape and motion estimated by our algorithm have the potential to aid in a variety of tasks including virtual change rooms and security applications.

## Acknowledgments

Funded by France National Research grant ANR-14-CE24-0030 ACHMOV. We thank Yannick Marion for help with code to efficiently fit an S-SCAPE model to a single frame, Leonid Pishchulin for helpful discussions, Alexandros Neophytou and Adrian Hilton for providing comparison data, and Mickaël Heudre, Julien Pansiot and volunteer subjects for help acquiring the database. The database was acquired using the Kinovis platform: `http://kinovis.inrialpes.fr/`.

[1] The benchmark can be downloaded at http://dressedhuman.gforge.inria.fr/.

## References


1. Zuffi, S., Black, M.: The Stitched Puppet: A graphical model of 3d human shape and pose. In: Conference on Computer Vision and Pattern Recognition. (2015) 3537–3546
2. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Conference on Computer Vision and Pattern Recognition. (2015) 343–352
3. Bogo, F., Black, M.J., Loper, M., Romero, J.: Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In: ICCV. (2015)
4. Weiss, A., Hirshberg, D., Black, M.: Home 3D body scans from noisy image and range data. In: International Conference on Computer Vision. (2011) 1951–1958
5. Helten, T., Baak, A., Bharai, G., Müller, M., Seidel, H.P., Theobalt, C.: Personalization and evaluation of a real-time depth-based full body scanner. In: International Conference on 3D Vision. (2013) 279–286
6. Wuhrer, S., Pishchulin, L., Brunton, A., Shu, C., Lang, J.: Estimation of human body shape and posture under clothing. Computer Vision and Image Understanding **127** (2014) 31–42
7. Neophytou, A., Hilton, A.: A layered model of human body and garment deformation. In: International Conference on 3D Vision. (2014) 171–178
8. Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., Schiele, B.: Building statistical shape spaces for 3d human modeling. Technical Report 1503.05860, arXiv (2015)
9. Balan, A.O., Black, M.J.: The naked truth: Estimating body shape under clothing. In: European Conference on Computer Vision. (2008) 15–29
10. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: shape completion and animation of people. ACM Transactions on Graphics **24**(3) (2005) 408–416 Proceedings of SIGGRAPH.
11. Chen, Y., Liu, Z., Zhang, Z.: Tensor-based human body modeling. In: Conference on Computer Vision and Pattern Recognition. (2013) 105–112
12. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P.: A statistical model of human pose and body shape. Computer Graphics Forum **28**(2) (2009) 337–346 Proceedings of Eurographics.
13. Jain, A., Thormählen, T., Seidel, H.P., Theobalt, C.: MovieReshape: tracking and reshaping of humans in videos. ACM Transactions on Graphics **29** (2010) 148:1–10 Proceedings of SIGGRAPH Asia.
14. Neophytou, A., Hilton, A.: Shape and pose space deformation for subject specific animation. In: International Conference on 3D Vision. (2013) 334–341
15. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.: SMPL: A skinned multi-person linear model. Transactions on Graphics **34**(6) (2015) 248:1–248:16 Proceedings of SIGGRAPH Asia.
16. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.: DYNA: a model of dynamic human shape in motion. Transactions on Graphics **34**(4) (2015) #120:1–14 Proceedings of SIGGRAPH.
17. Hasler, N., Stoll, C., Rosenhahn, B., Thormählen, T., Seidel, H.P.: Estimating body shape of dressed humans. Computers & Graphics **33**(3) (2009) 211–216 Proceedings of Shape Modeling International.
18. Wuhrer, S., Shu, C., Xi, P.: Posture-invariant statistical shape analysis using Laplace operator. Computers & Graphics **36**(5) (2012) 410–416 Proceedings of Shape Modeling International.

19. Stoll, C., Gall, J., de Aguiar, E., Thrun, S., Theobalt, C.: Video-based reconstruction of animatable human characters. ACM Transactions on Graphics **29**(6) (2010) #139:1–10 Proceedings of SIGGRAPH Asia.
20. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. ACM Transactions on Graphics **27**(3) (2008) #98,1–10 Proceedings of SIGGRAPH.
21. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. ACM Transactions on Graphics **27**(3) (2008) #97:1–10 Proceedings of SIGGRAPH.

$\therefore$

## A.5 EIGEN APPEARANCE MAPS OF DYNAMIC SHAPES

Adnane Boukhayma, Vagia Tsiminaki, Jean-Sébastien Franco, Edmond Boyer ECCV 2016 - European Conference on Computer Vision, Oct 2016, Amsterdam, Netherlands.

HAL
archives-ouvertes.fr

# Eigen Appearance Maps of Dynamic Shapes

Adnane Boukhayma, Vagia Tsiminaki, Jean-Sébastien Franco, Edmond Boyer

▶ **To cite this version:**

Adnane Boukhayma, Vagia Tsiminaki, Jean-Sébastien Franco, Edmond Boyer. Eigen Appearance Maps of Dynamic Shapes. ECCV 2016 - European Conference on Computer Vision, Oct 2016, Amsterdam, Netherlands. pp.230-245, 10.1007/978-3-319-46448-0_14 . hal-01348837

**HAL Id: hal-01348837**
**https://hal.inria.fr/hal-01348837**

Submitted on 7 Oct 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Eigen Appearance Maps of Dynamic Shapes

Adnane Boukhayma, Vagia Tsiminaki, Jean-Sébastien Franco, Edmond Boyer

LJK, Université Grenoble Alpes, Inria Grenoble Rhône-Alpes, France
firstname.lastname@inria.fr

**Abstract.** We address the problem of building efficient appearance representations of shapes observed from multiple viewpoints and in several movements. Multi-view systems now allow the acquisition of spatio-temporal models of such moving objects. While efficient geometric representations for these models have been widely studied, appearance information, as provided by the observed images, is mainly considered on a per frame basis, and no global strategy yet addresses the case where several temporal sequences of a shape are available. We propose a per subject representation that builds on PCA to identify the underlying manifold structure of the appearance information relative to a shape. The resulting eigen representation encodes shape appearance variabilities due to viewpoint and motion, with Eigen textures, and due to local inaccuracies in the geometric model, with Eigen warps. In addition to providing compact representations, such decompositions also allow for appearance interpolation and appearance completion. We evaluate their performances over different characters and with respect to their ability to reproduce compelling appearances in a compact way.

## 1 Introduction

The last decade has seen the emergence of 3D dynamic shape models of moving objects, in particular humans, acquired from multiple videos. These spatio-temporal models comprise geometric and appearance information extracted from images, and they allow for subject motions to be recorded and reused. This is of interest for applications that require real 3D contents for analysis, free viewpoint and animation purposes and also for interactive experiences made possible with new virtual reality devices. This ability to now record datasets of subject motions bolsters the need for shape and appearance representations that make optimal use of the massive amount of image information usually produced. While dynamic shape representations have been extensively studied, from temporally coherent representations over a single sequence, to shape spaces that can encode both pose and subject variabilities over multiple sequences and multiple subjects, appearance representations have received less attention in this context. In this paper, we investigate this issue.

Currently, appearance information is still most often estimated and stored once per frame, e.g. a texture map associated to a 3D model [1], and the leap to an efficient temporal appearance representation is still a largely open problem.

This is despite the obvious redundancy with which the appearance of subjects is observed, across temporal frames, different viewpoints of the same scene, and often several sequences of the same subject performing different actions or motions. At the opposite of the spectrum, and given registered geometries, one can store only one texture for a sequence or even for a subject in several sequences, hence dramatically reducing sizes, but in so doing would drop the ability to represent desirable appearance variations, such as change in lighting or personal expression of the subject.

In this paper, we advance this aspect by providing a view-independent appearance representation and estimation algorithm, to encode the appearance variability of a dynamic subject, observed over one or several temporal sequences. Compactly representing image data from all frames and viewpoints of the subject can be seen as a non-linear dimensionality reduction problem in image space, where the main non-linearities are due to the underlying scene geometry. Our strategy is to remove these non-linearities with state-of-the-art geometric and image-space alignment techniques, so as to reduce the problem to a single texture space, where the remaining image variabilities can be straightforwardly identified with PCA and thus encoded as Eigen texture combinations. To this goal, we identify two geometric alignment steps. First, we coarsely register geometric shape models of all time frames to a single shape template, for which we pre-computed a single reference surface-to-texture unwrapping. Second, to cope with remaining fine-scale misalignments due to registration errors, we estimate realignment warps in the texture domain. Because they encode low-magnitude, residual geometric variations, they are also advantageously decomposed using PCA, yielding Eigen warps. The full appearance information of all subject sequences can then be compactly stored as linear combinations of Eigen textures and Eigen warps. Our strategy can be seen as a generalization of the popular work of Nishino *et. al.* [2], which introduces Eigen textures to encode appearance variations of a static object under varying viewing conditions, to the case of fully dynamic subjects with several viewpoints and motions.

The pipeline is shown to yield effective estimation performance. In addition, the learned texture and warp manifolds allow for efficient generalizations, such as texture interpolations to generate new unobserved content from blended input sequences, or completions to cope with missing observations due to *e.g.* occlusions. To summarize our main contribution is to propose and evaluate a new appearance model that specifically addresses dynamic scene modeling by accounting for both appearance changes and local geometric inaccuracies.

## 2 Related work

Obtaining appearance of 3D models from images was first tackled from static images for inanimate objects, e.g. [3, 2], a case largely explored since e.g. [4, 5]. The task also gained interest for the case of subjects in motion, e.g. for human faces [6]. With the advent of full body capture and 3D interaction systems [7,

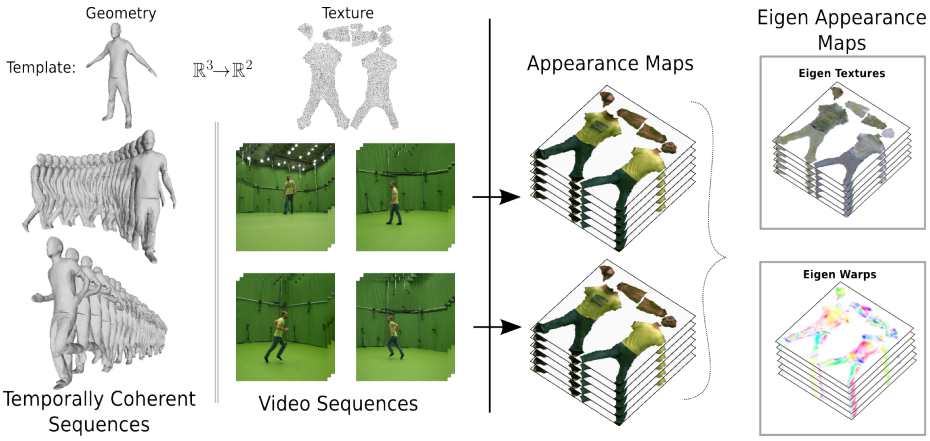



Fig. 1: Overview: Time consistent shape modeling provides datasets of appearance maps. Our proposed method exploits the manifold structure of these appearance information through PCA decomposition to generate the Eigen appearance maps relative to a shape.

1] the task of recovering appearance has become a key issue, as the appearance vastly enhances the quality of restitution of acquired 3D models.

A central aspect of the problem is how to represent appearance, while achieving a proper trade-off between storage size and quality. 3D capture traditionally generates full 3D reconstructions, albeit of inconsistent topology across time. In this context the natural solution is to build a representation per time frame which uses or maps to that instant's 3D model. Such per instant representations come in two main forms. View-dependent texturing stores and resamples from each initial video frame [8], eventually with additional alignments to avoid ghosting effects [9]. This strategy creates high quality restitutions managing visibility issues on the fly, but is memory costly as it requires storing all images from all viewpoints. On the other hand, one can compute a single appearance texture map from the input views in an offline process [1], reducing storage but potentially introducing sampling artifacts. These involve evaluating camera visibility and surface viewing angles to patch and blend the view contributions in a single common mapping space. To overcome the resolution and sampling limitations, 3D superresolution techniques have been devised that leverage the viewpoint multiplicity to build such maps with enhanced density and quality [10–12].

In recent years, a leap has been made in the representation of 3D surfaces captured, as they can now be estimated as a deformed surface of time-coherent topology [13, 14]. This in turns allows any surface unwrapping and mapping to be consistently propagated in time, however in practice existing methods have only started leveraging this aspect. Tsiminaki *et. al.* [11] examines small temporal segments for single texture resolution enhancement. Volino *et. al.* [15] uses a view-based multi-layer texture map representation to favour view-dependant dy-

namic appearance, using some adjacent neighbouring frames. Collet *et. al.* [1] use tracked surfaces over small segments to improve compression rates of mesh and texture sequences. Methods are intrinsically limited in considering longer segments because significant temporal variability then appears due to light change and movement. While global geometry consistency has been studied [16–18], most such works were primarily aimed at animation synthesis using mesh data, and do not propose a global appearance model for sequences. In contrast, we propose an analysis and representation spanning full sequences and multiples sequences of a subject.

For this purpose, we build an Eigen texture and appearance representation that extends concepts initially explored for faces and static objects [19, 6, 20, 2]. Eigenfaces [19] were initially used to represent the face variability of a population for recognition purposes. The concept was broadened to built a 3D generative model of human faces both in the geometry and texture domains, using the fact that the appearance and geometry of faces are well suited to learning their variability as linear subspaces [6]. Cootes *et. al.* [20] perform the linear PCA analysis of appearance and geometry landmarks jointly in their active appearance model. Nishino *et. al.* [2] instead use such linear subspaces to encode the appearance variability of static objects under light and viewpoint changes at the polygon level. We use linear subspaces for full body appearance and over multiple sequences. Because the linear assumption doesn't hold for whole body pose variation, we use state of the art tracking techniques [21] to remove the non-linear pose component by aligning a single subject-specific template to all the subject's sequence. This in turn allows to model the appearance in a single mapping space associated to the subject template, where small geometric variations and appearances changes can then be linearly modeled.

## 3 Method

To eliminate the main geometric non-linearity, we first align sequence geometries to a single template shape and extract the texture maps of a subject over different motion sequences in a common texture space using a state-of-the-art method [11]. Other per-frame texture extractions may be considered. From these subject specific textures, Eigen textures and Eigen warps that span the appearance space are estimated. The main steps of the method below are depicted in Figure 2 and detailed in the following sections.

1. Texture deformation fields that map input textures to, and from, their aligned versions are estimated using optical flows. Given the deformation fields, Poisson reconstruction is used to warp textures.
2. PCA is applied to the aligned maps and to the texture warps to generate the Eigen textures and the Eigen warps that encode the appearance variations due to, respectively, viewpoint, illumination, and geometric inaccuracies in the reference model.

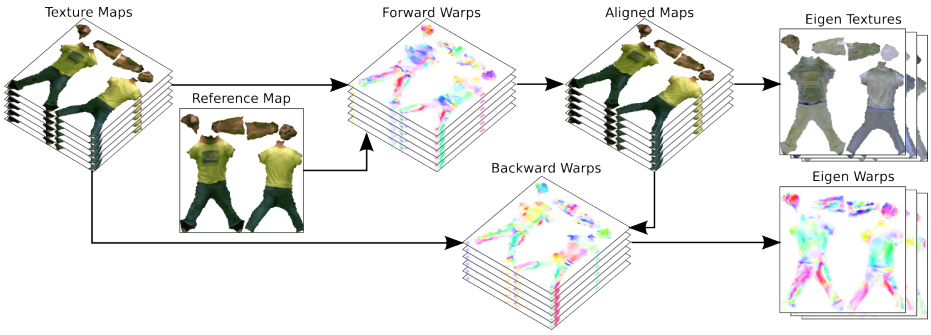



Fig. 2: Method pipeline from input textures (left) to eigen maps (right).

Hence, The main modes of variation of aligned textures and deformation fields, namely Eigen textures and Eigen warps respectively, span the appearance space in our representation. The main steps of this method are depicted in Figure 2 and detailed in the following sections.

Note that due to texture space discretization, the warps between textures are not one-to-one and, in practice, two separate sets of warps are estimated. Forward warps map the original texture maps to the reference map. Backward warps map the aligned texture maps back to the corresponding input textures (see Figure 2).

### 3.1 Aligning texture maps

Appearance variations that are due to viewpoint and illumination changes are captured through PCA under linearity assumption for these variations. To this purpose, textures are first aligned in order to reduce geometric errors resulting from calibration, reconstruction and tracking imprecisions. Such alignment is performed using optical flow, as described below, and with respect to a reference map taken from the input textures. An exhaustive search of the best reference map with the least total alignment error over all input textures is prohibitive since it requires $N^2$ alignments given $N$ input textures. We follow instead a medoid shift strategy over the alignment errors.

The alignment algorithm (see Algorithm 1) first initializes the reference map as one texture from the input set. All texture maps are then aligned to this reference map, and the alignment error is computed as the cumulative sum of squared pixel differences between the reference and the aligned texture maps. The medoid over the aligned texture maps, with respect to alignment error, then identifies the new reference map. These two steps, alignment and medoid shift, are iterated until the total alignment error stops decreasing.

**Data:** Texture maps $\{I_k\}_{k\in[\![1..N]\!]}$
**Result:** Reference map $A_{ref}$, aligned textures $A_k$
$A_{ref}, e_0$ initializations;
**while** $e_i < e_{i-1}$ **do**
  Compute alignment warps: $\{w_k\}_{k\in[\![1..N]\!]}$ s.t. $A_{ref} \approx I_k(x + w_k)$;
  Align texture maps: $A_k = I_k(x + w_k)$;
  Update alignment error: $e_i = \sum_k \|A_k - A_{ref}\|^2$;
  Set $A_{ref}$ as the texture that gives the medoid of the aligned textures:
    $A_{ref} = I_{k_0}$ s.t. $k_0 = arg\min_k \sum_l \|A_k - A_l\|^2$;
**end**

**Algorithm 1:** Texture alignment with iterative reference map selection.

**Dense texture correspondence with optical flow** The warps $\{w_k\}$ in the alignment algorithm, both forward and backward in practice, are estimated as dense pixel correspondences with an optical flow method [22]. We mention here that the optical flow assumptions: brightness consistency, spatial coherency and temporal persistence, are not necessarily verified by the input textures. In particular, the brightness consistency does not hold if we assume appearance variations with respect to viewpoint and illumination changes. To cope with this in the flow estimation, we use histogram equalization as a preprocessing step, which presents the benefit of enhancing contrast and edges within images. Additionally, local changes in intensities are reduced using bilateral filtering, which smooths low spatial-frequency details while preserving edges.

**Texture warping** Optical flows give dense correspondences $\{w\}$ between the reference map and the input textures. To estimate the aligned textures $\{A\}$, we cast the problem as an optimization that seeks the texture map which, once moved according to $w$, best aligns with the considered input texture both in the color and gradient domains. Our experiments show that solving over both color and gradient domains significantly improves results as it tends to better preserve edges than with colors only. This is also demonstrated in works that use the Poisson editing for image composition, e.g. [23, 24] or interpolation, e.g. [25, 26]. We follow here a similar strategy.

We are given an input texture map $I$, a dense flow $w$ from $A_{ref}$ to $I$, and the gradient image $\nabla I$. The aligned texture $A$ of $I$ with respect to $A_{ref}$ is then the map that minimizes the following term:

$$E(A) = \Sigma_x \|\nabla^2 A(x) - \overrightarrow{\nabla}.\nabla I(x + w)\|^2 + \lambda \|A(x) - I(x + w)\|^2, \qquad (1)$$

where $\nabla^2$ is the Laplacian operator, $\overrightarrow{\nabla}.$ the divergence operator, and $x$ denotes pixel locations in texture maps. The weight $\lambda$ balances the influence of color and gradient information. In our experiments, we found that the value 0.02 gives the best results with our datasets.

Using a vector image representation, the above energy can be minimized by solving, in the least-squares sense, the overdetermined $2N \times N$ system below, where $N$ is the active region size of texture maps:

$$\begin{pmatrix} L \\ \Lambda \end{pmatrix} A = \begin{pmatrix} \overrightarrow{\nabla}.\nabla I(x+w) \\ \Lambda I(x+w) \end{pmatrix}, \tag{2}$$

where $L$ is the linear Laplacian operator and $\Lambda = diag_N(\lambda)$. A solution for $A$ is easily found by solving the associated normal equations:

$$\left(L^T L + \Lambda^2\right) A = L^T \overrightarrow{\nabla}.\nabla I(x+w) + \Lambda^2 I(x+w). \tag{3}$$

Figure 3 shows an example where a texture map is warped, given a warp field, using both direct pixel remapping and Poisson warping. The latter strategy achieves visually more compelling and edge preserving results.


Fig. 3: Poisson versus direct texture warping.

### 3.2 Eigen Textures and Eigen Warps

Once the aligned textures and the warps are estimated, we can proceed with the statistical analysis of appearances. Given the true geometry of shapes and their motions, texture map pixels could be considered as shape appearance samples over time and PCA applied directly to the textures would then capture appearance variability. In practice, incorrect geometry causes distortions in the texture space and textures must be first aligned before any statistical analysis. In turn, de-alignment must be also estimated to map the aligned textures back to their associated input textures (see Figure 2). And these backward warps must be part of the appearance model to enable appearance reconstruction. In the following, warps denote the backward warps. Also, we consider vector representations of the aligned texture maps and of the warps. These representations include only pixels that fall inside active regions within texture maps. We perform Principal Component Analysis on the textures and on the warp data separately to find the orthonormal bases that encode the main modes of variation in the texture space and in the warp space independently. We refer to vectors spanning the texture space as Eigen textures, and to vectors spanning the warp space as Eigen warps.

Let us consider first texture maps. Assume $N$ is the dimension of the vectorized representation of active texture elements, and $F$ the total number of frames available for the subject under consideration. To give orders of magnitude for our datasets, $N = 22438995$ and $F = 207$ for the TOMAS dataset, and $N = 25966476$ and $F = 290$ for the CATY dataset that will be presented in the next section. We start by computing the mean image $\bar{A}$ and the centered data matrix $M$ from aligned texture maps $\{A_i\}_{i\in[1..F]}$:

$$\bar{A} = \frac{1}{F}\sum_k A_k \quad , \quad M = \begin{bmatrix} | & & | \\ A_1 - \bar{A} & ... & A_F - \bar{A} \\ | & & | \end{bmatrix}. \tag{4}$$

Traditionally, the PCA basis for this data is formed by the Eigen vectors of the covariance matrix $MM^T$, of size $N \times N$, but finding such vectors can easily become prohibitive as a consequence of the texture dimensions. However, it appears that the non zero eigen values of $MM^T$ are equal to the non zero Eigen values of $M^TM$, of size $(F \times F)$ this time, and that they are at most: $\min(F, N) - 1$. Based on this observation, and since $F << N$ in our experiments, we solve the characteristic equation $\det(MM^T - \alpha I_N) = 0$ by performing Singular Value Decomposition on the matrix $M^TM$, as explained in [27]:

$$M^TM = D\Sigma D^T \quad , \quad D = \begin{bmatrix} | & & | \\ V_1 & ... & V_F \\ | & & | \end{bmatrix} \tag{5}$$

where $D$ contains the $(F-1)$ orthonormal Eigen vectors $\{V_i\}$ of $M^TM$, and $\Sigma = diag(\alpha_i)_{1\leq i\leq F}$ contains the eigen values $\{\alpha_i\}_{1\leq i\leq F-1}$. We can then write:

$$M^TMV_i = \alpha_i V_i \quad , \quad i \in [1..F-1], \tag{6}$$

and hence:

$$MM^T \underbrace{MV_i}_{T_i} = \alpha_i \underbrace{MV_i}_{T_i} \quad , \quad i \in [1..F-1], \tag{7}$$

where $T_i$ are the Eigen vectors of $MM^T$ and therefore form the orthonormal basis of the aligned texture space after normalization, namely the Eigen textures.

In a similar way, we obtain the mean warp $\bar{w}$ and the orthonormal basis of the warp space $\{W_i\}_{1\leq i\leq F-1}$, the Eigen warps.

### 3.3 Texture generation

Given the Eigen textures and the Eigen warps, and as shown in Figure 4, a texture can be generated by first creating an aligned texture by linearly combining Eigen textures and second de-aligning this new texture using another linear combination of the Eigen warps.

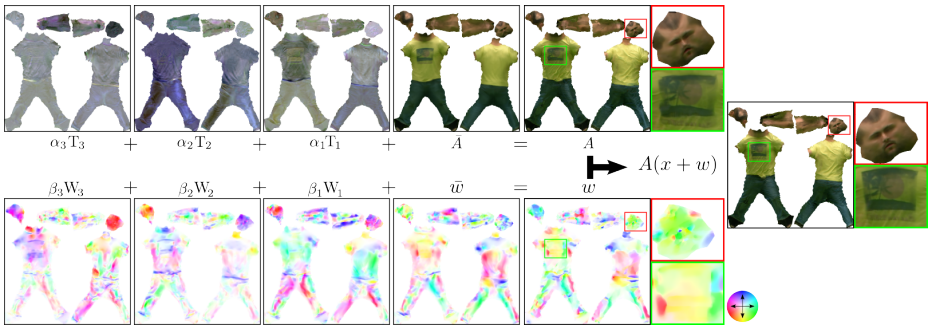



Fig. 4: Texture map generation by linear combination.

## 4 Performance Evaluation

To validate the estimation quality of our method, we apply our estimation pipeline to several datasets, project and warp input data using the built eigenspaces, then evaluate the reconstruction error. To distinguish the different error sources, we evaluate this error both in texture space before projection, and in image domain by projecting into the input views, as compared to the original views of the object and the texture before any reconstruction in texture space, estimated in our pipeline using [11]. For the image error measurement, we use the 3D model that was fitted to the sequence, as tracked to fit the test frames selected [21], and render the model as textured with our reconstructed appearance map, using a standard graphics pipeline. In both cases, we use the structural similarity index (SSIM) [28] as metric to compare to the original. All of our SSIM estimates are computed in the active regions of the texture and image domains, that is on the set of texels actually mapped to the 3D model in the texture domain, and only among actual silhouette pixels in the image domain.

We study in particular the compactness and generalization abilities of our method, by examining the error response as a function of the number of eigen components kept after constructing the linear subspaces, and the number of training images selected. For all these evaluations, we also provide the results for a naive PCA strategy, where only a set of eigen appearance maps are built in texture space and use to project and reconstruct textures, to show the performance contribution of including the Eigen warps.

For validation, we used two multi-sequence datasets: (1) the Tomas dataset which consists of 4 different sequences left, right, run and walk with 207 total number of frames and 68 input views each captured at resolution 2048x2048 pixels per frame; and (2) the Caty dataset: low, close, high and far jumping sequences with 290 total number of frames and 68 input views each captured at resolution 2048x2048 pixels per frame.

### 4.1 Estimation Quality and Compactness

We study the quality and compactness of the estimated representation by plotting the SSIM errors of reconstructed texture and image estimates of our method against naive PCA, for the two multi-sequence datasets (Figure 5). Note that all texture domain variability could be trivially represented by retaining as many Eigen textures as there are input images, thus we particularly examine how the quality degrades with the fraction of Eigen components kept. In the case of image domain evaluations, we plot the average SSIM among all viewpoints. Our method outperforms naive PCA in image and texture domains on both datasets, achieving higher quality with a lower number of Eigen components, and only marginally lower quality as the number of components grows, where the method would be anyway less useful. Higher number of Eigen components marginally favors naive PCA, because naive PCA converges to input textures when increasing the Eigen textures retained by construction, whereas our method hits a quality plateau due to small errors introduced by texture warp estimation and decomposition. For both datasets, virtually no error (0.98 SSIM) is introduced by our method in the texture domain with as low as 50 components, a substantially low fraction compared to the number of input frames (207 and 290). This illustrates the validity of the linear variability hypothesis in texture domain. The error is quite higher in the image domain (bounded by 0.7) for both our method and naive PCA, because measurements are then subject to fixed upstream errors due to geometric alignments, projections and image discretizations. Nevertheless, visually indistinguishable results are achieved with 50 Eigen components (images and warps), with a significant compactness gain.

### 4.2 Generalization ability

In the previous paragraph, we examined the performance of the method by constructing an Eigen space with all input frames. We here evaluate the ability of the model to generalize, *i.e.* how well the method reconstructs textures from input frames under a reduced number of examples that don't span the whole input set. For this purpose, we perform an experiment using a varying size training set, and a test set from frames not in the training set. We use a training set comprised of randomly selected frames spanning 0% to 60% of the total number of frames, among all sequences and frames of all datasets, and plot the error of projecting the complement frames on the corresponding Eigen space (Figure 6). The experiment shows that our representation produces a better generalization than naive PCA, *i.e.* less training frames need to be used to reconstruct a texture and reprojections of equivalent quality. For the Tomas dataset, one can observe than less than half training images are needed to achieve similar performance in texture space, and a quarter less with the Caty dataset.

## 5 Applications

We investigate below two applications of the appearance representation we propose. First, the interpolation between frames at different time instants and sec-

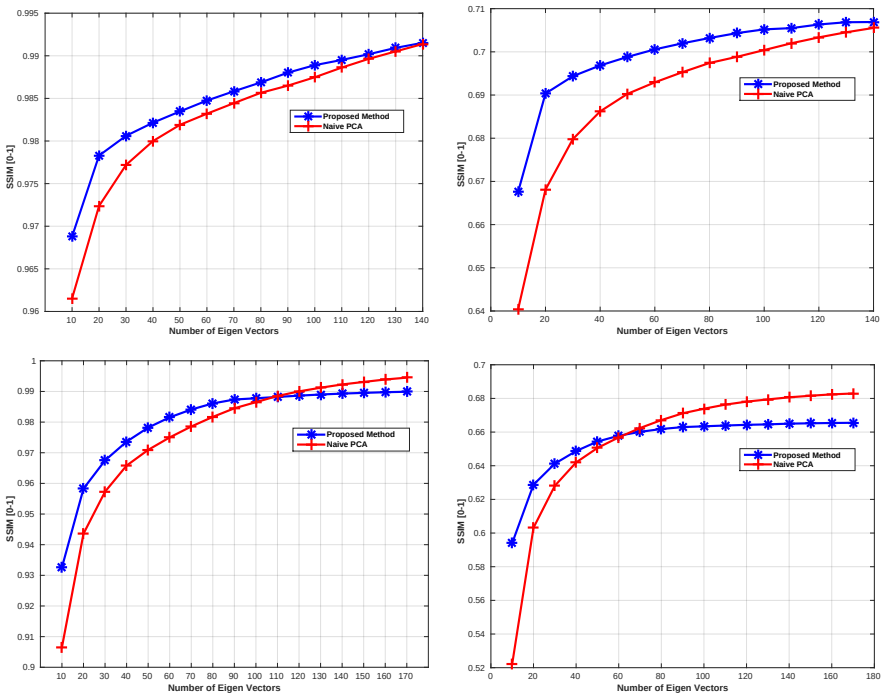



Fig. 5: Reconstruction Error for Tomas and Caty Dataset from top to down in Texture and Image Domain from left to right.

ond, the completion of appearance maps at frames where some appearance information is lacking due to occlusions or missing observations during the acquisition. Results are shown in the following section and the supplementary video.

### 5.1 Interpolation

In our framework, appearance interpolation benefits from the pre-computed warps and the low dimensionality of our representation to efficiently synthesize compelling new appearances with reduced ghosting-artefacts. It also easily enables extension of appearance interpolation from pairwise to multiple frames. Assume that shapes between two given frames are interpolated using a standard non-linear shape interpolation, for instance [29]. Consider then the associated aligned textures and associated warps at the given frames. We perform a linear interpolation in the Eigen texture and Eigen warp spaces respectively by blending the projection coefficients of the input appearance maps. Poisson warping, as introduced in section 3.1 is used to build de-aligned interpolated texture with the interpolated backward warp. Figure 7 compares interpolation using our pipeline to a standard linear interpolation for 4 examples with the Caty and Tomas datasets. Note that our method is also linear but benefits from the alignment

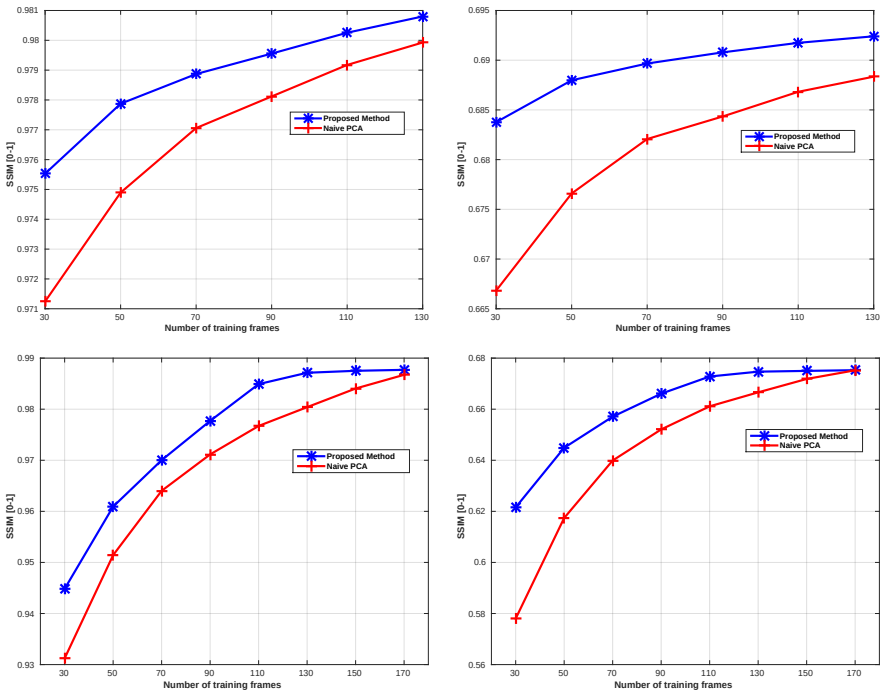



Fig. 6: Generalization Error for THOMAS and CATY Dataset from top to down in Texture and Image Domain from left to right.

performed in the texture space to reduce interpolation artefacts, as well as from the simplified computational aspects since interpolation applies to projection coefficients only.

### 5.2 Completion

As mentioned earlier, appearance maps can be incomplete due to acquisition issues. For instance, as shown in Figure 8, during the running sequence the actor TOMAS bends his knees in such a way that the upper parts of his left and right shins become momentarily hidden to the acquisition system. This results in missing information for those body parts in the texture maps and over a few frames. Such an issue can be solved with our texture representation by omitting the incomplete frames when building our appearance representations, and then projecting these incomplete appearance maps in the Eigen spaces and reconstructing them using the projection coefficients and Poisson texture warping. Figure 8 shows two examples of this principle with occluded regions. Note however, that while effectively filling gaps in the appearance map, this completion might yet loose appearance details in regions of the incomplete map where information is not duplicated in the training set.

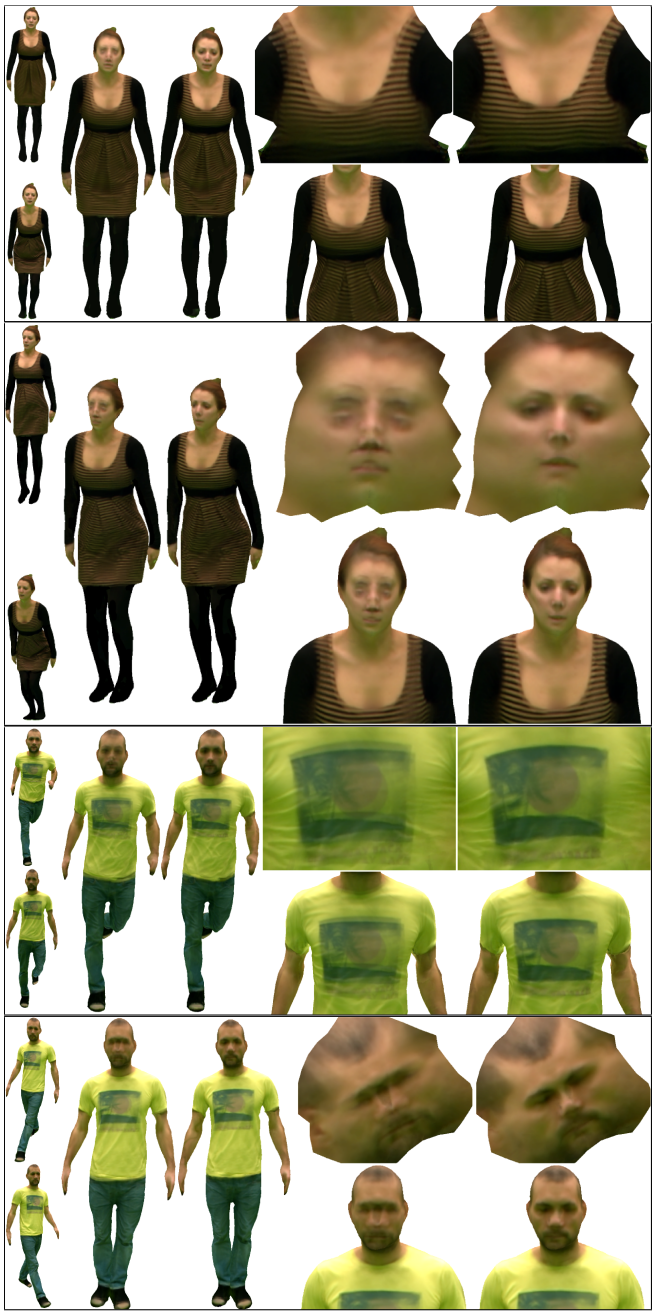


Fig. 7: Interpolation examples using linear interpolation (left) and our pipeline (right). From left to right: Input frames, Interpolated models, and a close-up on the texture maps (top) and the rendered images (bottom).

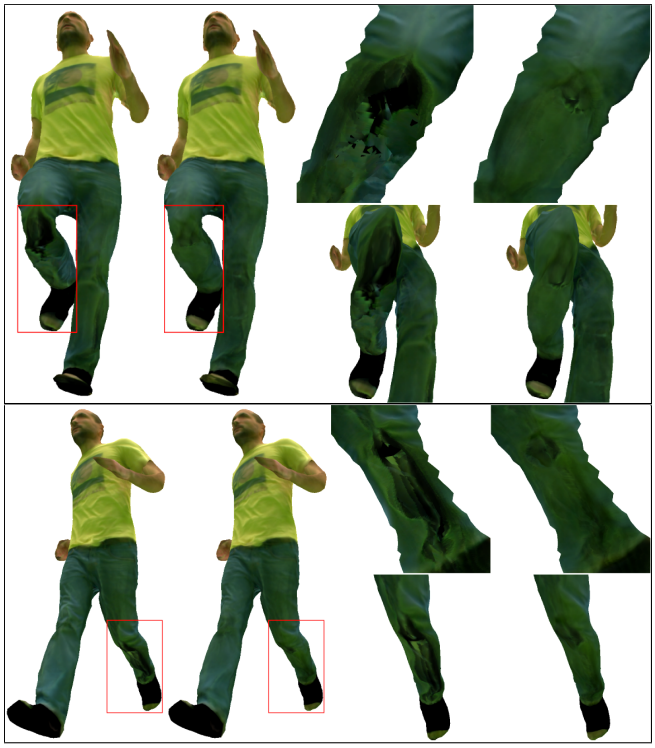


Fig. 8: Completion examples. From left to right: Input and completed models, close-up on input and completed texture maps (top) and rendered images (bottom).

## 6 Conclusion

We have presented a novel framework to efficiently represent the appearance of a subject observed from multiple viewpoints and in different motions. We propose a straightforward representation which builds on PCA and decomposes into Eigen textures and Eigen warps that encode, respectively, the appearance variations due to viewpoint and illumination changes and due to geometric modeling imprecisions. The framework was evaluated on 2 datasets and with respect to: (i) its ability to accurately reproduce appearances with compact representations; (ii) its ability to resolve appearance interpolation and completion tasks. In both cases, the interest of a global appearance model for a given subject was demonstrated. Among the limitations, the representation performances are dependent on the underlying geometries. Future strategies that combine both shape and appearance information would thus be of particular interest. The proposed model could also be extended to global representations over populations of subjects.

## References


1. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Trans. Graph. (2015)
2. Nishino, K., Sato, Y., Ikeuchi, K.: Eigen-texture method: Appearance compression and synthesis based on a 3d model. IEEE Trans. Pattern Anal. Mach. Intell. (2001)
3. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In: ACM SIGGRAPH. (1996)
4. Lempitsky, V.S., Ivanov, D.V.: Seamless mosaicing of image-based texture maps. In: CVPR. (2007)
5. Waechter, M., Moehrle, N., Goesele, M.: Let there be color! large-scale texturing of 3d reconstructions. In: ECCV. (2014)
6. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: ACM SIGGRAPH. (1999)
7. Carranza, J., Theobalt, C., Magnor, M.A., Seidel, H.P.: Free-viewpoint video of human actors. ACM Trans. Graph. (2003)
8. Zitnick, C., Kang, S., Uyttendaele, M., Winder, S., Szeliski, R.: High-quality video view interpolation using a layered representation. In: ACM SIGGRAPH. (2004)
9. Eisemann, M., De Decker, B., Magnor, M., Bekaert, P., de Aguiar, E., Ahmed, N., Theobalt, C., Sellent, A.: Floating textures. Computer Graphics Forum (Proc. of Eurographics) (2008)
10. Tung, T.: Simultaneous super-resolution and 3D video using graph-cuts. (2008)
11. Tsiminaki, V., Franco, J.S., Boyer, E.: High Resolution 3D Shape Texture from Multiple Videos. In: CVPR. (2014)
12. Goldlücke, B., Aubry, M., Kolev, K., Cremers, D.: A super-resolution framework for high-accuracy multiview reconstruction. International Journal of Computer Vision (2014)
13. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. ACM Trans. Graph. (2008)
14. Cagniart, C., Boyer, E., Ilic, S.: Free-from mesh tracking: a patch-based approach. In: CVPR. (2010)
15. Volino, M., Casas, D., Collomosse, J., Hilton, A.: Optimal representation of multiple view video. In: BMVC. (2014)
16. Boukhayma, A., Boyer, E.: Video based Animation Synthesis with the Essential Graph. In: 3DV. (2015)
17. Casas, D., Tejera, M., Guillemaut, J.Y., Hilton, A.: 4D Parametric Motion Graphs for Interactive Animation. In: ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games. (2012)
18. Casas, D., Volino, M., Collomosse, J., Hilton, A.: 4D Video Textures for Interactive Character Appearance. Computer Graphics Forum (Proc. of Eurographics) (2014)
19. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognitive Neuroscience (1991)
20. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. (2001)
21. Allain, B., Franco, J.S., Boyer, E.: An efficient volumetric framework for shape tracking. In: CVPR. (2015)
22. Snchez Prez, J., Meinhardt-Llopis, E., Facciolo, G.: TV-L1 Optical Flow Estimation. Image Processing On Line (2013)

23. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Trans. Graph. (2003)
24. Chen, T., Zhu, J.Y., Shamir, A., Hu, S.M.: Motion-aware gradient domain video composition. IEEE Transactions on Image Processing (2013)
25. Linz, C., Lipski, C., Magnor, M.: Multi-image interpolation based on graph-cuts and symmetric optical flow (2010)
26. Mahajan, D., Huang, F.C., Matusik, W., Ramamoorthi, R., Belhumeur, P.N.: Moving gradients: a path-based method for plausible image interpolation. ACM Trans. Graph. (2009)
27. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognitive Neuroscience (1991)
28. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing (2004)
29. Xu, D., Zhang, H., Wang, Q., Bao, H.: Poisson shape interpolation. In: ACM Symposium on Solid and Physical Modeling. (2005)

$\therefore$

## A.6 TRACKING-BY-DETECTION OF 3D HUMAN SHAPES : FROM SURFACES TO VOLUMES

Chun-Hao Huang, Benjamin Allain, Jean-Sébastien Franco, Nassir Navab, Slobodan Ilic, Edmond Boyer. IEEE Transactions on Pattern Analysis and Machine Intelligence, Institute of Electrical and Electronics Engineers, 2017,

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



# Tracking-by-Detection of 3D Human Shapes: from Surfaces to Volumes

Chun-Hao Huang, Benjamin Allain, Edmond Boyer, Jean-Sébastien Franco, Federico Tombari, Nassir Navab and Slobodan Ilic

**Abstract**—3D Human shape tracking consists in fitting a template model to temporal sequences of visual observations. It usually comprises an association step, that finds correspondences between the model and the input data, and a deformation step, that fits the model to the observations given correspondences. Most current approaches follow the Iterative-Closest-Point (ICP) paradigm, where the association step is carried out by searching for the nearest neighbors. It fails when large deformations occur and errors in the association tend to propagate over time. In this paper, we propose a discriminative alternative for the association, that leverages random forests to infer correspondences in one shot. Regardless the choice of shape parameterizations, being surface or volumetric meshes, we convert 3D shapes to volumetric distance fields and thereby design features to train the forest. We investigate two ways to draw volumetric samples: voxels of regular grids and cells from Centroidal Voronoi Tessellation (CVT). While the former consumes considerable memory and in turn limits us to learn only subject-specific correspondences, the latter yields much less memory footprint by compactly tessellating the interior space of a shape with optimal discretization. This facilitates the use of larger cross-subject training databases, generalizes to different human subjects and hence results in less overfitting and better detection. The discriminative correspondences are successfully integrated to both surface and volumetric deformation frameworks that recover human shape poses, which we refer to as 'tracking-by-detection of 3D human shapes.' It allows for large deformations and prevents tracking errors from being accumulated. When combined with ICP for refinement, it proves to yield better accuracy in registration and more stability when tracking over time. Evaluations on existing datasets demonstrate the benefits with respect to the state-of-the-art.

**Index Terms**—Shape tracking, random forest, centroidal Voronoi tessellation, 3D tracking-by-detection, discriminative associations.

## 1 Introduction

3D shape tracking is the process of recovering temporal evolutions of a template shape using visual information, such as images or 3D points. It finds applications in several domains including computer vision, graphics and medical imaging. In particular, it has recently demonstrated a good success in marker-less human motion capture (mocap). Numerous approaches assume a user-specific reference surface, with the objective to recover the skeletal poses [1], surface shapes [2], or both simultaneously [3]. A standard tracking process consists in an alternation of the following two steps. First, finding associations between the observed data, *e.g.* 3D points of the reconstructed visual hull, to the corresponding 3D template shape, typically based on the proximity in Euclidean space or a feature space. Second, given such associations, recovering the pose of the template under the constraint of a deformation model, typically based on the kinematic skeleton [1], [4], [5], [6], or the piecewise-rigid surface [2] parameterization, among others.

Most of these model-based methods can be viewed as extensions of Iterative-Closest-Point (ICP) framework [7], [8] to deformable shapes, which attempts to explain newly observed data using the previous outcomes. As long as the initialization is close to the optimum solution, it is able to produce outstanding results. However, they also suffer from inherent weaknesses of *generative* strategies, *e.g.* slow convergence. Moreover, when large deformations or many outliers occur, discovering associations becomes particularly difficult. Unreliable correspondences result in ambiguous situations that yield erroneous numerical solutions.

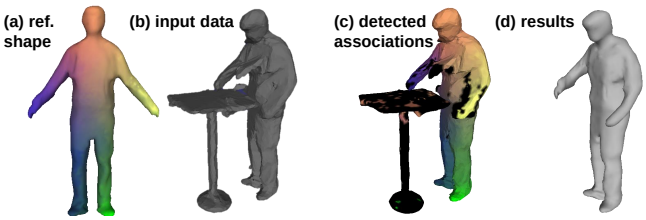



Fig. 1. Given a reference shape (a) and input data (b), our method discovers reliable data-model correspondences by random forests, color-coded in (c). This strategy detects user-specific shapes in a frame-wise manner, resulting in better sustainability. In (d) the reference model (a) is deformed with correspondences (c) to fit the input data (b).

Recently, a number of alternatives and enhancements have been explored for both association and deformation stages independently. On one hand, improvements have also been proposed for the association problem by discovering them discriminatively [6], [9], [16], This in turn yields the possibility for 3D tracking techniques that are robust to failure. In contrast to those generative ICP variants, these *discriminative* approaches that 'detect' rather than track models have shown better robustness over the past decade, for instance, in human pose estimation with 2.5D data from Kinect [6], [10]. These approaches usually consider foreground human subjects to be pre-segmented, which is not a

- *C.-H. Huang, F. Tombari and N. Navab are are with Computer Aided Medical Procedures, Technische Universität München, Germany, E-mail: see http://campar.in.tum.de/WebHome*
- *E. Boyer, J.-S. Franco and B. Allain are with Inria, LJK, France E-mail: see http://morpheo.inrialpes.fr/*
- *S. Ilic is with Siemens AG, Munich, Germany E-mail: see http://campar.in.tum.de/WebHome*

*Manuscript received XX.XX, 2017; revised XX.XX, 2017.*

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



favorable assumption in full 3D data that generally contains substantial amount of outliers like Fig. 1(b). Including non-human objects into the reference shape so that more points are explained, *i.e.* less outliers, is one workaround adopted by many existing multi-view methods [17], [18], with the downside that further post-processing is required to analyze only humans' movements. There is a growing need to facilitate robust frame-wise observation-model associations for reconstructed complete 3D human shapes. Although surface-based features are commonly used for this purpose in the context of shape matching [9], volumetric features have also proven to be a promising direction for 3D shape description with surface-based templates [11].

On the other hand, progress has also been made in the deformation stage by introducing volumetric deformation models instead of purely surface-based ones, mainly motivated by the observation that human movements are largely volume-preserving. It has shown significantly improved robustness to various tracking situations, such as shape folding and volume bias of observed shapes [12]. As volumetric deformation models are gradually used in capturing actors' motions due to their inherent local volume-preserving properties, facilitating volumetric discriminative correspondences can be favorable. We investigate this direction and make the following two contributions in this paper.

First, two volumetric features are designed for *human shape correspondence detection*, operating respectively on surface and volumetric meshes. Inspired by Taylor *et al.* [6], we apply regression forests to improve the associations, with two learning strategies devised for different shape parameterizations. In the case of surface mesh representations, we convert shapes to the volumetric Truncated Signed Distance Field (TSDF) [13], where each surface vertex is fed into user-specific forests to predict correspondences in *one shot*. Meanwhile, we also tessellate both the observed and template shapes as a set of uniform and anisotropic cells (see Fig. 2) from Centroidal Voronoi Tessellation (CVT) [14] and, again leverage the similar distance-transform representations to predict volumetric correspondences for all CVT cells.

Second, by integrating these one-shot associations into the respective deformation models, we further present a discriminative human mocap framework, as depicted in Fig. 1, termed tracking-by-detection of 3D human shapes. In contrast to the ICP-like methods [2], [3], [4], it does not require close initializations from a nearby frame to estimate correspondences and thus better handles large deformations. Experiments demonstrate that, when combined with a generative tracking approach, this *hybrid* framework leads to better or comparable results than purely generative ones, *e.g.* [2], [15], reducing error accumulations and hence increasing the stability. The regression entropy is also augmented with the classification one to identify outliers. Very few prior arts afford the tracking or matching situation where the input describes mainly irrelevant outliers. Notably, in the case of CVT, our method is a unified volumetric pipeline where the shape representation, deformation model, feature description, and points association are all built on a single CVT representation that brings benefits at all stages of the pipeline. This fully volumetric tracking-by-detection method shows improved accuracy and memory performance compared to the surface-based counterpart [11].

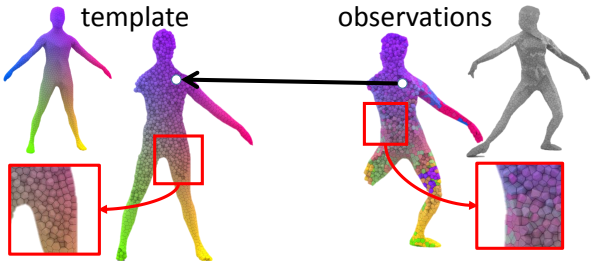



Fig. 2. Centroidal Voronoi tessellations yields volumetric cells of uniform shape and connectivity with controllable complexity. The cells of the observed shape are matched discriminatively to those of the template.

## 2 Related Work

Among the vast literature on human motion analysis [19], we focus on top-down approaches that assume a 3D template and deform it according to input data, either directly with pixels [4], [20], or with computed 3D points [2], [3], [15]. These methods typically decompose into two major steps: (1) data association, where observations are associated to the model, and (2) deformation stage, where motion parameters are estimated given the associations. As our primary objective in this paper is to improve the first part, existing approaches are discussed accordingly below.

### 2.1 Generative approaches

Methods of this category follow the association strategy in ICP while extending the motion model to more general deformations than the one in the original method [7], [8]. Correspondences are addressed by searching for closest points, with various distance measures such as point-to-point [2], point-to-plane [21], or Mahalanobis distances [20]. This strategy heavily relies on the fact that observations in consecutive frames are in vicinity. Klaudiny *et al.* [22], Huang *et al.* [3] and Collet *et al.* [17] generalize the idea from the previous frame to a certain key-frame in the considered sequences, finding the best non-sequential order to track, but the proximity assumption remains. On the other hand, since 3D data such as reconstructed point clouds often contain spurious fake geometries, another challenge consists in identifying online and dynamically irrelevant observations without any prior knowledge. Liu *et al.* [4] establish 3D-2D correspondences by considering both texture in images and contours in silhouettes and further include image segmentation information to differentiate multiple interacting subjects. Huang *et al.* [2], [3] relax the hard correspondence constraint to soft assignments and introduce an additional outlier class to reject noisy observations. Data is explained by Gaussian Mixture Models (GMM) in an Expectation-Maximization (EM) manner [23]. In [24], both source and target points are similarly modeled as GMMs and the registration problem is cast as minimizing the distance between two mixture models. Collet *et al.* [17] fuse information from various modalities attentively to generate high-quality textured meshes. Yet, to yield a temporal coherent mesh tessellation, the underlying tracking component is still ICP-based [25]. All these generative methods are highly likely to fail in large deformations. Furthermore, they are prone to error accumulations and, as a result of matching several successive frames wrongly (whether sequentially or not), they are prone to drift.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



### 2.2 Discriminative approaches and 3D descriptors

Recently, discriminative approaches have demonstrated their strengths in estimating human [6], [26] and hand [27] poses from depth images. With the initial intention to substitute ICP-based optimization, Taylor *et al.* [6] propose a frame-wise strategy that yields decent dense correspondences without iterative refinements. The method replaces the step of proximity search in ICP-based tracking methods by learning the mapping from input 3D points from depth sensors, to the human template surface domain, termed the Vitruvian manifold. Later, Pons-Moll *et al.* [5] train forests with a new objective on surface manifolds, and increase the precision by finishing convergence with an ICP-based loop after the discriminative association stage. Both approaches operate frame-independently and are generally drift free. Following the same weak pair-wise features and random forest framework, Dou *et al.* [18] learn to match two successive depth frames to avoid depending on a specific template.

More informative descriptors and matching strategies have long been studied for shape recognition or retrieval with meshes [28] and point clouds [29]. The well known heat kernel signatures (HKS) [30] and wave kernel signatures (WKS) [31] exploit the Laplacian-Beltrami operator, the extension of the Laplacian operator to surface embeddings. Rodola *et al.* [9] later apply forests to learn the parameters of WKS during training. These features are nonetheless known for their lack of resilience to significant topology changes, an artifact frequently seen in noisy surface acquisitions. MeshHoG [32] and SHOT [33] attach a local coordinate frame at each point to achieve invariant representations and reach better performance for noisy surfaces. To enforce consistent matches over the whole shape, Chen and Koltun [34] and Starck *et al.* [35] formulate the matching problem as the inference of Markov random field (MRF).

Besides hand-crafted features, there is a recent trend that applies Convolutional Neural Network (CNN) [36] to discover the deep representation of non-rigid human shapes. Wei *et al.* [16] render depth images in several viewpoints, where the CNN feature transformation takes place, and average the descriptors from multiple views. Boscaini *et al.* [37] stay in 3D space but define the convolution function in the intrinsic manifold domain. While showing encouraging results in handling missing data, these methods do not consider matching human shapes in the presence of large amount of outliers, *e.g.* un-subtracted furniture in the background, and thus do not fit to our 'detection' purpose.

Another common trait of the aforementioned approaches is that the computation involves only surface points. We show in our early work [11] that surface features can be built based on local coordinate frames in a regular-grid volume. In this paper, we not only improve this feature but also propose a new one to address the need of fully volumetric correspondences. Both features, implicitly or explicitly, leverage distance-transform volumes to describe 3D geometry. Taking only surface vertices into account, the existing approaches rely on heterogeneous shape representations, deformation models, target primitives and feature spaces. Instead, our CVT-based tracking-by-detection proposal builds a unified framework for all these purposes and takes advantage of volumetric tracking strategies.

## 3 Overview

We implement discriminative associations using two different volumetric representations. In the first case, we convert the triangular surface meshes to the Truncated Signed Distance Field (TSDF) constructed with the regular 3D volumetric grid. In the second case, we use CVT representation which is not bound to the regular grids. As in Fig. 2, the interior space of a triangular surface is tessellated into a set of *cells* of uniform anisotropic shape whose seed location coincides with its centers of mass. Such an optimal discretization yields lower memory footprint than regular-grid volumes, in turn accommodating more training meshes. Moreover, we also associate CVT cells discriminatively and present volumetric correspondences.

Formally, a humanoid shape describes a continuous volumetric domain in 3D $\Omega \subset \mathbb{R}^3$ whose border $\partial\Omega$ defines a 2-manifold surface. The discretized mesh representation $\mathcal{M}$ contains a set of 3D points $\mathbf{M}$ and their connectivity $\mathcal{T}$, *i.e.* $\mathcal{M} = (\mathbf{M}, \mathcal{T})$, where $\mathbf{M}$ is drawn from the surface ($\mathbf{M} \subset \partial\Omega$) or the whole volume ($\mathbf{M} \subset \Omega$ ). The goal of 3D shape tracking is to register a source reference[1] mesh $\mathcal{X} = (\mathbf{X}, \mathcal{T}_X)$ to the observed target mesh $\mathcal{Y} = (\mathbf{Y}, \mathcal{T}_Y)$, such as fitting the shape in Fig. 1(a) to the one in Fig. 1(b).

Our method starts with surface meshes reconstructed by shape-from-silhouette method [38]. We refer only to points on surfaces as *vertices* $v \in \mathcal{V}$, where $\mathcal{V}$ is the set of their indices. Suppose the reference surface $\mathcal{X}$ and the input visual hull $\mathcal{Y}$ are located at $\mathbf{X} = \{\mathbf{x}_v\}_{v \in \mathcal{V}_X}$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i \in \mathcal{V}_Y}$[2] respectively, the registration typically boils down to two steps: (1) *association*: matching each points in $\mathcal{Y}$ with those in $\mathcal{X}$ to build the correspondence set $\mathcal{C} = \{(i, v)\} \subset \mathcal{V}_Y \times \mathcal{V}_X$; and (2) *optimization*: estimating the motion parameter $\mathbf{\Theta}$ by minimizing an energy $E$ that describes the discrepancies between pairs in $\mathcal{C}$, *i.e.* $\hat{\mathbf{\Theta}} = \operatorname{argmin}_{\mathbf{\Theta}} E(\mathbf{\Theta}; \mathcal{C})$, such that $\mathbf{X}(\hat{\mathbf{\Theta}})$ resembles $\mathbf{Y}$ as much as possible.

To discover the correspondences $\mathcal{C}$ discriminatively, we adapt the *Vitruvian* strategy [6] from matching 2.5D against 3D to 3D against 3D. This amounts to *warping* the input mesh $\mathcal{Y}$ to the reference one $\mathcal{X}$, denoted as $\tilde{\mathcal{Y}} = (\tilde{\mathbf{Y}}, \mathcal{T}_Y) = (\mathbf{r}(\mathbf{Y}), \mathcal{T}_Y)$ where $\mathbf{r}$ is the warping function. A good $\mathbf{r}$ shall lead to a clean warp $\tilde{\mathcal{Y}}$ as in Fig. 3. Incorrect warped points, however, can still be told from huge edges. Specifically, this $\mathbb{R}^3 \rightarrow \mathbb{R}^3$ mapping $\mathbf{r}$ is learned by a regression forest [39]. We convert each surface into an implicit representation, a distance field, which is usually defined volumetrically. As stated above, we investigate two ways to define the volumetric elements $s$. The first one is a voxel from a regular axis-aligned volume, *i.e.* $s \in \mathbb{N}^3$, while the second one is a cell from a volumetric mesh, *i.e.* $s \in \mathcal{S}$, where $\mathcal{S}$ is a group of CVT cells that tessellate only the surface interiors. Depending on the choice of $s$, our volumetric feature $\mathbf{f}$ is hence also realized in two different forms. Taking the feature $\mathbf{f}$ as input, multiple binary decision trees are trained with previously observed meshes. In the online testing phase, a input point obtains a prediction $\tilde{\mathbf{y}}_i = \mathbf{r}(\mathbf{y}_i)$ that indicates the locations of potential matches since the warp $\tilde{\mathbf{Y}}$ is

1. Several terms are used interchangeably in this paper: reference and template; correspondences and associations; point and primitive.

2. The observations are always indexed by $i$ regardless of the parameterization.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



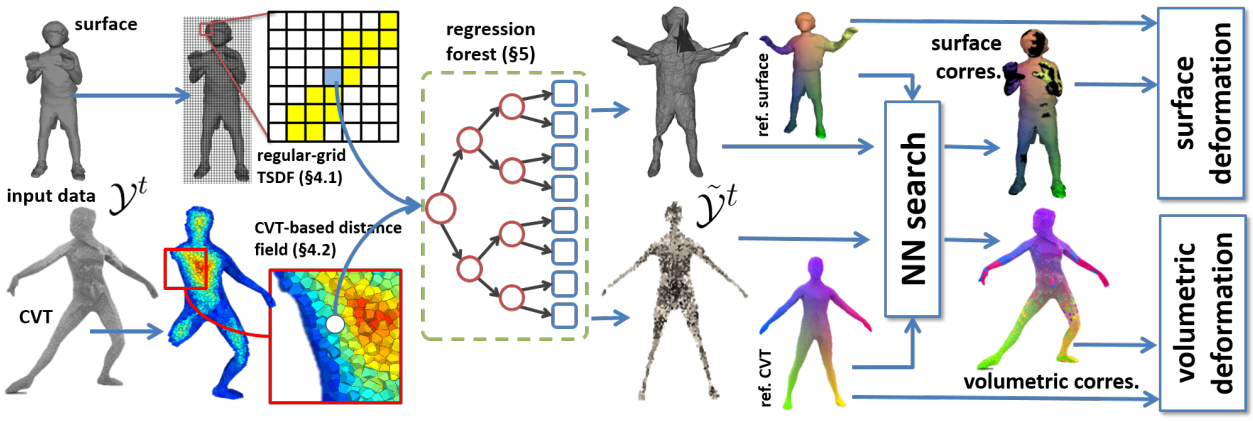



Fig. 3. The pipeline of our tracking-by-detection framework. Data-model associations are visualized in the same color. Upper row: surface-based associations (black means no correspondence found for that vertex); bottom row: volumetric associations.

learned to resemble $\mathbf{X}$. Thus, $\mathcal{C}$ can be built swiftly by doing nearest neighbor search between $\tilde{\mathbf{Y}}$ and $\mathbf{X}$ just once and the deformation parameter $\mathbf{\Theta}$ that encodes the *shape pose* of the template is estimated accordingly. Notably, in the case of CVT, since cells comprise a volumetric mesh, the whole pipeline (discovering $\mathcal{C}$ and estimating $\mathbf{\Theta}$) can instead be conducted in a fully volumetric fashion. Fig. 3 illustrates this *correspondence detection* process. The details of training, prediction and deformation models are provided in § 5.

## 4 Volumetric Features

The two volumetric features are introduced in this section. Although both taking a volumetric point $s$ as input, the first one actually aims to match surface vertices $v$, denoted as $\mathbf{f}(v) := \mathbf{f}(s_v)$ while the second one matches $s$ directly, *i.e.* $\mathbf{f}(s)$. Both are designed to be incorporated into forest training and prediction. A great advantage of decision trees is to learn the most discerning attributes among a large feature bank. One does not have to prepare the whole high-dimensional vector $\mathbf{f}$ to draw predictions, because only a few learned attributes $\kappa$ are needed to traverse the trees. As a result, features can be computed *on the fly* during testing. To make use of such property, the calculation of each $f_\kappa$ is assumed to independent. We hence avoid the histogram-based descriptors that requires normalization, such as MeshHOG [32] or SHOT [33], and resort to offset comparison features used in [40] for $\mathbf{f}(s_v)$ and Haar feature in [41] for $\mathbf{f}(s)$.

### 4.1 Regular-voxel-based features

Our first approach to discriminative associations considers regular-grid volumes (upper row in Fig. 3, $s \in \mathbb{N}^3$). The warping function $\mathbf{r}$ is modeled as a composite one: $\mathbf{r} : \mathbb{R}^3 \to \mathbb{N}^3 \to \mathbb{R}^3$, where the former is voxelization and the regression trees account for only the latter. We first cast each mesh $\mathcal{M}$ into a volumetric scalar field $D : \mathbb{N}^3 \subset \mathbb{R}^3 \to \mathbb{R}$.

#### 4.1.1 Truncated signed distance transform (TSDT)

Voxelizing a surface in general comprises two parts: (1) determining which voxel $s$ that every vertex $v$ maps to, and (2) testing the overlap between triangles and voxels. The first part can be viewed as a quantization mapping from Euclidean space to a discretized space $s : \mathbb{R}^3 \to \mathbb{N}^3$. The size of the volume is large enough to include all possible pose variations, and its center is aligned with the barycenter of the surfaces. The voxel size is chosen to be close to the average edge length of meshes, so that a single voxel is not mapped by too many vertices. To check the intersection of triangles with voxels, we apply *separating axis theorem* which is known to be efficient for collision detection [42].

Voxels occupied by the surface are referred to as $s_{\text{suf}}$. We further identify voxels located inside and outside the surface, denoted respectively as $s_{\text{in}}$ and $s_{\text{out}}$. Together they define a directional truncated signed distance transform:

$$D(s) = \begin{cases} +\epsilon & \text{if } s_{\text{out}} \text{ and } d(s,\mathcal{M}) > \epsilon. \\ +d(s,\mathcal{M}) & \text{if } s_{\text{out}} \text{ and } d(s,\mathcal{M}) \leq \epsilon. \\ 0 & \text{if } s_{\text{suf}} \\ -d(s,\mathcal{M}) & \text{if } s_{\text{in}} \text{ and } d(s,\mathcal{M}) \leq \epsilon. \\ -\epsilon & \text{if } s_{\text{in}} \text{ and } d(s,\mathcal{M}) > \epsilon. \end{cases} \quad (1)$$

$d(s,\mathcal{M})$ denotes the shortest Euclidean distance from the voxel center to the mesh, which can be computed efficiently via AABB trees using `CGAL` library. If the distance is bigger than a threshold $\epsilon$, we store only $\pm\epsilon$ to indicate the inside/outside information. It is empirically set to be three times the physical length of diagonal of voxels. In the earlier version of this work [11], we store averaged surface normals at each $s_{\text{suf}}$. However, such representations yield high memory footprint and thus limit the amount of training meshes we can incorporate later in § 5. The TSDT representation naturally encodes the spatial occupancies of a mesh and the required memory footprint is only one-third of the former (each voxel stores now just a scalar, not a vector). It shares a similar spirit with implicit surface representations, *e.g.* level-set, and has been widely employed in RGBD-based tracking or reconstruction [43], [44].

#### 4.1.2 Pair-wise offset features

Next, we present the features $\mathbf{f}$ for describing TSDT, which are later used to train the forests. Since we are interested in predicting correspondences for vertices instead of triangles, from now on we concentrate only on those surface voxels $s_{\text{suf}}$ occupied by mesh vertices $v$, denoted as $s_v$. The feature is thus defined as a function of $s_v$, *i.e.* $\mathbf{f}(v) := \mathbf{f}(s_v)$.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



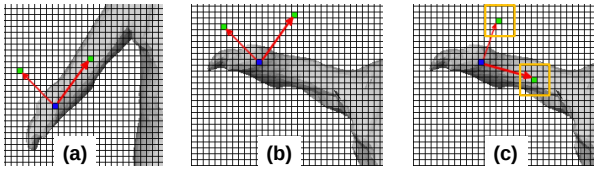



Fig. 4. The intuition of adjusting offsets. (a) original offset pair $\psi$. (b) $\eta = 0$ results in $\psi$ without re-orientation, *i.e.* $\mathbf{R} = \mathbf{I}$. (c) $\eta = 1$. $\psi$ is orientated by a rotation matrix $\mathbf{R} = [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]$ characterized by a LCF.

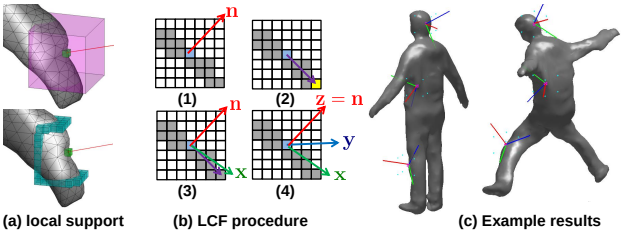



Fig. 5. Our method leads to quasi pose-covariant LCFs.

As depicted in Fig. 4, for each surface voxel $s_v$ (blue), we shoot two offsets (red vectors) $\psi = (\mathbf{o}_1, \mathbf{o}_2) \in \mathbb{N}^3 \times \mathbb{N}^3$, reaching two neighboring voxels (green). To describe the local geometry, we take the TSDT values within a cuboid around two respective voxels (yellow squares), perform element-wise subtractions and sum them up. Let $\varepsilon$ denotes this sum-of-difference operation. By definition, $\varepsilon$ from different offsets $\psi$ can be evaluated independently and thus fully parallelizable, which is an useful trait since this computation will be carried out multiple times during training with thousands of randomly generated $\psi$ for the same $s_v$.

The feature vector $\mathbf{f}$ consist of $\varepsilon$ resulted from many offset pairs $\psi$. More precisely, it is a function of $s_v$ but takes an offset pair $\psi$, a binary variable $\eta$ (whether to use *Local Coordinate Frame* (LCF) or not), and a rotational matrix $\mathbf{R} \in SO(3)$ (the orientation of LCF) as parameters. Every possible combination of offset pairs $\psi$ and binary variables $\eta$ results in one independent feature attribute $\kappa$, in notations: $f_\kappa(s_v) = \varepsilon(s_v; \mathbf{R}^\eta(\psi))$. The dimensionality of $\mathbf{f}$ is virtually infinite. Binary variables $\eta$ determines the alignment of the offset $\psi$ with respect to a LCF, whose transformation is specified by $\mathbf{R}$. The intuition behind this adjustment is to make features $\mathbf{f}$ invariant to poses, c.f. Fig. 4(b) and (c). Without re-orientations, $\psi$ might land on different types of voxel pairs, c.f. Fig. 4(a) and (b), and hence cause different feature responses $\varepsilon$, despite the fact that the current voxels are located on the same position on the body. Both offset pairs $\psi$ and binary variables $\eta$ are learned during forest training, while the rotational matrix $\mathbf{R}$ is characterized by a LCF obtained as follows.

#### 4.1.3 Local coordinate frame

Defining local coordinate frames for 3D primitives (voxels, vertices, points) has long been studied and usually comes with their 3D descriptor counterparts, see [45] for a comprehensive review. An ideal LCF is supposed to follow whatever transformations the meshes undergo, namely, as *co-variant* as possible, such that the consequent feature representations are as *invariant* as possible. Constructing a LCF boils down to defining three orthonormal vectors as $[x, y, z]$ axes. To do that, the state-of-the-art methods in the field of LCFs for rigid matching of 3D meshes and point clouds mainly rely on the neighboring points within a *local support* [33], [46], [47], [48]. The way they leverage spatial distributions can in general be classified into two categories: (1) EigenValue-Decomposition (EVD) [33], [47], [49], and (2) signed distance (SignDist.) [46], [48]. Since it is impractical to repeat EVD process for all surface voxels $s_v$, in the following, we propose an adaptation of SignDist. approach to our volumetric representations [50]. This conclusion is drawn after an extensive study and comparison of three LCF approaches presented in our early work [50].

Specifically, for each $s_v$, we consider its surface normals $\mathbf{n}_v$ as $z$ axis, and obtain $y$ axis by $z \times x$. The task left is to identify a repeatable $x$ axis. To this end, the class of SignDist. approaches look for a discerning point within the support (yellow voxel in Fig. 5(b)). We first open an local cuboid support (pink) around each $s_v$ (green) as visualized in Fig. 5(a). The search involves only the peripheral voxels $\tilde{s}$ (cyan) lying on the intersection of support borders and the surface. The discernibility is defined as the maximum signed distance to the tangent plane [46]:

$$\hat{s} = \arg\max_{\tilde{s} \in \tilde{\mathcal{S}}} \left( (\tilde{s} - s_v)^\top \mathbf{n}_v \right), \tag{2}$$

where $\tilde{\mathcal{S}}$ is the intersection of support borders and the surface. The $x$ axis is the projection of the vector directed from $s_v$ towards $\hat{s}$. Fig. 5(b) illustrates the full procedure. Note that there is no guarantee that the discerning point $\hat{s}$ from Eq. 2 is always repeatable: in particular, if different directions yield similar values of the signed distance, the $x$ axis will be ambiguous, hence the resulting LCFs could rotate about the $z$ axis. Therefore, as shown in Fig. 5(c), this approach produces LCFs quasi-covariant to pose changes, and as a result, only quasi-pose-invariant features $\mathbf{f}$. We leave such noise for forests to take care of during learning.

### 4.2 CVT-based features

The feature $\mathbf{f}(s_v)$ above describes surface geometries in volumes but is devised to match only surface vertices $v$. A more intriguing question is: can one match these points $s$ directly? In other words, instead of an auxiliary role of matching surfaces, can they also be associated to the template discriminatively and even participate in shape deformations (bottom row of Fig. 3)? We investigate this direction with a *volumetric* representation from centroidal Voronoi tessellations that haven shown some recent success in various applications [51], [52], *i.e.* $s$ is a CVT cell.

We use it to sample a distance field where every cell $s$ stores the Euclidean distance from the centroid to the surface $\partial\Omega$: $d(\mathbf{x}_s, \partial\Omega) = \min_{p \in \partial\Omega} d(\mathbf{x}_s, p)$, yielding a distance-transform like representation similar to the TSDT above.

#### 4.2.1 Haar-like spherical feature

The offset feature $\mathbf{f}(s_v)$ above is nevertheless not applicable here since it relies on regular grids. We propose a new feature $\mathbf{f}(s)$ with the following principles in mind. It should be able to characterize the local neighborhood of any point of the volumetric shape. This rules out the descriptors that rely on surface normals such as MeshHOG [32] and SHOT [33].

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



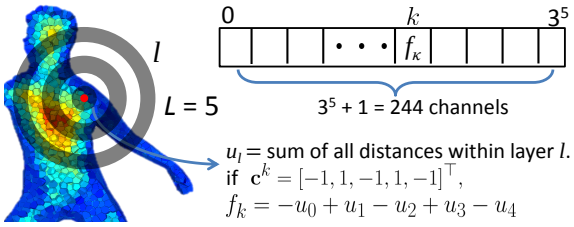



Fig. 6. CVT-based feature. Left: CVT cells $\mathcal{S}$ sample a distance field, where each cell stores the distance $d(\mathbf{x}_s, \partial\Omega)$. Blue to red colors means from close to far. Red dot: cell $s$ to be described. Right: a toy example of our feature $\mathbf{f}$, where $L = 5$. Shadowed and transparent layers have coefficients $c_l = -1$ and 1 respectively. See text for more explanations.

To be able to match any deformed pose with the template, we would like our feature to be pose-invariant. Therefore, we build it on the distance transform because it naturally encodes the relative location with respect to the surface and it is invariant to rotations, translations and quasi-invariant to pose changes. Finally, our feature needs to be robust to the topological noise present in the input data.

Given a distance field sampled by CVT cells $\mathcal{S}$, our feature is similar in spirit to Haar feature in the Viola-Jones face detector [41], except that the rectangular neighborhood is replaced with a sphere. As depicted in Fig. 6, we open an $L$-layer spherical support region in the Euclidean space around each cell. An $L$-dimensional vector $\mathbf{u}$ is defined accordingly, where each element $u_l$ is the sum of the distances of all cells falling within layer $l$. The feature value is the linear combination of all $u_l$, with coefficients $c_l$ chosen from a set $\Upsilon = \{-1, 0, 1\}$. Formally, suppose $\mathbf{c}$ are $L$-dimensional vectors whose elements are the bootstrap samples of $\Upsilon$. Let $\mathbf{c}^\kappa$ denote one particular instance of $\mathbf{c}$, *i.e.* , $\mathbf{c}^\kappa \in \Upsilon^L$. The feature value is then expressed as an inner product: $\mathbf{u}^\top \mathbf{c}^\kappa$, corresponding to one feature attribute $\kappa$. We consider all possible $\mathbf{c}^\kappa$ and also take the distance $d$ itself into account. $\mathbf{f}$ is hence a vector of $(3^L + 1)$ dimensions, where $3^L$ is the cardinality of $\Upsilon^L$ and each element $f_\kappa$ is defined as:

$$f_\kappa \triangleq \begin{cases} \mathbf{u}^\top \mathbf{c}^\kappa = \sum_l c_l^\kappa u_l, & \kappa < 3^L,\ c_l^\kappa \in \{-1, 0, 1\} \\ d(\mathbf{x}_s, \partial\Omega), & \kappa = 3^L \end{cases}. \quad (3)$$

Since each dimension $f_\kappa$ is computation-wise independent, $\mathbf{f}$ is suitable for decision forests, which select feature channels $\kappa$ randomly to split the data during training. Being derived from $d(\mathbf{x}_s, \partial\Omega)$, $\mathbf{f}$ inherits the invariance to rigid-body motions. As opposed to the early version of this work [53], we normalize the distances with respect to the averaged edge length of cells, achieving invariance to the body size to a certain extent. However, $\mathbf{f}$ is not invariant to pose changes as the contained cells in each layer vary with poses. Although considering geodesic spherical supports instead of Euclidean ones would overcome this issue and yield quasi-invariance to pose changes, the resulting feature would be highly sensitive to topological noise. Thus, we keep the Euclidean supports and let forests take care of the variations caused by pose changes in learning.

## 5 Correspondences Inference

Now that the features for both surface and volumetric associations, $\mathbf{f}(v)$ and $\mathbf{f}(s)$, are defined, we proceed on using them to train a regression forest, an ensemble of $T$ binary decision trees, to learn the mapping $\mathbf{r} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ from the observation domain to the template domain. During training each tree learns the split functions that best separate data recursively at branch nodes, while during testing the input point is routed through each tree, reaching $T$ leaves that store statistics as predictions. We discuss in § 5.1 a generic learning framework that applies to both shape parameterizations. A CVT-specific multi-template strategy is presented in § 5.2 to generalize the Vitruvian framework from single mesh connectivity to multiple ones.

### 5.1 Training and prediction

Broadly speaking, training a regression forest amounts to determining the following components: sample-label pairs, split functions, learning objectives and leaf-node statistical models. Readers are referred to [39] for a comprehensive analysis on different choices of these components.

#### 5.1.1 Training data and split functions

First we elaborate the training scenario for surface representations. Since forests aim to map an observed 3D vertex back to the template domain $\partial\Omega_X$, usually chosen to be in the rest (T or A) pose, it requires meshes in various poses but with the same connectivity for training. To incorporate abundant training variations, we animate the template $\mathbf{X}^0 = \{\mathbf{x}_v^0\} \subset \partial\Omega_X$ to a variety of poses with a method similar to [54]. After voxelizing all animated meshes, we associate each surface voxel to their locations at the rest pose, obtaining a pool of sample-label pairs $\mathcal{D} = \{(s_v, \mathbf{x}_v^0)\}$. Each tree is trained with a randomly bootstrapped subset of $\mathcal{D}$. While the split function may be arbitrarily complex, a typical choice is a stump where one single dimension $\kappa$ is compared to a threshold $\tau$, *i.e. axis-aligned thresholding*. Our splitting candidate $\phi$ is hence the pair of testing channels $\kappa$ and thresholds $\tau$, $\phi = (\kappa, \tau)$, where $\kappa$ is represented by offset pairs $\psi$ and binary variables $\eta$ in § 4.1. Let $\mathcal{D}_N$ denotes the samples arriving at a certain branch node. The training process is to partition $\mathcal{D}_N$ recursively into two subsets $\mathcal{D}_L$ and $\mathcal{D}_R$, based on randomly generated $\phi$:

$$\mathcal{D}_L(\phi) = \{s_v \in \mathcal{D}_N | f_\kappa(s_v) = \varepsilon(s_v; \mathbf{R}^\eta(\psi)) \geq \tau\}, \quad (4a)$$
$$\mathcal{D}_R(\phi) = \{s_v \in \mathcal{D}_N | f_\kappa(s_v) = \varepsilon(s_v; \mathbf{R}^\eta(\psi)) < \tau\}. \quad (4b)$$

Similarly, given a set of CVTs corresponding to the template volumes $\Omega_X$ deformed in various poses, we associate each cell $s \in \mathcal{S}_X$ to its locations in the rest pose, denoted as $\mathbf{x}_s^0 \in \mathbf{X}^0 \subset \Omega_X$, forming a pool of sample-label pairs $\mathcal{D} = \{(s, \mathbf{x}_s^0)\}$ as the dataset. The split candidate $\phi$ is again the pair of thresholds and feature attributes, $\phi = (\kappa, \tau)$, where features are instead computed according to Eq. 3 but the thresholding criteria in Eqs. 4a and 4b follows.

#### 5.1.2 Learning objectives and leaf predictions

At branch nodes, many candidates $\phi$ are randomly generated and the one that maximizes the information gain $I$, $\phi^* = \text{argmax}_\phi I(\phi)$, is stored for the later prediction use. We follow the classic definition of information gain:

$$I(\phi) = H(\mathcal{D}_N) - \sum_{i \in \{L,R\}} \frac{|\mathcal{D}_i(\phi)|}{|\mathcal{D}_N|} H(\mathcal{D}_i(\phi)), \quad (5)$$

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



where $H$ is the entropy, measured as the variance in Euclidean space, *i.e.* $H = \sigma^2$ for both parameterizations. The tree recursively splits samples and grows until one of the following stopping criteria is met: (1) it reaches the maximum depth, or (2) the number of samples $|\mathcal{D}_N|$ is too small. A Mean-Shift clustering [55] is performed in a leaf node to represent the distributions of $\mathbf{x}^0$ as a set of confidence-weighted modes $\mathcal{H} = \{(\mathbf{h}, \omega)\}$. $\mathbf{h} \in \mathbb{R}^3$ is the mode location and $\omega$ is a scalar weight.

In the prediction phase, a 3D input point $i \in \mathcal{V}_Y$ or $i \in \mathcal{S}_Y$ traverses down the trees and lands on $T$ leaves containing different collections of modes: $\{\mathcal{H}_1 \cdots \mathcal{H}_T\}$. The final regression output $\mathbf{r}_i$ is the cluster centroid with largest weight obtained by performing Mean-Shift [55] on them. Each observed point then gets a closest point $p$ in the reference shape $\mathbf{X}^0$, either in surfaces, $p = \operatorname{argmin}_{v \in \mathcal{V}_X} \left\|\mathbf{r}_i - \mathbf{x}_v^0\right\|_2$, or in CVTs, $p = \operatorname{argmin}_{s \in \mathcal{S}_X} \left\|\mathbf{r}_i - \mathbf{x}_s^0\right\|_2$. The correspondence pair $(i, p)$ serves as input to the subsequent deformation framework described in § 6.

Outliers such as false geometries, or un-removed background elements often exist in 3D data, drastically deteriorating tracking results. If their models are available, we also include them in the training process, so that forests can identify and reject them online. In this case, the goodness of a split $\phi$ is evaluated in terms of both classification and regression. We follow Fanelli *et al.* [56] and extend the entropy to be:

$$H(\mathcal{D}) = -\sum_c p(c|\mathcal{D}) \log p(c|\mathcal{D}) + (1 - e^{\frac{\delta}{\alpha}})\sigma^2(\mathcal{D}), \quad (6)$$

where $p(c|\mathcal{D})$ is the class probability of being foreground or background. It is the weighted sum of the aforementioned regression measure $\sigma^2$ and the classification entropy measure. Forests trained with Eq. 6 are often referred to as *Hough forests*. During training it learns simultaneously (1) how to distinguish between valid and invalid samples (outliers) and (2) how to match valid samples to the template. The regression part gets increasing emphasis when the current depth $\delta$ gets larger (*i.e.* the tree grows deeper), and the steepness is controlled by the parameter $\alpha$.

### 5.2 Learning across multiple volumetric templates

So far we know how to utilize Vitruvian-based learning framework to match surface or volumetric data against the template. For the training purposes, one has to deform the reference mesh into various poses such that all meshes share a consistent topology $\mathcal{T}_X$ and one can easily assign each sample a continuous label which is its rest-pose position $\mathbf{X}^0$. In this regards, the trained forest applies only to one mesh connectivity $\mathcal{T}_X$. Nevertheless, the amount of training data for one single template is often limited. To avoid over-fitting, the rule of thumb is to incorporate as much variation as possible into training. This motivates us to devise an alternative that learns across different template connectivities $\mathcal{T}_X$. Due to the high memory footprint of regular voxel grids, this strategy is unfortunately less practical for the surface feature $\mathbf{f}(v)$ in § 4.1 and we implement it only with CVTs.

Given $U$ distinct CVT templates: $\{\mathcal{S}^\mu\}_{\mu=1}^U$[3], whose temporal evolutions are recovered with the method in [51], resulting in a collection of different templates deformed in various poses: $\{\{\mathbf{X}_1^t\} \cdots \{\mathbf{X}_U^t\}\}$ as our dataset. To include all of them into training, we take one generic template $\hat{\mathcal{S}}$ as the reference. Intuitively, if there exists a mapping $g$ that brings each cell $s \in \mathcal{S}^\mu$ to a new cell $g(s) = \hat{s} \in \hat{\mathcal{S}}$, one only needs to change the template-specific labels $\mathbf{x}_s^0$ to the corresponding $\mathbf{x}_{\hat{s}}^0$, which are common to all templates, and the training process above can again be applied. In other words, we align topologies by matching every template $\mathcal{S}^\mu$ to $\hat{\mathcal{S}}$. Fig. 7 depicts this multi-template learning scheme.

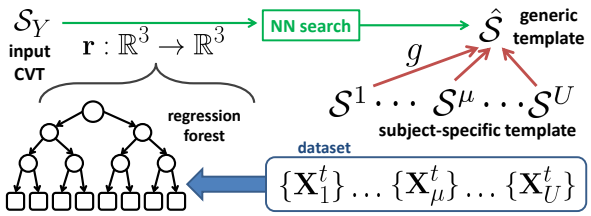



Fig. 7. The schematic flowchart of the multi-template learning framework. Red arrows: mappings $g^\mu$ that associate the indices from each subject-specific template $\mathcal{S}^\mu$ to the common one $\hat{\mathcal{S}}$. $\mathbf{X}_\mu^t$ are the temporal evolutions of each template. Blue: training; green: prediction.

Although various approaches for matching surface vertices exist, only a handful of works discuss matching voxels/cells. Taking *skinning weights* [57] as an example, we demonstrate in the following how to adapt a surface descriptor to CVTs. Note that our goal is not to propose a robust local 3D descriptor. With proper modifications, other descriptors can be used as well for shape matching.

#### 5.2.1 Generalized skinning weights.

Skinning weights are originally used for skeleton-based animations, aiming to blend the transformations of body parts (bones). Usually coming as a side product of the skeleton-rigging process [58], it is a vector $\mathbf{w}$ of $B$-dimensions, each corresponding to a human bone $b$ and $B$ is the number of bones. The non-negative weight $w_b$ indicates the dependency on that part and is normalized to sum up to one, *i.e.* $\sum_b w_b = 1$. As such, a skinning weight vector $\mathbf{w}$ is actually a probability mass function of body parts, offering rich information about vertex locations. To extend it from surface vertices to CVT cells, we first relax the unity-summation constraint as $\mathbf{w}$ is not used to average transformations of bones but only as a descriptor here. The intuition behind the adaptation is that, a CVT cell should have bone dependencies similar to the closest surface point. Therefore, for a cell whose normalized distance to the surface is $d$, its skinning weight is simply the one of its closest surface point, scaled by $e^d$. We tackle scale changes by normalizing the distance field with the averaged edge length of cells in the shape. Since the shortest distance usually hits a triangle rather than a single vertex, we use barycentric coordinates as the coefficients to linearly combine the skinning weights of the three vertices. Note that this does not violate the unity-summation constraint for surface vertices as their distance $d$ is still zero. We illustrate this concept in Fig. 8(a). The mapping $g$ is then determined by searching for the nearest neighbor in the skinning weight space: $g(s) = \arg\min_{\hat{s} \in \hat{\mathcal{S}}} \|\mathbf{w}_{\hat{s}} - \mathbf{w}_s\|_2$.

In practice, we use `Pinocchio` [58] to computes skinning weights, extend them from surface vertices to CVT

3. The template suffix $X$ is dropped to keep notations uncluttered.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



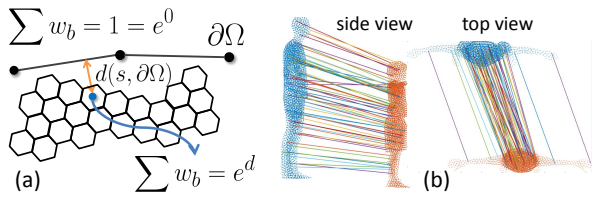



Fig. 8. (a): illustration of our strategy adapting skinning weights to CVT cells. Distances to the surface $d(\mathbf{x}_s, \partial\Omega)$ are reflected in the normalization constants $e^d$. (b): result of matching two templates.

cells, and match all cells to those of the common template $\hat{\mathcal{S}}$. The resulting skeletons are not used in our method. Fig. 8(b) visualizes one example of matching results. Our approach yields reasonable matches, regardless of the difference in body sizes. Due to the descriptiveness of skinning weights, symmetric limbs are not confused. Note that this computation is performed only between user-specific templates $\mathcal{S}^\mu$ and the generic one $\hat{\mathcal{S}}$ off-line once. Input data $\mathcal{S}_Y$ cannot be matched this way, because rigging a skeleton for shapes in arbitrary poses remains a challenging task.

## 6 Tracking

Recall that our goal is not only to detect the associations $\mathcal{C}$ but eventually to estimate the deformation parameter $\hat{\mathbf{\Theta}}$ via $\hat{\mathbf{\Theta}} = \operatorname{argmin}_{\mathbf{\Theta}} E(\mathbf{\Theta}; \mathcal{C})$, such that the resulting $\mathbf{X}(\hat{\mathbf{\Theta}})$ best explains $\mathbf{Y}$. The choice of $\mathbf{\Theta}$ could be raw point positions [59], [60], skeletal kinematic chains [4], [61] and cage [62]. We opt for a patch-based deformation framework [2] for surfaces and a CVT cluster-based method [51] for volumetric meshes respectively. Both group the 3D points into a higher-level structure, where shape deformations are represented as the ensemble of their rigid-body motions $\boldsymbol{\theta}$. We briefly explain here the basic principles and how to apply the predicted correspondences in § 5 to track a sequence of temporally inconsistent observations.

### 6.1 Surface-based deformation

In [2], the reference surface is decomposed into several patches $k$. It serves as a intermediate deformation structure between vertex positions and anatomical skeletons. Without any prior knowledge of motion, patches are preferred to be distributed uniformly over $\mathcal{X}$. Given correspondences $\mathcal{C}$ from above, a data term is formulated as:

$$E_{data}(\mathbf{\Theta}; \mathcal{C}) = \sum_{(i,p)\in\mathcal{C}} w_{ip} \|\mathbf{y}_i - \mathbf{x}_p(\mathbf{\Theta})\|_2^2, \quad (7)$$

which is a standard sum of weighted squared distances.

Since evolving a surface with discrete observations (even with a good $\mathcal{C}$) is ambiguous by nature, regularization terms are usually introduced to exert soft constraints. Given a vertex $v$, the rigidity constraint enforces the predicted positions $\mathbf{x}_v(\boldsymbol{\theta}_k)$ and $\mathbf{x}_v(\boldsymbol{\theta}_l)$ from two adjacent patches $P_k$ and $P_l \in \mathcal{N}_k$ to be consistent:

$$E_r(\mathbf{\Theta}) = \sum_{k=1}^{K} \sum_{P_l\in\mathcal{N}_k} \sum_{v\in P_k\cup P_l} w_{kl} \left\|\mathbf{x}_v(\boldsymbol{\theta}_k) - \mathbf{x}_v(\boldsymbol{\theta}_l)\right\|_2^2, \quad (8)$$

where $\mathbf{\Theta}$ is implicitly encoded in $\mathbf{x}_v(\boldsymbol{\theta}_k)$ and $\mathbf{x}_v(\boldsymbol{\theta}_l)$.

Given a fixed input $\mathbf{Y}$, the regression forest returns a fixed response $\tilde{\mathbf{Y}}$, and in turn a fixed $\mathcal{C}$. We therefore apply standard Gauss-Newton method directly to find the minimizer of final energy: $E(\mathbf{\Theta}; \mathcal{C}) = \lambda E_{data}(\mathbf{\Theta}; \mathcal{C}) + E_r(\mathbf{\Theta})$, where $\lambda$ defines the softness of the template and is empirically set to 10 throughout our experiments. Note anyway that refining $\mathcal{C}$ like non-rigid ICP does is always possible. In this case, our method provides better initializations than using last frame results, reducing the needed ICP-iterations.

### 6.2 Volumetric deformation

On the other hand, a similar deformation framework can be formulated for CVTs as well, only that surface patches $k$ are replaced by clusters of cells. We follow [51] which is essentially a non-rigid ICP method. As opposed to the extensive correspondence search, we again directly use the association pairs $(i, p)$ detected by the forest as initializations. This results in a faster pose estimation.

## 7 Experiments

The presented method is evaluated extensively in this section. We verify the merits of the discriminative associations as well as the complete 3D tracking-by-detection pipeline, in both surface and CVT parameterizations. As summarized in Table 1 in the supplemental material, in total 15 datasets are considered for various evaluation purposes. Due to the availability of ground-truths, the input in § 7.1 is the non-rigid registration, whereas in § 7.2 it is the reconstructed visual hull from [38] or raw tessellated CVT from [63].

### 7.1 Discriminative associations

Recall that the goal of discriminative correspondences is to guide the shape deformation not to match non-rigid 3D shapes accurately. We aim only to show that (1) the presented features are more or at least equally informative for matching humanoid surfaces than the existing state-of-the-arts 3D descriptors, *e.g.* Heat Kernel Signature (HKS) [30], [64] or Wave Kernel Signature (WKS) [31] and (2) CVT-based associations are more reliable than the surface-based counterparts. We describe every vertices with HKS, WKS, and our pair-wise offset features $\mathbf{f}(v)$ in § 4.1. CVT cells are, on the other hand, described by the Haar-like spherical features $\mathbf{f}(s)$ in § 4.2. The forests learn to match these 3D primitives against their own learning template, either a generic reference surface (*FAUST*) or a subject-specific CVT template (*Goalkeeper*, *Ballet* and *Thomas*).

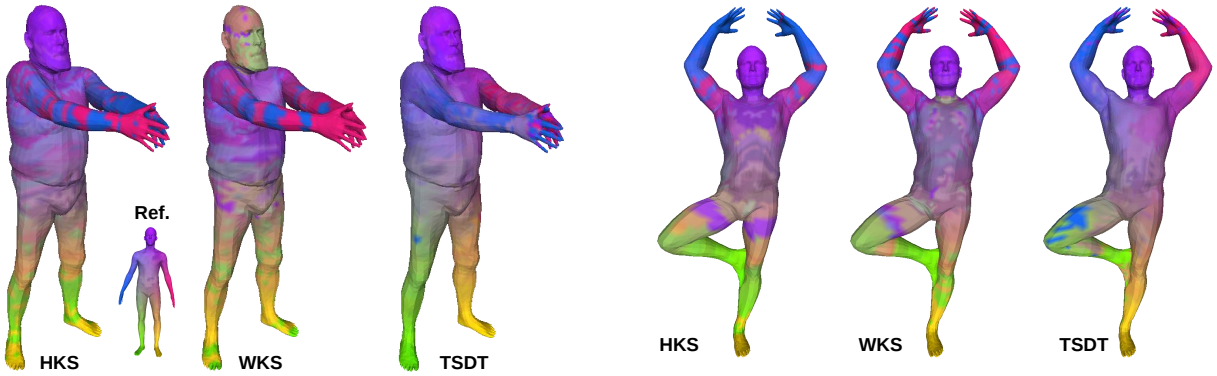
#### 7.1.1 Surface-based correspondences

Surface correspondences are validated on the publicly-available dataset *FAUST* [65]. Following [16], [34], we use only the training set because of the availability of ground-truth vertex indices. It comprises 100 static 3D scans from 10 subjects in 10 poses. The accuracy on *FAUST* indicates how well the proposed method deals with human shape variations. Specifically, half the registrations (50 meshes) are chosen to train $T = 50$ trees and the other half are left out for testing. At branch nodes, 5000 splitting candidates $\phi$ are randomly generated and the best one is stored. The error measure is the geodesic distance between predicted vertices

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308






(a) unseen poses (b) unseen shapes

Fig. 9. Qualitative results of surface matching on *FAUST*. Best viewed in pdf.

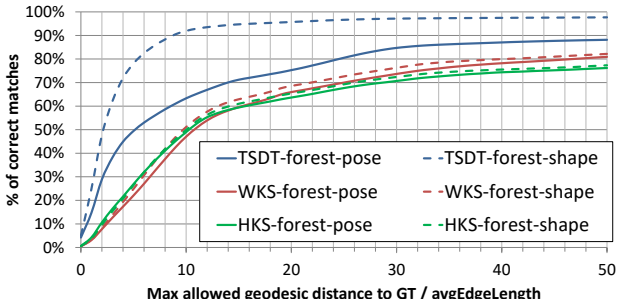



Fig. 10. Cumulative errors on *FAUST* [65].

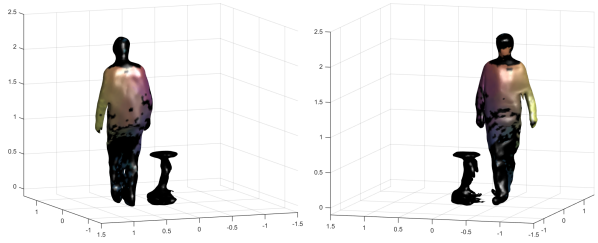


Fig. 11. Predicted data-model associations on noisy visual hulls with Hough forests. Black color means that the points are either outliers, or the inferred correspondences are rejected due to incompatible normals.

and ground-truths. If the distance is smaller than a certain threshold, we consider the point correctly matched. The percentage of correct matches in varying thresholds characterizes the performance of one algorithm and is commonly used in many matching papers [16], [34].

The results are shown in Fig. 10, where $x$-axis is normalized by the averaged edge length of the template. We partition the 100 meshes in two ways to test the generalization to unseen shapes or poses. The keyword *pose* means that the forest is trained with meshes in all 10 subjects but in only 5 poses, whereas *shape* represents the opposite. To compare fairly with other existing methods, we keep the Vitruvian-manifold label space unchanged (*i.e.* the same learning template) while replacing the voxel-based features with 30-dimensional scale-invariant HKS or WKS feature vectors. The proposed TSDT-forest combination yields overall best accuracy in Fig. 10, suggesting that the voxel-based TSDT feature is indeed more informative than H/WKS in the chosen parameter range. Comparing the blue solid curve to the dashed one, we notice that our approach handles unseen shapes better than unseen poses. This is due to the fact that our feature relies mainly on 3D geometry, in which pose variations cause more significant changes than shape variations. Although this phenomenon is not observed in the curves of H/WKS because they exploit the spectral domain for better pose invariance, they suffer from the confusion between symmetric parts as visualized in Fig. 9.

We further visualize in Fig. 11 the predicted associations on noisy reconstructed visual hulls with outliers, where no ground truths are available. Black colors means that the predicted correspondences are either rejected by simple normal compatibility check [2] like those on the body, or rejected because they are recognized as the chair. In this experiment, we include chair meshes into training data and follow Eq. 6 as the entropy measure to grow the trees. As a result, we can identify observations on the chair online and remove them, so that they do not affect the subsequent tracking stage. The task of trees here is throwing away the points of known outlier classes and in the meanwhile also predicting correspondences for the remaining points.

As one can see, our approach is capable of predicting reasonable associations for noisy visual hulls while rejecting outliers. This is of importance since they are the real input data of the final tracking-by-detection pipeline. HKS and WKS are known for their sensitivity to topological noises, *e.g.* the merging of arms and torso. We however would like to remark that, as oppose to our feature vector $\mathbf{f}(v)$ that has a dimensionality virtually longer than 5000 from the randomly generated splitting candidates at each branch node, HKS and WKS are only 30-dimensional in our experiment. To fully conclude that the presented voxel-based feature is certainly better than HKS or WKS requires a more fair and thorough comparison but is not the main goal of this paper.

### 7.1.2 *Volumetric correspondences*

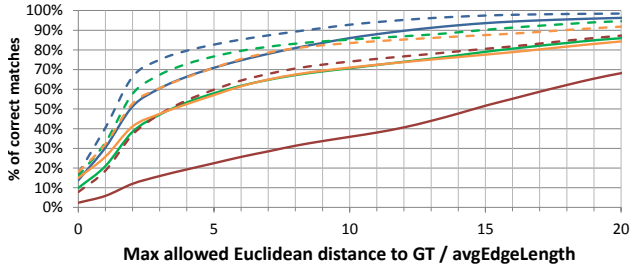
The discriminative CVT-based correspondence in § 4.2 is validated with 6 sequences from 3 subjects: *Goalkeeper*, *Ballet* and *Thomas*. We register each template to the corresponding raw CVT sequences using a EM-ICP based

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

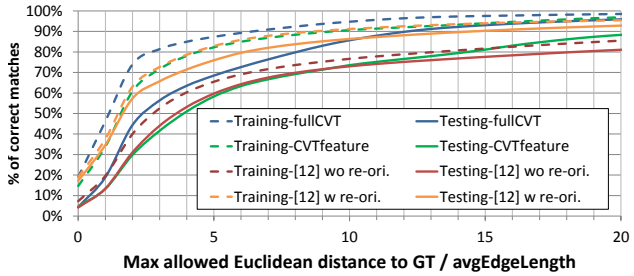
This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308




Fig. 12. Qualitative results of volumetric matching on the raw CVTs. Best viewed in pdf.




(a) *Thomas*




(b) *Ballet*

Fig. 13. Cumulative matching accuracy of different approaches. The $x$-axis is normalized w.r.t. the average edge length of the templates. The number of trees $T$ is 20 in this experiment. Dashed and solid lines are accuracies on training (Tr) and testing (Te) sequences respectively.

method [51] to recover temporal coherent volumetric deformations (tracked CVTs). For each subject, up to 250 tracked CVTs are randomly chosen from the first sequence to form the training set, while the second sequences are completely left out for testing. We open $L = 8$ sphere layers for the feature computation. Each tree is grown up to depth 20 with $30\%$ bootstrap samples randomly chosen from the dataset.

The contributions of CVT on improving the correspondences detection are evaluated in two aspects. First, we keep using the Vitruvian manifold $\partial\Omega$ as the regressing domain but replace the voxel-based features $\mathbf{f}(v)$ with the spherical feature $\mathbf{f}(s)$, denoted as *CVTfeature*. Next, we further change the label space from surfaces $\partial\Omega$ to volumes $\Omega$, termed *fullCVT*. We test on the tracked CVTs and report the results on all frames of both training sequences (Tr) and testing ones (Te). The drop between them indicates the ability to generalize. The same error measure as in the previous subsection is applied, only the geodesic distances are replaced by Euclidean ones. To yield a fair comparison with [11], here the forests are subject-specific and consist of $T = 20$ trees.

Fig. 13 shows the percentage of correct matches in varying thresholds for *Thomas* and *Ballet*. Since *CVTfeature*

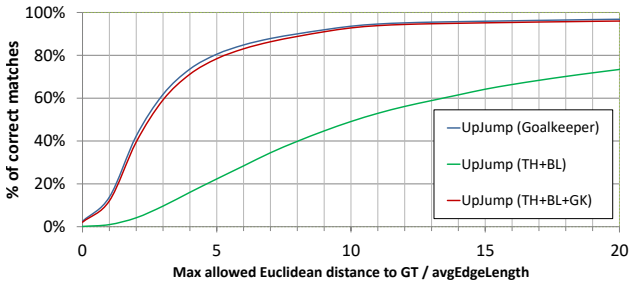



Fig. 14. Cumulative matching accuracy on *Goalkeeper*.

and [11] are regressing to surfaces whereas *fullCVT* regresses to volumes, we normalize the $x$-axis by the average edge length of templates to yield fair comparisons. While the results of *CVTfeature* are comparable to [11] (green vs. red or orange), *fullCVT* attains the improved accuracies (blue vs. red or green), demonstrating the advantages of our fully volumetric framework. Some visual results of the *fullCVT* approach on raw CVT input are shown in Fig. 12.

It is worth a closer analysis to highlight the advantages of CVT-based feature $\mathbf{f}(s)$ against the voxel-based one $\mathbf{f}(v)$. Our early work [11] applied $\mathbf{f}(v)$ that takes $150^3$ voxels for $\mathbf{f}(v)$ to describe a human shape, while CVT needs only 5k cells[4]. Consequently, [11] is not able to include a sufficient amount of training shapes, leading to a major drawback that forests are limited to one single subject. To further decrease the needed number of training meshes, [11] exploits skeletal poses to cancel the global orientation. This in turn makes every mesh in the training dataset face the same direction and we learn merely pose variations. It follows that during tracking the input data has to be re-oriented likewise using the estimated skeletal poses from the last frame. The CVT-based feature $\mathbf{f}(s)$, on the other hand, considers distance fields of cells which is naturally invariant to rotations and hence does not require re-orientations. We anyway compare to [11] in both settings. Orange curves in Fig. 13 shows the results with re-orientation, which is better than the proposed strategy in *Ballet*. Nonetheless, without re-orienting data, the accuracy drops substantially during testing (compare red to orange). The efficiency on memory and the invariance of our features are two determining reasons why the presented method is better than [11]. With the multi-template learning strategy in § 5.2, it takes just one forest for different subjects in the tracking-by-detection experiment in § 7.2.

4. Further note that [11] stores a 3D vector in each voxel, whereas we store a scalar in each CVT cell. So the ratio is $3 \times 150^3$ to 5k.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



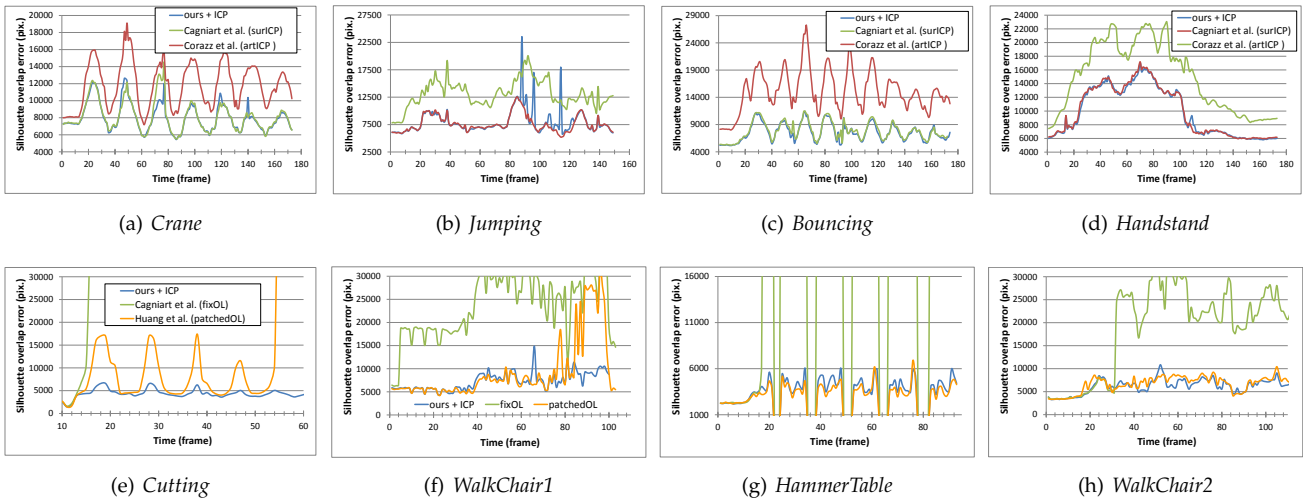



Fig. 15. Pixel overlap error of 8 sequences, averaged over all cameras. Image resolution: (a-d): 1920×1080; (e-h): 1000×1000. Best viewed in pdf.

Next, we use the sequences of *Goalkeeper* to verify the merits of this multi-template learning framework, which is unfeasible for voxel-based feature $\mathbf{f}(v)$ due to the high memory footprint. It is a particularly difficult dataset because motions in the testing sequence *UpJump* have little overlap with those in the training *SideJump*. We report in Fig. 14 the correctness of correspondences for *UpJump* (unseen during training) in *fullCVT* setting. Three situations are taken into account: training with the tracked CVTs of all three subjects (red), training only with those from *Goalkeeper* (blue) and without *Goalkeeper* (green). For red and green curves, we choose the *Goalkeeper* template as the common one $\hat{\mathcal{S}}$ and follow the strategy in § 5.2 to align distinct CVT tessellations. Comparing the red curve to the blue one confirms the advantage of this cross-template approach, leading to a forest that applies to all three subjects without trading off much accuracy. Nonetheless, the training data of the testing subject is still indispensable, as the accuracy drops when there is no tracked CVTs of *Goalkeeper* (green vs. red or blue), even if the forest of green curve is trained with twice the amount of CVTs compared to the blue curve. This is partially due to the fact that template of *Goalkeeper* has much smaller size than the other two and suggests that the proposed Haar-like feature in Eq. 3 captures more shape than pose information.

## 7.2 Tracking-by-detection

Now we move on to evaluate the full tracking-by-detection pipeline. The predicted associations $\mathcal{C}$ of two parameterizations are fed into their respective shape deformation frameworks in § 6 and the tracking is carried out on a frame-by-frame basis. The fidelity of estimated shapes is verified by the widely-used silhouette overlap error.

### 7.2.1 Surface-based

An individual forest is trained for each subject with up to 200 meshes, depending on the number of vertices in the template. For *Baran* and *Vlasic*, we train standard regression forest; for *Lionel* and *Ben* we apply the adaptation in Eq. 6 ($\alpha = 5$) due to the un-properly segmented chairs and tables in input data. Growing $T = 20$ trees to depth 25 with 5000 testing offset pairs $\psi$ takes about 3 hours. Although it is not the aim of this paper, we anyway augment the energies in § 6.1 with the skeleton energy in [2] and validate the estimated human poses in 2D.

TABLE 1
Average silhouette overlap error in pixels 4 sequences at low frame rate. Image resolution: 1920×1080.

| | *Crane* | *Jumping* | *Bouncing* | *Handstand* |
|---|---|---|---|---|
| ours | 7746.40 | 9148.94 | 6847.72 | 9279.57 |
| surICP [2] | 8295.58 | 16759.29 | 9400.76 | 11690.61 |

For sequences without outliers, we compare with surface-based ICP (surICP) [2] and articulated ICP (artICP) [15], both of which explain data with GMM using the Expectation-Maximization algorithm. We run an additional ICP step to reduce the errors (ours + ICP) for all testing sequences. The averaged overlap errors are shown in Fig. 15(a-d). In general, our method performs much better than artICP and attains comparable results with surICP. However, ICP-based methods often fail when large deformation occurs between consecutive frames, which is usually the case in videos of low frame rates. We simulate this by tracking only every three frames. As reported in Table 1, surICP now yields higher errors because local proximity search fails to estimate correspondences properly, while our approach is able to handle large jumps between successive input.

Four of our testing sequences, *Cutting*, *WalkChair1*, *HammerTable*, and *WalkChair2* contain tables or chairs in observations, which play the roles as static outliers. We compare with other outlier rejection strategies such as, fixed outlier proportion (fixOL) [2], removing outliers by body-part classifications with SVM (bpSVM) [66], and modeling outlier likelihood dynamically by aggregating over all patches (patchedOL) [3]. As shown in Fig. 15(e-h), conventional outlier strategy fixOL drifts quickly and soon fail to track (green curves). ICP with robust outlier treatment, patchedOL, is able to sustain noisy input to a certain extent. Once it starts drifting, the error only gets higher due to its ICP nature (yellow curves). When subjects and outliers are sperate

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



components in visual hulls, we cast them into separately TSDT, and feed them into the joint classification-regression forest. If they are connected to each other, forests inevitably associate some outliers to the humanoid template, leading to undesirable deformations as suggested by the spike in blue curves in Fig. 15(f). Nonetheless, since we rely less on previous frames for data associations, the results can always get recovered when they are separate again. In average, we still yield low errors throughout the whole sequences. We remark that such ability to recover is the essence of our discriminative approach, which is the biggest advantages over the existing generative methods. The recovered shapes and poses, superimposed on original images, are also presented in Fig. 2(c) in the supplementary material.

#### 7.2.2 Fully volumetric tracking-by-detection

After evaluating the surface-based tracking-by-detection framework, now we turn to evaluate the volumetric one. We compare in two quantitative metrics against the whole pipeline in [11], which is the early version of our surface-based tracking-by-detection approach.

Unlike the matching experiment in the previous subsection, here we apply the multi-template strategy in § 5.2 to train one universal regression forest, with *Goalkeeper* chosen as the common template $\hat{\mathcal{S}}$. Training $T = 50$ trees up to depth 20 where each one is grown with around 200 CVTs (approximately one million samples) takes about 15 hours on a 24-core Intel Xeon CPU machine. For each subject, we track the testing sequence, which is not part of the training set. Tracking inputs are raw CVTs that have no temporal coherence. The number of clusters $K$ is 250 for *Ballet* and *Goalkeeper* and 150 for *Thomas*. We evaluate our tracking approach with two different metrics. On one hand, evaluation with marker-based motion capture evaluates the correctness of the surface pose, but only for a sparse set of surface points. On the other hand, the silhouette overlap error evaluates the shape estimate but not the estimated pose. Hence these metrics are complementary.

Some visual results are shown in Fig. 3 in the supplemental material and video[5]. Our approach is able to discover volumetric associations even in challenging poses found in *Thomas* and deform the templates successfully. As shown in Table 2-4, we evaluate the results by computing the overlap error between the ground truth silhouette and the projection of the estimated surface. The metric we use is the pixel error (number of pixels that differ). Statistics are computed on all frames of all cameras. The *Ballet/Seq2* sequence has marker-based motion capture data: fifty markers were attached to the body of the subject, providing a sparse ground truth for surface tracking. First, each marker is associated to a vertex of the template surface. Then, for each marker, we measure the distance between its location and the estimated vertex location. Statistics on the distance are reported on Table 5. We observe that our approach attains slightly better performances than a state of the art ICP-based approach [51] and outperforms a surface-based tracking-by-detection [11] which mostly fails to correctly register the legs of the subject.

5. https://hal.inria.fr/hal-01300191

TABLE 2
Silhouette pixel error on sequence *Goalkeeper/UpJump*. Image size is 2048×2048.

| method | mean | stddev. | median | max |
|---|---|---|---|---|
| Proposed | 15221 | 6843 | 14754 | 57748 |
| Huang *et al.* [11] | 19838 | 14260 | 15607 | 109428 |
| Allain *et al.* [51] | 14773 | 6378 | 14355 | 43359 |

TABLE 3
Silhouette pixel error on sequence *Ballet/Seq2*. Image size is 1920×1080.

| method | mean | stddev. | median | max |
|---|---|---|---|---|
| Proposed | 2620 | 1041 | 2557 | 8967 |
| Huang *et al.* [11] | 5427 | 2809 | 4863 | 39559 |
| Allain *et al.* [51] | 2606 | 1008 | 2571 | 7642 |

TABLE 4
Silhouette pixel error on sequence *Thomas/Seq2*. Image size is 2048×2048.

| method | mean | stddev. | median | max |
|---|---|---|---|---|
| Proposed | 9991 | 7089 | 7968 | 78242 |
| Huang *et al.* [11] | 28731 | 23421 | 22991 | 354293 |
| Allain *et al.* [51] | 10199 | 7379 | 8022 | 81649 |

TABLE 5
Statistics of surface registration error at marker locations, on the *Ballet/Seq2* sequence.

| method | mean (mm) | stddev. (mm) |
|---|---|---|
| Proposed | 26.37 | 16.67 |
| Huang *et al.* [11] | 124.02 | 200.16 |
| Allain *et al.* [51] | 27.82 | 18.39 |

#### 7.2.3 Discussion

Last but not lease, we make a short comparison between the two presented features. As discussed above, voxel-based volume in § 4.1 has the downside of high memory footprint, which limits the allowed training variations. Aligning the orientations is one way to reduce the training variation such that forests only need to learn the pose variations of one single subject. One has to repeat the same thing likewise during the testing phase. In [11], we rely on the skeletal poses of previous frames for this purpose and thus the forest predictions are not fully frame-independent, exposing tracking subject to the potential risk of drifting. To facilitate a fully 3D tracking-by-detection framework, the information of previous frames is preferred no to participate in the discriminative correspondence estimation. On the other hand, the spherical feature presented in § 4.2 attempts to incorporate rotational, pose, and even shape variations during training, yielding completely frame-wise forest predictions.

As reported in Fig. 13, without aligning rotations, the accuracies of correspondences drop substantially on the testing sequences for the method in [11]. This means that voxel-based framework and the corresponding features do not generalize well to unseen rotations. When deployed in tracking applications, such unreliable associations eventually result in failure. In particular, one can observe in Table 4 that [11] attains high silhouette overlap discrepancy, most likely due to the fact that the subject rotates himself in many orientations and thus confuses the forest. From these observations, we conclude that the CVT-based Haar-like feature and the derived fully volumetric tracking-by-detection framework is better than the voxel-based counterpart.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



## 8 Conclusion

In this paper, we present two features for surface and CVT shape parameterizations respectively, both making use of volumetric distance fields to describe 3D geometry. Aiming to integrate with random forests, each feature attribute is computationally independent and can be obtained on the fly in testing. They facilitate the surface-based and CVT-based discriminative associations and in turn lead to the corresponding tracking-by-detection frameworks for 3D human shapes. While CVT-based approach is more robust to the surface counterpart, we show that both yield more stability compared with the respective generative ICP extensions. The reliability of the proposed method is confirmed by the experiments on numerous public sequences. Future directions include alleviating problems of topological changing and incorporating photometric information.

## Acknowledgments

Several datasets proposed in this paper have been acquired using the Kinovis platform at Inria Grenoble Rhône-Alpes (http://kinovis.inrialpes.fr).

## References

[1] D. Vlasic, I. Baran, W. Matusik, and J. Popović, "Articulated mesh animation from multi-view silhouettes," *TOG*, vol. 27, no. 3, p. 97, 2008.
[2] C.-H. Huang, C. Cagniart, E. Boyer, and S. Ilic, "A bayesian approach to multi-view 4d modeling," *IJCV*, vol. 116, no. 2, pp. 115–135, 2016.
[3] C.-H. Huang, E. Boyer, N. Navab, and S. Ilic, "Human shape and pose tracking using keyframes," in *CVPR*. IEEE, 2014, pp. 3446–3453.
[4] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt, "Markerless motion capture of multiple characters using multiview image segmentation," *TPAMI*, vol. 35, no. 11, pp. 2720–2735, 2013.
[5] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon, "Metric regression forests for human pose estimation," in *BMVC*, 2013.
[6] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *CVPR*. IEEE, 2012, pp. 103–110.
[7] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *PAMI*, 1992.
[8] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image and vision computing*, vol. 10, no. 3, pp. 145–155, 1992.
[9] E. Rodola, S. R. Bulo, T. Windheuser, M. Vestner, and D. Cremers, "Dense non-rigid shape correspondence using random forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 4177–4184.
[10] J. Shotton, A. Fitzgibbon, M. Cook, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*. IEEE, 2011, pp. 1297–1304.
[11] C.-H. Huang, E. Boyer, B. do Canto Angonese, N. Navab, and S. Ilic, "Toward user-specific tracking by detection of human shapes in multi-cameras," in *CVPR*. IEEE, 2015, pp. 4027–4035.
[12] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H. P. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," in *ACM SIGGRAPH 2008*, 2008.
[13] S. Osher and R. Fedkiw, *Level set methods and dynamic implicit surfaces*. Springer Science & Business Media, 2006, vol. 153.
[14] Q. Du, V. Faber, and M. Gunzburger, "Centroidal Voronoi tessellations: Applications and algorithms," *SIAM review*, vol. 41, pp. 637–676, 1999.
[15] S. Corazza, L. Mündermann, E. Gambaretto, G. Ferrigno, and T. P. Andriacchi, "Markerless motion capture through visual hull, articulated icp and subject specific model generation," *IJCV*, 2010.
[16] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li, "Dense human body correspondences using convolutional networks," in *CVPR*, 2016.
[17] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan, "High-quality streamable free-viewpoint video," *TOG*, vol. 34, no. 4, p. 69, 2015.
[18] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor *et al.*, "Fusion4d: Real-time performance capture of challenging scenes," *TOG*, vol. 35, no. 4, p. 114, 2016.
[19] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2, pp. 90–126, 2006.
[20] M. Straka, S. Hauswiesner, M. Rüther, and H. Bischof, "Simultaneous shape and pose adaption of articulated models using linear optimization," in *ECCV*. Springer, 2012.
[21] Y. Chen and G. Medioni, "Object modeling by registration of multiple range images," in *ICRA*. IEEE, 1991.
[22] M. Klaudiny, C. Budd, and A. Hilton, "Towards optimal non-rigid surface tracking," in *ECCV*. Springer, 2012, pp. 743–756.
[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," in *Journal of the royal statistical society*, 1977.
[24] B. Jian and B. C. Vemuri, "Robust point set registration using gaussian mixture models," *IEEE TPAMI*, vol. 33, no. 8, pp. 1633–1645, 2011.
[25] H. Li, B. Adams, L. J. Guibas, and M. Pauly, "Robust single-view geometry and motion reconstruction," in *TOG*, vol. 28, no. 5. ACM, 2009, p. 175.
[26] X. Wei, P. Zhang, and J. Chai, "Accurate realtime full-body motion capture using a single depth camera," *TOG*, vol. 31, no. 6, p. 188, 2012.
[27] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt, "Fast and robust hand tracking using detection-guided optimization," in *CVPR*, 2015, pp. 3213–3221.
[28] E. Boyer, A. M. Bronstein, M. M. Bronstein, B. Bustos, T. Darom, R. Horaud, I. Hotz, Y. Keller, J. Keustermans, A. Kovnatsky *et al.*, "Shrec 2011: robust feature detection and description benchmark," in *Eurographics 3DOR Workshop*. Eurographics Association, 2011, pp. 71–78.
[29] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. M. Kwok, "A comprehensive performance evaluation of 3d local feature descriptors," *IJCV*, vol. 116, no. 1, pp. 66–89, 2016.
[30] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," *CGF*, vol. 28, no. 5, pp. 1383–1392, 2009.
[31] M. Aubry, U. Schlickewei, and D. Cremers, "The wave kernel signature: A quantum mechanical approach to shape analysis," in *ICCV Workshops*. IEEE, 2011.
[32] A. Zaharescu, E. Boyer, and R. Horaud, "Keypoints and local descriptors of scalar functions on 2d manifolds," *International Journal of Computer Vision*, vol. 100, no. 1, pp. 78–98, 2012.
[33] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *ECCV*. Springer, 2010, pp. 356–369.
[34] Q. Chen and V. Koltun, "Robust nonrigid registration by convex optimization," in *ICCV*, 2015, pp. 2039–2047.
[35] J. Starck and A. Hilton, "Correspondence labelling for wide-timeframe free-form surface matching," in *ICCV*. IEEE, 2007, pp. 1–8.
[36] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
[37] D. Boscaini, J. Masci, E. Rodolà, and M. M. Bronstein, "Learning shape correspondence with anisotropic convolutional neural networks," in *NIPS*, 2016.
[38] J.-S. Franco and E. Boyer, "Efficient polyhedral modeling from silhouettes," *PAMI*, vol. 31, no. 3, 2009. [Online]. Available: https://hal.inria.fr/inria-00349103
[39] A. Criminisi and J. Shotton, *Decision forests for computer vision and medical image analysis*. Springer, 2013.
[40] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu, "Regression forests for efficient anatomy detection and localization in ct studies," in *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*. Springer, 2011.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2017.2740308



[41] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.
[42] T. Akenine-Möller, "Fast 3d triangle-box overlap testing," in *SIGGRAPH 2005 Courses*. ACM, 2005.
[43] W. Kehl, N. Navab, and S. Ilic, "Coloured signed distance fields for full 3d object reconstruction," in *BMVC*, 2014.
[44] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *ISMAR*. IEEE, 2011.
[45] A. Petrelli and L. Di Stefano, "On the repeatability of the local reference frame for partial shape matching," in *ICCV*. IEEE, 2011, pp. 2244–2251.
[46] C. S. Chua and R. Jarvis, "Point signatures: A new representation for 3d object recognition," *IJCV*, vol. 25, no. 1, pp. 63–85, 1997.
[47] A. Mian, M. Bennamoun, and R. Owens, "On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes," *IJCV*, vol. 89, no. 2-3, pp. 348–361, 2010.
[48] A. Petrelli and L. Di Stefano, "A repeatable and efficient canonical reference for surface matching," in *3DimPVT*. IEEE, 2012.
[49] J. Novatnack and K. Nishino, "Scale-dependent/invariant local 3d shape descriptors for fully automatic registration of multiple sets of range images," in *ECCV*. Springer, 2008.
[50] C.-H. Huang, F. Tombari, and N. Navab, "Repeatable local coordinate frames for 3d human motion tracking: From rigid to non-rigid," in *3DV*. IEEE, 2015, pp. 371–379.
[51] B. Allain, J.-S. Franco, and E. Boyer, "An efficient volumetric framework for shape tracking," in *CVPR*. IEEE, 2015. [Online]. Available: https://hal.inria.fr/hal-01141207
[52] L. Wang, F. Hétroy-Wheeler, and E. Boyer, "On volumetric shape reconstruction from implicit forms," in *ECCV 2016-European Conference on Computer Vision*, 2016.
[53] C.-H. Huang, B. Allain, J.-S. Franco, N. Navab, S. Ilic, and E. Boyer, "Volumetric 3d tracking by detection," in *CVPR*, 2016.
[54] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," in *TOG*. ACM, 2004.
[55] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *TPAMI*, vol. 24, no. 5, pp. 603–619, 2002.
[56] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3d face analysis," *IJCV*, vol. 101, no. 3, pp. 437–458, 2013.
[57] J. P. Lewis, M. Cordner, and N. Fong, "Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM, 2000.
[58] I. Baran and J. Popović, "Automatic rigging and animation of 3d characters," in *TOG*, 2007.
[59] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel, "Laplacian surface editing," in *Eurographics*, 2004.
[60] M. Zollhöfer, M. Nießner, S. Izadi, C. Rhemann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, and M. Stamminger, "Real-time non-rigid reconstruction using an rgb-d camera," *TOG*, vol. 33, no. 4, 2014.
[61] M. Ye, Y. Shen, C. Du, Z. Pan, and R. Yang, "Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera," *TPAMI*, vol. 38, no. 8, pp. 1517–1532, 2016.
[62] E. Duveau, S. Courtemanche, L. Reveret, and E. Boyer, "Cage-based motion recovery using manifold learning," in *3DimPVT*, 2012, pp. 206–213.
[63] L. Wang, F. Hétroy-Wheeler, and E. Boyer, "A hierarchical approach for regular centroidal Voronoi tessellations," in *CGF*, 2015.
[64] M. M. Bronstein and I. Kokkinos, "Scale-invariant heat kernel signatures for non-rigid shape recognition," in *CVPR*. IEEE, 2010, pp. 1704–1711.
[65] F. Bogo, J. Romero, M. Loper, and M. Black, "FAUST: Dataset and evaluation for 3D mesh registration," in *CVPR*, 2014, pp. 3794–3801.
[66] C.-H. Huang, E. Boyer, and S. Ilic, "Robust human body shape and pose tracking," in *3DV*. IEEE, 2013.

**Chun-Hao Huang** received the MSc degree in computer and communication engineering from National Cheng-Kung University in 2010. After one year in Academia Sinica as a research assistant, he started his doctoral study in Technische Universität München in 2012 and obtained the Ph.D. degree in 2016. His research interests include 2D/3D conversion, human motion capture and other 3D-vision related topics. He received Studying Abroad Scholarship from Taiwan government and the best paper award runner-up in 3DV'13.

**Benjamin Allain** received the MSc degree in computer science from Ensimag - Grenoble INP, France, in 2012. Then he joined the Morpheo group at Inria Grenoble Rhône-Alpes and obtained his PhD degree in computer science from Université Grenoble Alpes in 2017. Since November 2016, he has been working as a research scientist at Smart Me Up, France. His research interests include non-rigid motion tracking of 3D shapes from multiview video sequences.

**Edmond Boyer** is a senior research scientist at the INRIA where he leads the MORPHEO research team. He obtained his PhD from the Institut National Polytechnique de Lorraine in 1996. He was research assistant at Cambridge university in 1998 before joining the INRIA. His fields of competence cover computer vision, computational geometry and virtual reality. He has done pioneering work in the area of geometric 3D reconstruction with focus on objects with complex shapes like the humans. Edmond Boyer is co-founder of the 4D View Solutions (http://www.4dviews.com/), one of the leading companies worldwide specialized in multi-view acquisition and processing. His current research interests are on 3D dynamic modeling from images and videos, motion perception and analysis from videos.

**Jean-Sebastien Franco** is assistant professor of computer science at the Ensimag (School of Computer Science and Applied Mathematics, Grenoble Universities), and a researcher at the Inria Grenoble Rhône-Alpes and LJK lab, France, with the Morpheo team since 2010. He obtained his Ph.D. from the Institut National Polytechnique de Grenoble in 2005 with the Inria MOVI / Perception team. He started his professional career as a postdoctoral research assistant at the University of North Carolinas Computer Vision Group in 2006, and as assistant professor at the University of Bordeaux, with the IPARLA team, INRIA Bordeaux Sud-Ouest. His expertise is in the field of computer vision, with several internationally recognized contributions to dynamic 3D modeling from multiple views and 3D interaction.

**Federico Tombari** is a research scientist and team leader at the Computer Aided Medical Procedures Chair of the Technical University of Munich (TUM). He has co-authored more than 110 refereed papers on international conferences and journals in the field computer vision and robotic perception, on topics such as visual data representation, RGB-D object recognition, 3D reconstruction and matching, stereo vision, deep learning for computer vision. He got his Ph.D at 2009 from University of Bologna, at the same institution he was Assistant Professor from 2013 to 2016. He is a Senior Scientist volunteer for the Open Perception foundation and a developer for the Point Cloud Library, for which he served, in 2012 and 2014, respectively as mentor and administrator in the Google Summer of Code. In 2015 he was the recipient of a Google Faculty Research Award. His works have been awarded at conferences and workshops such as 3DIMPVT'11, MICCAI'15, ECCV-R6D'16.

**Nassir Navab** is a professor of computer science and founding director of Computer Aided Medical Procedures and Augmented Reality (CAMP) laboratories at Technische Universitt Mnchen (TUM) and Johns Hopkins University (JHU). He also has secondary faculty appointments at TUM and JHU medical schools. He received the Ph.D. degree from INRIA and University of Paris XI, France, and enjoyed two years of postdoctoral fellowship at MIT Media Laboratory before joining Siemens Corporate Research (SCR) in 1994. At SCR, he was a distinguished member and received the Siemens Inventor of the Year Award in 2001. In 2012, he was elected as a fellow member of MICCAI society. He received the 10 year lasting Impact Award of IEEE ISMAR in 2015. He holds 45 granted US patents and over 40 European ones. He has served on the program and organizational committee of over 80 international conferences including CVPR, ICCV, ECCV, MICCAI, IPCAI, and ISMAR. He has also been co-author of more than 20 awarded papers in international conferences. His current research interests include robotic imaging, computer aided surgery, computer vision and augmented reality.

**Slobodan Ilic** is currently senior key expert research scientist at Siemens Corporate Technology in Munich, Perlach. He is also a visiting researcher and lecturer at Computer Science Department of TUM and closely works with the CAMP Chair. From 2009 until end of 2013 he was leading the Computer Vision Group of CAMP at TUM, and before that he was a senior researcher at Deutsche Telekom Laboratories in Berlin. In 2005 he obtained his PhD at EPFL in Switzerland under supervision of Pascal Fua. His research interests include: 3D reconstruction, deformable surface modelling and tracking, real-time object detection and tracking, human pose estimation and semantic segmentation.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

$\therefore$

## A.7 MOULDING HUMANS : NON-PARAMETRIC 3D HUMAN SHAPE ESTIMATION FROM SINGLE IMAGES

Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, Gregory Rogez. ICCV 2019 - International Conference on Computer Vision, Oct 2019, Seoul, South Korea. pp.1-10

# Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images

Valentin Gabeur[1,2] Jean-Sébastien Franco[1] Xavier Martin[1] Cordelia Schmid[1,2] Grégory Rogez[3,†]
[1] Inria[*] [2] Google Research [3] NAVER LABS Europe

## Abstract

*In this paper, we tackle the problem of 3D human shape estimation from single RGB images. While the recent progress in convolutional neural networks has allowed impressive results for 3D human pose estimation, estimating the full 3D shape of a person is still an open issue. Model-based approaches can output precise meshes of naked under-cloth human bodies but fail to estimate details and un-modelled elements such as hair or clothing. On the other hand, non-parametric volumetric approaches can potentially estimate complete shapes but, in practice, they are limited by the resolution of the output grid and cannot produce detailed estimates. In this work, we propose a non-parametric approach that employs a double depth map to represent the 3D shape of a person: a visible depth map and a "hidden" depth map are estimated and combined, to reconstruct the human 3D shape as done with a "mould". This representation through 2D depth maps allows a higher resolution output with a much lower dimension than voxel-based volumetric representations. Additionally, our fully derivable depth-based model allows us to efficiently incorporate a discriminator in an adversarial fashion to improve the accuracy and "humanness" of the 3D output. We train and quantitatively validate our approach on SURREAL and on 3D-HUMANS, a new photorealistic dataset made of semi-synthetic in-house images annotated with 3D ground truth surfaces.*

## 1. Introduction

Recent works have shown the success of deep network architectures for the problem of retrieving 3D features such as kinematic joints [4, 33] or surface characterizations [43] from single images, with extremely encouraging results. Such successes, sometimes achieved with simple, standard network architectures [30], have naturally motivated the applicability of these methodologies for the more challenging problem of end-to-end full 3D human shape retrieval [2, 18]. The ability to retrieve such information from single images or videos is relevant to a broad number of applications, from self-driving cars, where spatial understanding of surrounding obstacles and pedestrians plays a key role, to animation or augmented reality applications such as virtual change rooms that can offer the E-commerce industry a virtual fitting solution for clothing or bodywear.

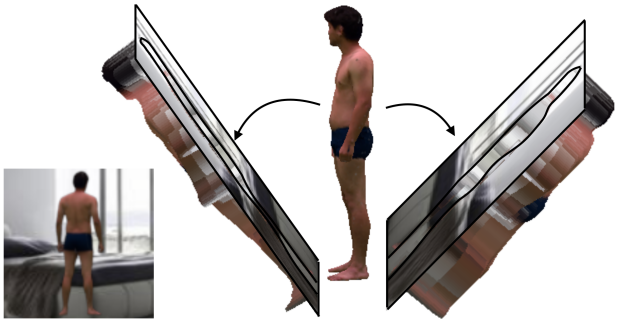


Figure 1. Our non-parametric representation for human 3D shape: given a single image, we estimate the "visible" and the "hidden" depth maps from the camera point of view. The two depth maps can be seen as the two halves of a virtual "mould". We show this representation for one of the images of our new dataset.

Designing a deep architecture that produces full 3D shapes of humans observed in an input image or a sequence of input images raises several key challenges. First, there is a representational issue. While the comfort zone of CNNs is in dealing with regular 2D input and output grids, the gap must be bridged between the 2D nature of inputs and the 3D essence of the desired outputs. One solution is to follow a parametric method and estimate the deformation parameters of a predefined human 3D model [2, 18]. These methods are limited to the level of details covered by the model. In contrast, non parametric approaches can potentially account for shape surface details but are prone to produce physically-impossible body shapes. This is the case of the recent volumetric approach proposed in [38] that encodes the human body as a voxel grid whose dimensionality directly impacts

[*] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France. [†] Most of the work was done while the last author was a research scientist at Inria.

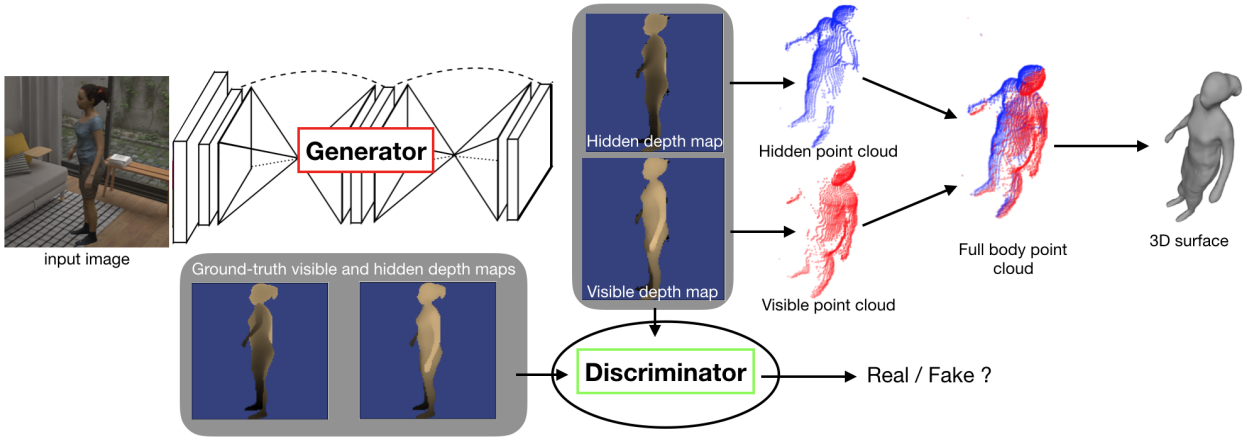



Figure 2. Overview. Given a single image, we estimate the "visible" and the "hidden" depth maps. The 3D point clouds of these 2 depth maps are combined to form a full-body 3D point cloud, as if lining up the 2 halves of a "mould". The 3D shape is then reconstructed using Poisson reconstruction [19]. An adversarial training with a discriminator is employed to increase the humanness of the estimation.

the precision of the estimation. This highlights a second challenge: the dimensionality of the problem is considerably higher than what existing networks have been shown to handle, because the parametrisation sought is no longer restricted to a subset of the variability, e.g. kinematic pose of humans or body shape parameters, but to an intrinsically finer description of the body. Finally, the training data for this problem, yet to be produced, requires a particularly demanding definition and acquisition effort. The large data variability of 3D problems has motivated some initial efforts to produce fully synthetic training sets [39], where such variability can be partially scripted. Yet recent successful methods underscore the necessity for realistic data, for both the general applicability of the estimation, and to keep the underlying network architecture simple, as devoid as possible of any domain specific adaptations.

In order to overcome these difficulties, we propose a non-parametric approach that employs a double depth map representation to encode the 3D shape of a person: a "visible" depth map capturing the observable human shape and a "hidden" depth map representing the occluded surface are estimated and combined to reconstruct the full human 3D shape. In this encoding of the 3D surface, the two depth maps can be seen as the two halves of a virtual "mould", see Figure 1. This representation allows a higher resolution output, potentially the same as the image input, with a much lower dimension than voxel-based volumetric representations, i.e. $O(N^2)$ vs $O(N^3)$. We designed an encoder-decoder architecture that takes as input a single image and simultaneously produces an estimate for both depth maps. These depth maps are then combined to obtain a point cloud of the full 3D surface which can be readily reconstructed using Poisson reconstruction [19]. Importantly, our fully differentiable depth-based model allows us to efficiently incorporate a discriminator in an adversarial fashion to improve the accuracy and "humanness" of the 3D output, especially in the case of strong occlusions. See Figure 2. To train and quantitatively evaluate our network in near real-world conditions, we captured a large-scale dataset of textured 3D meshes that we augment with realistic backgrounds. To account for the large variability in human appearance, we took special care in capturing data with enough variability in movements, clothing and activities. Compared to parametric methods, our method can estimate detailed 3D human shapes including hair, clothing and manipulated objects.

After reviewing the related work in Section 2, we present our two-fold contribution: our new non-parametric 3D surface estimation method is explained in Section 3 while our large-scale dataset of real humans with ground-truth 3D data is detailed in Section 4. Experiments are presented in Section 5 and conclusions drawn in Section 6.

## 2. Related Work

**3D object from single images.** Various representations have been adopted for 3D object shape estimation. Voxel-based representations [5] consist in representing the 3D shape as an occupancy map defined on a fixed resolution voxel grid. Octree methods [36] improve the computability of volumetric results by reducing the memory requirements. Point-clouds are another widely employed representation for 3D shapes. In [7], Fan et al. estimate sets of 1024 points from single images. Jiang et al. [15] build on this idea and incorporate a geometric adversarial loss (GAL) to improve the realism of the estimations. AtlasNet [10] directly estimates a collection of parametric surface elements to represent a 3D shape. Our representation combines two complementary depth maps aligned with the image, similar in spirit to the two halves of a "mould", and shares the resolution of the input image, capturing finer details while keeping output dimensionality reasonable.

Similarly to the work of Tatarchenko et al. [35] on reconstructing vehicle images from different viewpoints, we combine the estimation of several depth maps to obtain a 3D shape. For human shape estimation, however, we work on a deformable object. Also, we focus on the visible and hidden depth maps rather than any other because of their direct correspondence with the input image. Our two depth maps being aligned with the image, details as well as contextual image information are directly exploited by the skip connections to estimate the depth values. Multi-views [35] do not necessarily have pixel-to-pixel correspondences with the image making depth prediction less straightforward.

**3D human body shape from images.** Most existing methods for body shape estimation from single images rely on a parametric model of the human body whose pose and shape parameters are optimized to match image evidence [2, 11, 20, 29]. This optimization process is usually initialised with an estimate of the human pose supplied by the user [11] or automatically obtained through a detector [2, 20, 29] or inertial sensors [42]. Instead of optimizing mesh and skeleton parameters, recent approaches proposed to train neural networks that directly predict 3D shape and skeleton configurations given a monocular RGB video [37], multiple silhouettes [6] or a single image [18, 25, 28]. Recently, BodyNet [38] was proposed to infer the volumetric body shape through the generation of likelihoods on the 3D occupancy grid of a person from a single image.

A large body of work exists to extract human representations from multiple input views or sensors, of which some recently use deep learning to extract 3D human representations [8, 13, 21]. While they intrinsically aren't designed to deal with monocular input as proposed, multi-view methods usually yield more complete and higher precision results as soon as several viewpoints are available, a useful feature we leverage for creating the 3D HUMANS dataset.

More similar to ours are the methods that estimate projections of the human body: in [39], an encoder-decoder architecture predicts a quantized depth map of the human body while in DensePose [12] a mapping is established between the image and the 3D surface. Our method also makes predictions aligned with the input image but the combination of two complementary "visible" and "hidden" depth maps leads to the reconstruction of a full 3D volume. In [24], the authors complete the 3D point cloud built from the front facing depth map of a person in a canonical pose by estimating a second depth map of the opposite viewpoint. We instead predict both depth maps simultaneously from a single RGB image and consider a much wider range of body poses and camera views. All these methods rely on a parametric 3D model [2, 18, 20] or on training data annotated [12] or synthesised [39] using such a model. These models of humans built from thousands of scans of naked people such as the SMPL model [23] lack realism in terms of appearance. We instead propose to tackle real-world situations, modeling and estimating the detailed 3D body shape including clothes, hair and manipulated objects.

**3D human datasets.** Current approaches for human 3D pose estimation are built on deep architectures trained and evaluated on large datasets acquired in controlled environments with Motion Capture systems [1, 14, 34]. However, while the typology of human poses on these datasets captures the space of human motions very well, the visual appearance of the corresponding images is not representative of the scenarios one may find in unconstrained real-world images. There has been a recent effort to generate in-the-wild data with ground truth pose annotation [26, 32]. All these datasets provide accurate 3D annotation for a small set of body keypoints and ignore 3D surface with the exception of [20] and [41] who annotate the SMPL parameters in real-world images manually or using IMU. Although the resulting dataset can be employed to evaluate under-cloth 3D body shapes, its annotations are not detailed enough, and importantly, its size is not sufficient to train deep networks.

To compensate for the lack of large scale training data required to train CNNs, recent work has proposed to generate synthetic images of humans with associated ground truth 3D data [4, 31, 39]. In particular, the Surreal dataset [39], produced by animating and rendering the SMPL model [23] on real background images, has proven to be useful to train CNN architectures for body parts parsing and 2.5D depth prediction [39], 3D pose estimation [31, 33], or 3D shape inference [38]. However, because it is based on the SMPL model, this dataset is not realistic in terms of clothing, hair or interactions with objects and cannot be used to train architectures that target the estimation of a detailed 3D human shape. We propose to bridge this gap by leveraging multi-camera shape data capture techniques [3, 40], introducing the first large scale dataset of images showing humans in realistic scenes, i.e. wearing real clothes and manipulating real objects, dedicated to training with full 3D mesh and pose ground-truth data. Most similar to ours are the CMU Panoptic dataset [17] that focus on social interactions and the data of [45] that contains dense unstructured geometric motion data for several dressed subjects.

## 3. Methodology

In this section, we present our new non-parametric 3D human shape representation and detail the architecture that we designed to estimate such 3D shape from a single image.

### 3.1. "Mould" representation

We propose to encode the 3D shape of a person through a double 2.5D depth map representation: a "visible" depth map that depicts the elements of the surface that are directly observable in the image, and a "hidden" depth map that characterises the occluded 3D surface. These two depth

maps can be estimated and combined to reconstruct the complete human 3D shape as done when lining up the two halves of a "mould". See example in Figure 2.

Given a 3D mesh, obtained by animating a 3D human model or by reconstructing a real person from multiple views, and given a camera hypothesis, i.e. location and parameters, we define our two 2D depth maps $z_{vis}$ and $z_{hid}$ by ray-tracing. Specifically, we cast a ray from the camera origin, in the direction of each image pixel location $(u, v)$ and find the closest intersecting point on the mesh surface:

$$z_{vis}[u, v] = \min_{k \in \text{Ray}(u,v)} ||\mathbf{p}_k||_2 \quad (1)$$

for the visible map, and the furthest one for the hidden map:

$$z_{hid}[u, v] = \max_{k \in \text{Ray}(u,v)} ||\mathbf{p}_k||_2, \quad (2)$$

where 3D points $\{\mathbf{p}_i\} = \{(p_{x,i}, p_{y,i}, p_{z,i})\}$ are expressed in camera coordinate system and the L2-norm $||.||_2$ is the distance to the camera center. $\text{Ray}(u, v)$ denotes the set of points $\mathbf{p}_i$ on the ray passing through pixel $(u, v)$ obtained by hidden surface removal and visible surface determination.

To be independent from the distance of the person to the camera, we center the depth values on the center of mass of the mesh, i.e. $z_{orig} : z_{vis}[u, v]' = z_{vis}[u, v] - z_{orig}\ \forall u, v$, and similarly for $z_{hid}[u, v]$. Since they are defined with respect to the same origin, the 2 depth maps $z_{vis}[u, v]$ and $z_{hid}[u, v]$ can be readily combined in 3D space by merging their respective 3D point clouds into a global one:

$$\{\mathbf{p}_i\} = \{\mathbf{p}_i\}_{vis} \bigcup \{\mathbf{p}_i\}_{hid} \quad (3)$$

An example of such a point cloud is depicted in Figure 2, where points corresponding to $z_{vis}[u, v]$ and $z_{hid}[u, v]$ are respectively colored in red and blue. In practice, to keep the depth values within a reasonable range and estimate them more accurately, we place a flat background a distance $L$ behind the subject to define all pixels values in the depths maps in the range $[-z_{orig} \dots L]$. Points $\mathbf{p}_i$ of the point clouds are then selected as belonging to the human surface if $p_{z,i} \leq L - \epsilon$.

As in volumetric representation through voxel grid, our method also encodes 3D surfaces and point clouds of diverse sizes into a fixed size representation, making a 3D surface easier to consider as a deep network target. However, in our case, we can work at the image resolution with a much lower output dimensionality $O(N^2)$ than voxel-based volumetric representations $O(N^3)$, N being the size of the bounding box framing the human in the input image.

We numerically validated the benefit of our representation compared to a voxel grid approach by encoding a random set of 100 meshes (picked from our 3D HUMANS dataset presented in Section 4) at different resolutions and computing the 3D reconstruction error (average Chamfer distance) between ground-truth vertices and the resulting point clouds. This comparison is shown in Figure 3. The error obtained with our mould-representation decreases and converges to a minimum value that corresponds to surface details that cannot be correctly encoded even with high resolution depth maps, i.e. when some rays intersect more than twice with the human surface for particular poses. In practice, we show in Section 5 that this can be solved by employing a Poisson reconstruction to obtain a smooth 3D surface, including those areas. We can extrapolate from Figure 3 that voxel grids can reach perfect results with an infinity of voxels, but for manageable sizes, our representation allows to capture more details.

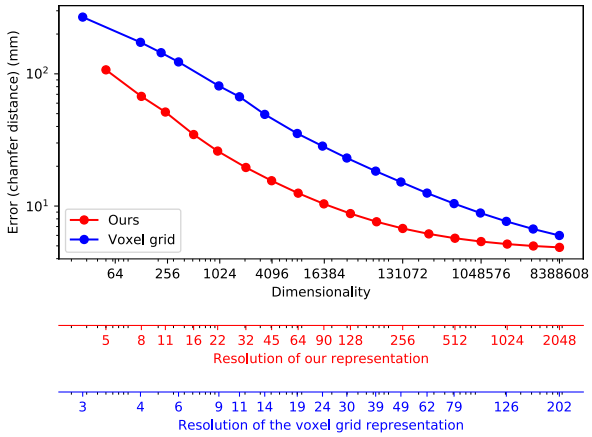



Figure 3. Reconstruction error for voxel grid and our "mould" when augmenting the dimensionality D of the representation, D=$N^3$ for voxels grid and D=$2N^2$ for ours.

### 3.2. Architecture

We formulate the 3D shape estimation problem as a pixel-wise depth prediction task for both visible and hidden surfaces. Our framework builds on the stacked hourglass network proposed by Newell et al. [27] that consists of a sequence of modules shaped like an hourglass, each taking as input the prediction from the previous module. Each of these modules has a set of convolutional and pooling layers that process features down to a low resolution and then upsample them until reaching the final output resolution. This process, combined with intermediate supervision through skip connections, implicitly captures the entire context of the image. Originally introduced for the task of 2D pose estimation and employed later for part segmentation and depth prediction [39], this network is an appropriate choice as it predicts a dense pixel-wise output while capturing spatial relationships associated with the entire human body.

We designed a 2-stack hourglass architecture that takes as input an RGB image $I$ cropped around the human and outputs the 2 depths maps $z_{vis}$ and $z_{hid}$ aligned with $I$. We

use a $\mathcal{L}_{L1}$ loss function defined on all pixels of both depth maps. The loss function to be minimized is thus the average distance between the ground truth $z_p$ and the estimation $\widehat{z_p}$:

$$\mathcal{L}_{L1} = \frac{1}{P}\sum_{p=1}^{P}|z_p - \widehat{z_p}|, \tag{4}$$

with $P$ being the number of pixels in the batch and $\widehat{z_p}$ the network output for pixel $p$, including pixels in both $z_{vis}[u,v]$ and $z_{hid}[u,v]$ maps.

We also experimented with an $\mathcal{L}_{L2}$ loss but found that it overly penalizes outliers, i.e. pixels incorrectly assigned to background and vice versa, and therefore focuses only on that task. By using the $\mathcal{L}_{L1}$ norm, we force the network to not only segment the image correctly, i.e discriminate the subject from the background, but also provide an accurate estimation of the depth at each pixel.

### 3.3. Adversarial training

As observed with other non-parametric methods [38] but also with approaches relying on a model [18], our network can sometimes produce implausible shapes that do not look human, especially when a limb is entirely occluded by other parts of the body. To improve the accuracy and the "humanness" of our prediction, we follow an adversarial training procedure inspired by the Generative Adversarial Networks (GAN) [9]. Our fully derivable depth-based model allows us to efficiently incorporate a discriminator in an adversarial fashion, i.e., the goal for the discriminator will be to correctly identify ground truth depth maps from generated ones. On the other hand, the generator objective will be two-fold: fitting the training set distribution through the minimization of the $\mathcal{L}_{L1}$ loss (Equation 4) and tricking the discriminator into classifying the generated depth maps as ground truth depth maps through the minimization of the $\mathcal{L}_{GAN}$ loss:

$$\mathcal{L}_{GAN}(G,D) = E_{I,z}[\log D(I,z)] + E_I[\log(1 - D(I,G(I))]. \tag{5}$$

Our discriminator $D$ will be trained to maximize the $\mathcal{L}_{GAN}$ loss by estimating 1 when provided with ground-truth depth maps $z$ and estimating 0 when provided with generated depth maps $G(I)$. In order to weigh the contribution of each loss, we will use a factor $\lambda$, our full objective being modeled as a minimax game:

$$(G^*, D^*) = \arg\min_G \max_D (\mathcal{L}_{GAN}(G,D) + \lambda \mathcal{L}_{L1}(G)). \tag{6}$$

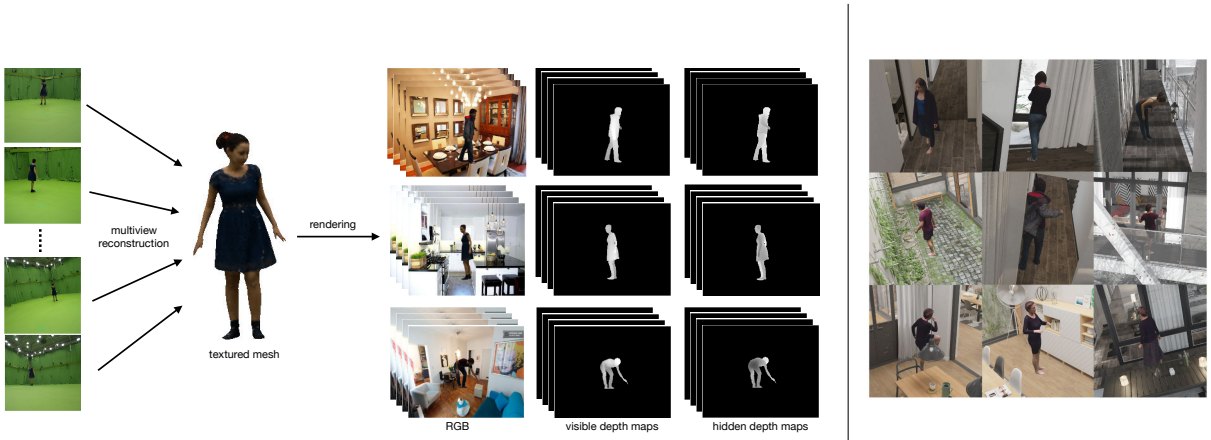
The $\mathcal{L}_{L1}$ loss will be used to learn the training set distribution by retrieving the low-frequency coefficients while the $\mathcal{L}_{GAN}$ loss will entice the generator into predicting realistic and precise depth maps. It is important to note that the discriminator is only used to guide the generator during the learning. The discriminator is not used at test time.

The architecture employed as our discriminator is a 4 stack CNN. Each stack is composed of a convolutional layer (kernel size 3, stride 1), a group normalization layer (32 groups), a ReLu activation function and a MaxPool 2x2 operation. There are 64 channels for the first convolution and the number of channels is multiplied by 2 at each stack until reaching 512 for the 4th and last stack convolution. We then connect our 8x8x512 ultimate feature map with 2 fully-connected layers of size 1024 and 512 neurons and then our final output neuron on which we apply a binary cross entropy loss. We jointly trained our generator and discriminator on 50,000 images for 40 epochs. Training is performed on batches of size 8 with the Adam optimizer. Given our small training batch size, we found the use of group norm [44] to be a great alternative to batch norm that was producing training instabilities. The learning rate is kept constant at 1e-4 during the first 20 epochs and is then decreased linearly to zero during the following 20 epochs. In practice, since our $\mathcal{L}_{L1}$ loss is much smaller than the $\mathcal{L}_{GAN}$ loss, we multiply the $\mathcal{L}_{L1}$ loss by a $\lambda$ factor equal to 1e4. With this adversarial training, we observed that the results are sharper and more realistic. In cases of deformed or missing limb, e.g. the legs in Figure 7 right, the use of a discriminator forces the generator to produce a better prediction.

## 4. Dataset generation

We introduce 3D HUMANS (HUman Motion, Activities aNd Shape), a realistic large-scale dataset of humans in action with ground-truth 3D data (shape and pose). It consists of semi-synthetic videos with 3D pose and 3D body shape annotations, as well as 3D detailed surface including cloths and manipulated objects. First, we captured 3D meshes of humans in real-life situations using a multi-camera platform. We then rendered these models on real-world background scenes. See examples in Figure 4a.

**Capture.** We employed a state of the art 3D capture equipment with 68 color cameras to produce highly detailed shape and appearance information with 3D textured meshes. The meshes are reconstructed frame by frame independently. They are not temporally aligned and do not share any common topology. We divided the capture into 2 different subsets: in the first one, 13 subjects (6 male and 7 female) were captured with 4 different types of garments (bathing suit/tight clothing, short/skirt/dress, wide cloths and jacket) while performing basic movements e.g., walk, run, bend, squat, knees-up, spinning. In the second subset, 6 subjects, 4 male and 2 female, were captured while performing 4 different activities (talking on the phone, taking pictures, cleaning a window, mopping the floor) in 2 different ways: standing/sitting for talking on the phone, standing/kneeling for taking picture, etc. More than 150k meshes were reconstructed. The dataset was collected at Inria from consenting and informed participants.




(a) (b)

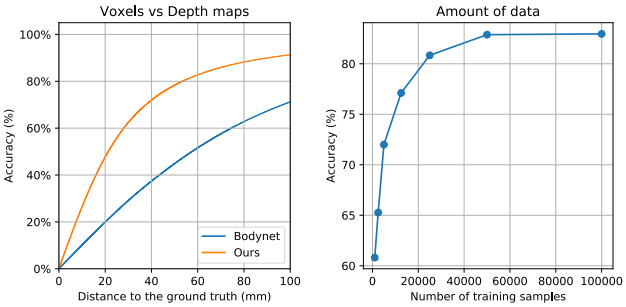
Figure 4. Data generation. (a) We captured 3D meshes of humans, wearing real clothes, moving and manipulating objects using a multi-camera platform. We then rendered these models on real-world background scenes and computed ground-truth visible and hidden depths maps. (b) We also generated a test set by rendering our meshes on realistic 3D environments.

**Rendering.** We rendered all our videos at a 320 x 240 resolution using a camera of sensor size 32mm and focal length 60mm. Our videos are 100 frames in length and start with the subject at the center of the frame. For the first frame of the sequence, the subject is positioned at a distance of 8 meters of the camera, with a standard deviation of 1 meter. We used the images of the LSUN dataset [46] for background.

**Annotations.** We augment our dataset with ground-truth SMPL pose and body parameters. To do so, we use the Human3.6M [14] environment as a "virtual MoCap room": we render the 3D meshes for which we want to estimate the 3D pose within that environment, generate 4 views using camera parameters and background images from the dataset and estimate the 2D/3D poses by running LCR-Net++, an off-the-shelf 3D pose detector particularly efficient on Human3.6M. An optimum 3D pose is then computed using multi-view 3D reconstruction and used as initialization to fit the SMPL model, estimating pose and shape parameters that better match each mesh. The SMPL model is fitted to the point clouds both for naked and dressed bodies. Keeping the body parameters fixed (obtained from fits in minimal clothing) resulted in a lower performance of the baseline when evaluated against ground truth dressed bodies.

## 5. Experiments

We analyse quantitatively and compare our approach to the state-of-the-art on two datasets. First, the SURREAL dataset [39], a synthetic dataset obtained by animating textured human models using MoCap data and rendering them on real background images, and our 3D HUMANS dataset introduced in this paper. While SURREAL covers a wider range of movements since it has been rendered using thousands of sequences from [1], our data better covers shape details such as hair and clothing. In the following experiments, both training and test images are tightly cropped around the person using subjects segmentation. The smallest dimension of the image is extended to obtain a square image that is then resized to 256x256 pixels to serve as input for our network. Performance is computed on both 128x128 output depth maps as the distance between each ground truth foreground pixel and its corresponding pixel in the predicted depth map. Background depth $L$ is set at 1.5m.

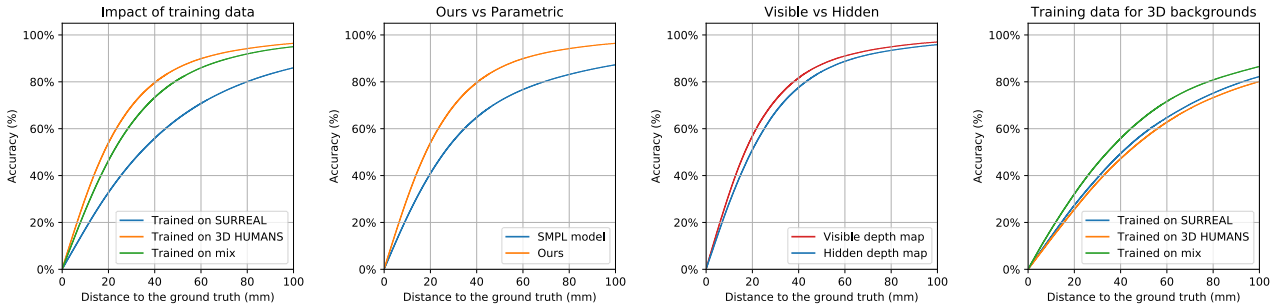



(a) (b)

Figure 5. Comparaison with state-of-the-art on the SURREAL dataset: (a) we first compare against the BodyNet [38] baseline. (b) We analyse the impact of varying the size of the training set on performance on our new 3D HUMANS dataset.

### 5.1. SURREAL

Recent methods [38, 39] evaluate their performance on this dataset. First, we evaluated the performance of our architecture when estimating quantized depth values (19+1 for background) through classification as in [39] and our proposed regression method: with a maximum distance to groundtruth of 30mm, the quantity of pixels with a correct depth estimation increase by 5% when using regression instead of classification. Then, we compare in Figure 5a our




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

Figure 6. Evaluation on our 3D HUMANS dataset: we first analyze the influence of the training data on performance (a). Then, we compare against the SMPL baseline (b). We compare the performance on visible and hidden depth map separately (c). Finally, we analyse the training data on a dataset rendered in realistic backgrounds and observe that SURREAL data is important for generalisation (d).

performance against the recent BodyNet voxel grid-based architecture from [38] who also reported numerical performance on SURREAL. Although good 3D performances are reported in the paper, we can see that when evaluating in the image domain, i.e., comparing depth maps, the performance of BodyNet drops. Our method makes 3D estimations aligned with the image and better recover details, outperforming BodyNet quite substantially.

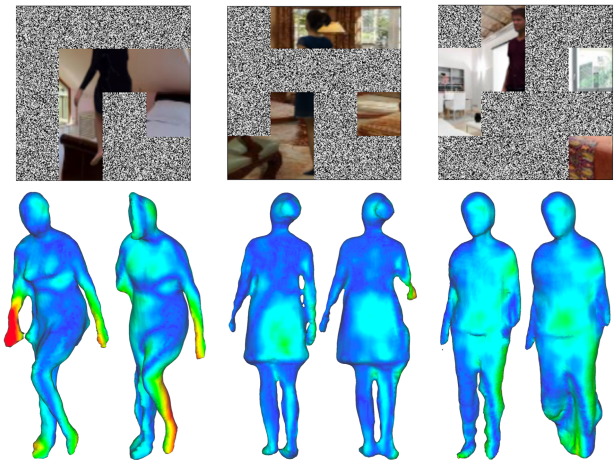


Figure 7. Performance on 3D-HUMANS dataset in presence of severe occlusions on three frames: (top) input images, (left) with GAN, (right) without GAN. Errors above 15cm are shown in red. The GAN helps increase the "humanness" of the predictions.

## 5.2. 3D HUMANS

We consider 14 subjects (8 male, 6 female) for training and the remaining 5 subjects (2 male, 3 female) for test. An interesting aspect of synthetic datasets is that they offer an almost unlimited amount of training data. In our case, the data generation relies on a capture process with a non-negligible acquisition effort. It is therefore interesting to analyze how adding more training data impacts the performance. Our results in Figure 5b show that training our architecture on 50,000 images is sufficient and that using more training images does not improve much the performance. The appearance of our images being quite different from SURREAL data, we first compare the performance of our method when considering different training strategies: training on SURREAL, training on 3D HUMANS, or training on a mix of both datasets. In Figure 6a, we can see that the best performances are obtained when SURREAL images are not used. The appearance of the images is too different and our architecture cannot recover details such as clothes or hair when trained on data obtained by rendering the SMPL model. This is verified by the result depicted in Figure 6b: we outperform, by a large margin, a baseline obtained by fitting the SMPL model on the ground truth meshes, effectively acting as an upper bound for all methods estimating SMPL meshes [2, 18]. It shows the inefficiency of these methods to estimate clothed body shape since clothes are not included in the SMPL model.

Finally, we analyse in Figure 6c how much our performance varies between front and back depth maps. As expected, we better estimate the visible depth map, but our hidden depth maps are usually acceptable. See examples in Figure 7 and Figure 8. The quality of the 3D reconstructions is remarkable given the low dimensionality of the input. Main failures occur when a limb is completely occluded. In such cases, the network can create non-human shapes. We proposed to tackle this issue by considering an adversarial training that we analyse in the next section. We note a higher performance on 3D HUMANS than on SURREAL. We attribute that to several factors including the higher pose variability in SURREAL (some subjects are in horizontal position) and the absence of lighting in 3D HUMANS. We also analyzed the results on different subsets of the evaluation set @50mm and obtained with/without clothing: 83.30% and 85.43% respectively and with/without object: 79.11% and 84.55% respectively, confirming the nuisance introduced by these elements.

Figure 8. Generalisation to previously unobserved data. We apply our pipeline to images with 3D realistically rendered backgrounds (left), and with 3 real-world images from the LSP dataset (right). These poses, in particular the baseball player, have not been seen at training time but our model still generalizes well.

### 5.3. GAN

Severe occlusions (self- or by other elements of the scene) are a limitation of our model that we address with adversarial training. We carried out a dedicated experiment where we artificially generated such occlusions in train/test images to quantify improvements. We obtain a 7% chamfer distance error drop with adversarial training and a clear qualitative improvement which we illustrate in Figure 7. We highlight the differences by showing an error heat-map over a Poisson reconstruction of the point cloud for better visualization. The quantitative gain is limited due to the network sometimes hallucinating plausible limbs far from groundtruth (red hand in the left Figure 7), resulting in higher error than a network without GAN that does not estimate any limb at all. This is because the metric does not evaluate the overall plausibility of the produced estimation.

### 5.4. Generalisation

In order to quantitatively measure its generalisation capability, we have evaluated our network on an additional dataset: instead of static background images, we have rendered the meshes in realistic 3D environments obtained on the internet (examples in Figure 4b). The results (Figure 6d) show that a mix training on both SURREAL and 3D HUMANS is ideal for generalisation. We suspect that jointly rendering the subject and the 3D background at the same time creates a more realistic image where the subject is more complicated to segment, hence the need for more variability in the training data. We also generated qualitative results for LSP images [16], depicted in Figure 8, and for the DeepFashion dataset [22], shown in Figure 9 where we compare our approach with HMR [18] and BodyNet [38]. We can observe that our approach captures more details, including hair, shirt and the belly of the pregnant woman (up), hair, skirt and body pose (middle) and dress (bottom).

Figure 9. Comparison between HMR [18] (left), Bodynet [38] (middle) and our method (right). Unlike [18, 38], we do not train on in-the-wild images but estimate 3D shapes of clothed subjects.

## 6. Conclusion

We have proposed a new non-parametric approach to encode the 3D shape of a person through a double 2.5D depth map representation: a "visible" depth map depicts the elements of the surface that are directly observable in the image while a "hidden" depth map characterises the occluded 3D surface. We have designed an architecture that takes as input a single image and simultaneously produces an estimate for both depth maps resulting, once combined, in a point cloud of the full 3D surface. Our method can recover detailed surfaces while keeping the output to a reasonable size. This makes the learning stage more efficient. Our architecture can also efficiently incorporate a discriminator in an adversarial fashion to improve the accuracy and "humanness" of the output. To train and evaluate our network, we have captured a large-scale dataset of textured 3D meshes that we rendered on real background images. This dataset will be extended and released to spur further research.

**Acknowledgements.** We thank Pau De Jorge and Jinlong Yang for their help in capturing the data employed in this paper. The dataset was acquired using the Kinovis[2] platform. This work was supported in part by ERC advanced grant Allegro.

[2]https://kinovis.inria.fr

## References


[1] CMU motion capture dataset. http://mocap.cs.cmu.edu. the database was created with funding from nsf eia-0196217. Technical report. 3, 6

[2] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 1, 3, 7

[3] D. Casas, M. Volino, J. Collomosse, and A. Hilton. 4D Video Textures for Interactive Character Appearance. *Computer Graphics Forum*, 2014. 3

[4] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3D pose estimation. In *3DV*, 2016. 1, 3

[5] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*, 2016. 2

[6] E. Dibra, H. Jain, A. C. Öztireli, R. Ziegler, and M. H. Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *3DV*, 2016. 3

[7] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017. 2

[8] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton. Volumetric performance capture from minimal camera viewpoints. In *ECCV*, 2018. 3

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*. 2014. 5

[10] T. Groueix, M. Fisher, V. G. Kim, B. Russell, and M. Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR*, 2018. 2

[11] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *ICCV*, 2009. 3

[12] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. *arXiv*, 2018. 3

[13] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, C. Ma, L. Luo, and H. Li. Deep volumetric video from very sparse multi-view performance capture. In *ECCV*, 2018. 3

[14] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. PAMI*, 2014. 3, 6

[15] L. Jiang, S. Shi, X. Qi, and J. Jia. GAL: Geometric adversarial loss for single-view 3D-object reconstruction. In *ECCV*, 2018. 2

[16] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 8

[17] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 3

[18] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 3, 5, 7, 8

[19] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3):29:1–29:13, July 2013. 2

[20] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 3

[21] V. Leroy, J.-S. Franco, and E. Boyer. Shape Reconstruction Using Volume Sweeping and Learned Photoconsistency. In *ECCV*, 2018. 3

[22] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 8

[23] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics*, 2015. 3

[24] N. Lunscher and J. Zelek. Deep learning whole body point cloud scans from a single depth map. In *CVPRW*, 2018. 3

[25] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017. 3

[26] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt. Single-shot multi-person 3D pose estimation from monocular rgb. In *3DV*, sep 2018. 3

[27] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 4, 10

[28] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018. 3

[29] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. *General Automatic Human Shape and Motion Capture Using Volumetric Contour Cues*. 2016. 3

[30] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3D pose estimation in the wild. In *NIPS*, 2016. 1

[31] G. Rogez and C. Schmid. Image-based synthesis for deep 3D human pose estimation. *IJCV*, 126(9):993–1008, 2018. 3

[32] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net: Localization-Classification-Regression for Human Pose. In *CVPR*, 2017. 3

[33] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE Trans. PAMI*, 2019. 1, 3

[34] L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4–27, 2010. 3

[35] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3D models from single images with a convolutional network. In *ECCV*, 2016. 3

[36] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In *ICCV*, 2017. 2

[37] H. Tung, H. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 3

[38] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 1, 3, 5, 6, 7, 8

[39] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from Synthetic Humans. In *CVPR*, 2017. 2, 3, 4, 6

[40] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.*, 27(3):97:1–97:9, Aug. 2008. 3

[41] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 3

[42] T. von Marcard, G. Pons-Moll, and B. Rosenhahn. Human pose estimation from video and imus. *IEEE Trans. PAMI*, 38(8):1533–1547, 2016. 3

[43] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li. Dense human body correspondences using convolutional networks. In *CVPR*, 2016. 1

[44] Y. Wu and K. He. Group normalization. In *ECCV*, 2018. 5, 10

[45] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of Human Body Shape in Motion with Wide Clothing. In *ECCV*, 2016. 3

[46] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6

# Annex

## 7. Architecture details

### 7.1. Generator

The main difference between our generator architecture and the stacked hourglass by Newell et al. [27] is the output dimension. Newell et al. estimate a 64x64 resolution heatmap for each body joint. In our case, we estimate 2 depth maps and aim at a higher 128x128 resolution. Our hourglass output dimension is 128x128x2. Because of this difference in output resolution, we apply the following modifications to the stacked hourglass [27] architecture: We do not use a maxpooling operation after layer1, we increase the depth of the hourglasses from 4 to 5 skipped connections, we project the hourglass result on 2 channels (one for each depth map). Also, we use 2 stacked hourglasses and we replace batch normalization by group normalization [44] that performs better on small training batches. See architecture details in Table 1.

| **Layer** | **Layer type** | **Output shape** |
|---|---|---|
| Input | Input | 256x256x3 |
| Conv1 | Conv 7x7 stride=2, GroupNorm, Relu | 128x128x64 |
| Layer1 | Residual module expanded | 128x128x128 |
| Layer2 | Residual module expanded | 128x128x256 |
| Layer3 | Residual module | 128x128x256 |
| Hg1 | Hourglass, skipped connections = 5 | 128x128x2 |
| Hg2 | Hourglass, skipped connections = 5 | 128x128x2 |

Table 1. Generator architecture.

### 7.2. Discriminator

For our discriminator, we employed a 4 stacks CNN. It takes as input a set of 2 depth maps at resolution 128x128 and outputs a scalar: close to 1.0 if it believes they are sampled from the ground truth depth maps and close to 0 if it believes they have been generated by the generator. See Table 2 for details.

| **Layer** | **Layer type** | **Output shape** |
|---|---|---|
| Input | Input | 128x128x2 |
| Conv1 | Conv 3x3 stride=1, GroupNorm, Relu | 128x128x64 |
| MP1 | MaxPool 2x2 | 64x64x64 |
| Conv2 | Conv 3x3 stride=1, GroupNorm, Relu | 64x64x128 |
| MP2 | MaxPool 2x2 | 32x32x128 |
| Conv3 | Conv 3x3 stride=1, GroupNorm, Relu | 32x32x256 |
| MP3 | MaxPool 2x2 | 16x16x256 |
| Conv4 | Conv 3x3 stride=1, GroupNorm, Relu | 16x16x512 |
| MP4 | MaxPool 2x2 | 8x8x512 |
| FC1 | Fully connected layer | 1024 |
| FC2 | Fully connected layer | 512 |
| FC3 | Fully connected layer | 1 |

Table 2. Discriminator architecture.

$\therefore$

## A.8 VOLUME SWEEPING : LEARNING PHOTOCONSISTENCY FOR MULTI-VIEW SHAPE RECONSTRUCTION

Vincent Leroy, Jean-Sébastien Franco, Edmond Boyer. International Journal of Computer Vision, 2020

**Noname manuscript No.**
(will be inserted by the editor)

# Volume Sweeping: Learning Photoconsistency for Multi-View Shape Reconstruction

Vincent Leroy[1,2] · Jean-Sébastien Franco[1] · Edmond Boyer[1]

the date of receipt and acceptance should be inserted later

**Abstract** We propose a full study and methodology for multi-view stereo reconstruction with performance capture data. Multi-view 3D reconstruction has largely been studied with general, high resolution and high texture content inputs, where classic low-level feature extraction and matching are generally successful. However in performance capture scenarios, texture content is limited by wider angle shots resulting in smaller subject projection areas, and intrinsically low image content of casual clothing. We present a dedicated pipeline, based on a per-camera depth map sweeping strategy, analyzing in particular how recent deep network advances allow to replace classic multi-view photoconsistency functions with one that is learned. We show that learning based on a volumetric receptive field around a 3D depth candidate improves over using per-view 2D windows, giving the photoconsistency inference more visibility over local 3D correlations in viewpoint color aggregation. Despite being trained on a standard dataset of scanned static objects, the proposed method is shown to generalize and significantly outperform existing approaches on performance capture data, while achieving competitive results on recent benchmarks.

**Keywords** Multi View Stereo Reconstruction · Learned Photoconsistency · Performance Capture · Volume Sweeping

Vincent Leroy
E-mail: vincent.leroy@naverlabs.com
Jean-Sébastien Franco
E-mail: jean-sebastien.franco@inria.fr
Edmond Boyer
E-mail: edmond.boyer@inria.fr
[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK 38000 Grenoble, France * Institute of Engineering Univ. Grenoble Alpes
[2] NAVER LABS Europe, 6 chemin de Maupertuis, Meylan, 38240, France

## 1 Introduction

In this paper, we examine the problem of multi-view shape reconstruction of production-realistic performance capture sequences. Such sequences may contain arbitrary casual clothing and motions, and have specific capture set assumptions due to the particular lighting and camera positioning of these setups. Multi-view 3D reconstruction is a popular and mature field, with numerous applications involving the recording and replay of captured 3D scenes, such as 3D content creation for broadcast and mobile applications, or the increasingly popular virtual and augmented reality applications with 3D user avatars. An essential and still improvable aspect in this matter, in particular with performance capture setups, is the fidelity and quality of the recovered shapes, our goal in this work.

Multi-view stereo (MVS) based methods have attained a good level of quality with pipelines that typically comprise feature extraction, matching stages and 3D shape extraction. Interestingly, very recent works have re-examined stereo and MVS by introducing features and similarity functions automatically inferred using deep learning. The main promise of this methodological shift is to include better data-driven priors, either in 2D [1,2,3,4] as improvement over classic 2D features, or in 3D to account for relative view placement and local or global shape priors [5,6,7]. These novel MVS methods have been shown to outperform classic learning-free methods on static scene benchmarks [8].

Our main goal is to examine whether these data-driven improvements transfer to the more complex case of live performance capture, where a diverse set of additional difficulties arise with respect to typical MVS setups. Typical challenges for these capture situations include smaller visual projection areas of objects of in-

**Fig. 1** Challenging scene captured with a passive RGB multi-camera setup [9]. (*left*) one input image, (*center*) reconstructions obtained with classical 2D features [10], (*right*) proposed solution. Our results validate the key improvement of a CNN-learned disparity to MVS for performance capture scenarios. Results particularly improve in noisy, very low contrast and low textured regions such as the arm, the leg or even the black skirt folds, which can be better seen in a brightened version of the picture in Figure 17.

terest, due to wider necessary fields of view for capturing motion; occlusion and self-occlusion of several subjects interacting together; lack of texture content typical of real-life subject appearance and clothing; or motion blur with fast moving subjects such as sport action scenes (see Figure 14). To the best of our knowledge, existing learning-based MVS schemes report results on static datasets such as DTU [11] or ShapeNet [12] but have not yet been demonstrated on performance capture data with the aforementioned typical issues.

We present a detailed framework for this purpose, which casts the problem as a fusion of per view depth maps as inspired by recent fusion methods [13], each depth map extracted using a learned multi-view photoconsistency function. Our approach performs multi-view matching within local volumetric units of inference. Contrary to previous methods, our volumetric unit is defined in a given view's own reference, so as to capture camera inherent 3D dependencies, specifically for the purpose of per-view decision. Instead of inferring occupancies, we infer disparity scores to ease training and to focus the method more on photometric configurations than local shape patterns. We sweep viewing rays with this volumetric receptive field, a process we coin *volume sweeping*, and embed the algorithm in a multi-view depth-map extraction and fusion pipeline followed by a geometric surface reconstruction. With this strategy, we validate that CNN-based MVS outperforms classical MVS approaches in performance capture scenarios. In particular, we obtain high precision geometric results on complex sequences, outperforming both existing CNN-based and classic non-learning methods on a large set of capture datasets. These diverse results are obtained using only a DTU subset as training data, which evidences the generalization capabilities of our network.

This article is an extended version of [14] that provides a complete and self-contained description of the proposed method, with more details about the pipeline from [10] along with the detailed volume sweeping and surface extraction algorithms. Several supplementary experiments were performed to give more insights on the contribution and study the influence of the parameters. We finally challenged the generalization properties of our network on multiple dataset that were not seen during training with competitive results compared to both hand-crafted and learned state of the art.

## 2 Related Work

**Multi-view stereo reconstruction** is an active and longstanding vision problem [15]. Stereo and MVS-

based approaches are increasingly being used for high fidelity capture applications [16,17,18,19,11,20,21], possibly complementing other strategies such as depth-based reconstruction [13,22,23,24] by addressing shortcomings that include limited range, sensitivity to high contrast lighting, and interference when increasing the number of viewpoints. While considering various shape representations, for instance point clouds [16], fused depth maps [25], meshes [26,27], or volumetric discretizations [28,29,30], most MVS methods infer 3D shape information by relying on the photoconsistency principle that rays observing the same scene point should convey similar photometric information.

In it simplest form, such similarity can be measured by considering projected color variances among views, as used in early works [28] with limited robustness. In stereo and short baseline situations, simple normalized forms of 2D window correlation are sufficient to characterize similarity under simple lighting and constrast changes, using *e.g.* ZNCC, SSD, SHD. For broader geometric and photometric resilience, various features based on scale-invariant gradient characterizations [31,32,33] have been designed, some specialized for the dense matching required for the MVS problem [34]. More recently, image features have been successfully applied to performance capture sequences in *e.g.* [20,10]. Generally, MVS methods characterize photoconsistency either with a symmetric, viewpoint agnostic, combination of all pairwise similarities [35], or with a per image depth map determination through sweeping strategies [36,25]. The latter sweeping approaches have the advantage of simplifying the scene parametrization of occlusions [37,38], which we leverage for our approach and show to yield a robustness advantage over other strategies in our experiments.

While classic MVS approaches have been generally successful, recent works aimed at learning stereo photoconsistency have underlined that additional priors and more subtle variability co-dependencies are still discoverable in real world data. Several works leverage this by learning how to match 2D patch pairs for short baseline stereo, letting deep networks infer what features are relevant [1,2,3,4]. More recent works extend this principle to short baseline MVS, with symmetric combination of 2D learned features [39], or wide baseline sparse capture scenarios [40,41].

Most of these methods however use a 2D receptive field for stereo matching. The intuition that volumetric 3D receptive fields may be more informative and ease CNN inference and has been explored by some recent works [5,6,7,8], an assertion that the presented approach further verifies. While casting correlations in 3D as well, our approach proposes several key differences. Contrary to the latter, our volumetric receptive field is projective in the camera coordinate frame, similar to some binocular stereo [42] or image-based rendering [43] works. This allows for sweeping along viewing rays, which was proven to be a robust search strategy for binocular stereo plane sweeping [38]. It also enables a per frame approach, with depth estimations, that appears to be more flexible than a global reasoning over all frames. This scheme also avoids decorrelating camera resolution and 3D receptive field resolution, as with *e.g.* voxels, the volumetric receptive field being defined as a backprojection along pixel rays. Additionally, this volumetric receptive field learns local pairwise correlations, a lower level and easier task than learning occupancy grid patterns. Our evaluation substantiates the aforementioned robustness benefits on a number of qualitative (7.3) and quantitative experiments (7.2) with challenging dynamic capture datasets, showing in particular the improvements over 2D receptive fields (7.1).

## 3 Method Overview

Our main objective is to study multi-view photoconsistency within the context of multi-view stereo reconstructions. We consider for that purpose the reconstruction framework, largely adopted over the last decade, that consists in first estimating per camera depth maps, followed by depth fusion and surface extraction. This framework allows to reason at the pixel level, enabling therefore each camera to provide local details on the observed surface with local estimations. This is in contrast with global strategies that consider photoconsistency at the shape level, with for instance voxels as in [6]. Comparisons between strategies are provided in the experiment section (see section 7.2).

Regarding depth map estimation, we propose to replace the traditional handcrafted photoconsistency measures used to estimate depths with a learned version. This version is based on CNNs and exploits their ability to learn local photometric configurations near surfaces observed from multiple viewpoints. As depicted in Figure 2, our approach takes as input a set of calibrated images and outputs a 3D mesh obtained by fusing depth maps. Depths along pixel viewing rays are obtained using a volume sweeping strategy that samples multi-view photoconsistency along rays and identifies the maxima. For a depth point candidate along a viewing ray, the photoconsistency is estimated using a discretized 3D volumetric projective grid centered on that point. In such a 3D grid, color inputs from the primary camera are paired with color inputs from another camera at each volume element of the grid around the depth point candidate. For a given depth candidate, we

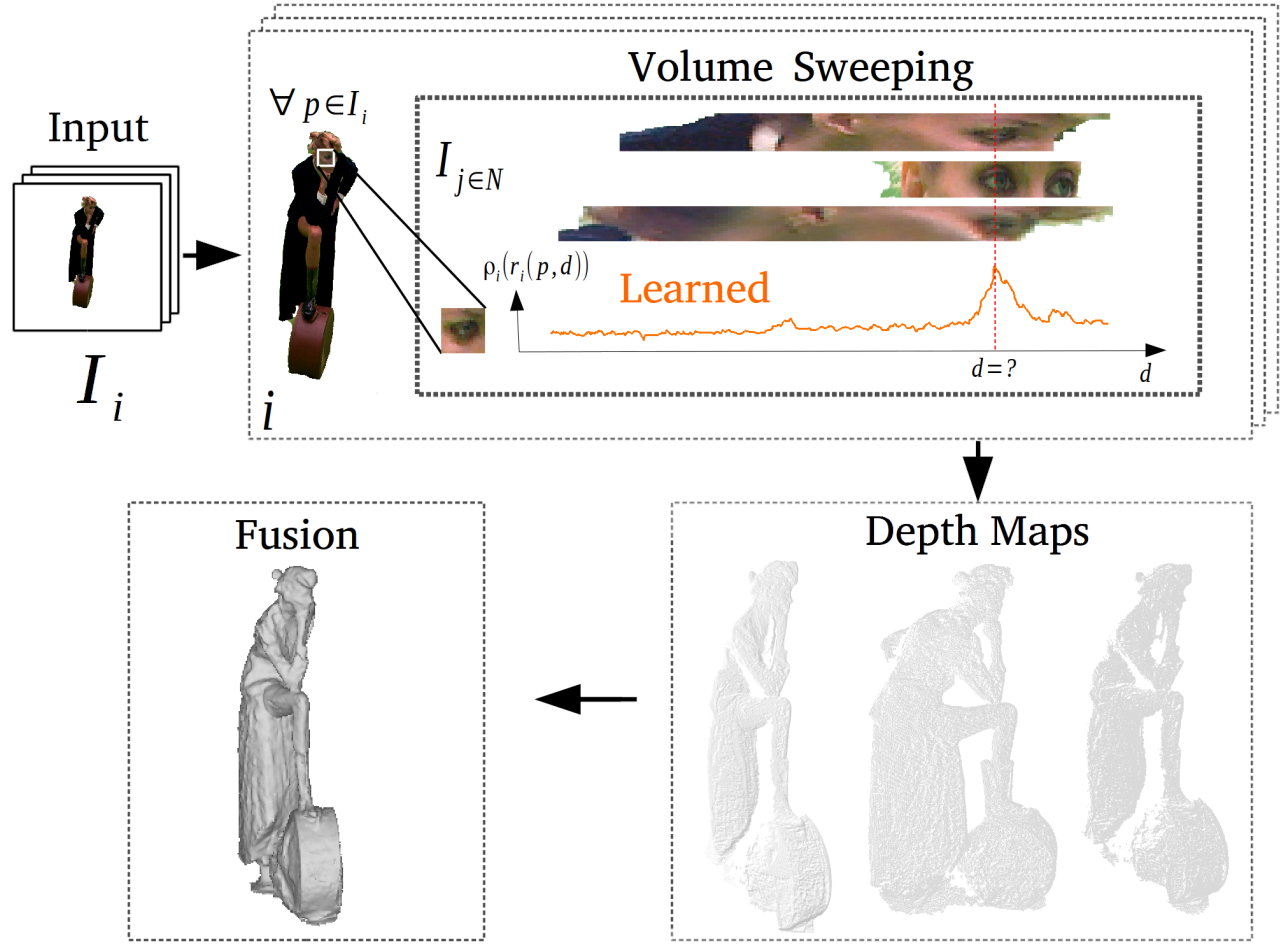



**Fig. 2** Method overview. Depth maps, for all input image $I_i$, are obtained by maximizing, along viewing lines, a learned function that measures photoconsistency at a given depth $d$ along the viewing line of a given pixel $p$. Depth maps are then fused into an implicit form from which the zero set surface is extracted.

collect all such paired color volume grids for every other camera than the primary. A trained CNN is used to recognize the photoconsistent configurations given pairs of color samples within the 3D grid. The key aspects of this strategy are:

- The per camera approach, which, by construction, samples the photoconsistency at a given location as captured and thus enables more local details to be revealed compared to a global approach, as shown in Figure 17.
- The 3D receptive field for the photoconsistency evaluation, which resolves some 2D projection ambiguities that hindered 2D based strategies.
- The learning based strategy using a convolutional neural network, which outperforms traditional photometric features when evaluating the photoconsistency in dynamic captured scenes, as demonstrated by our experiments.

The following sections focus on our main contributions, namely the 3D volume sampling in section 4.1 and the learning based approach in sections 4.2,4.3 for the photoconsistency evaluation. We then describe in section 5 our depth map evaluation procedure, derived from a *winners-take-all* strategy suitable to our capture scenario. These depth maps are then fused into an implicit form, from which, without loss of generality, we extract the zero-level set using the surface extraction technique described in section 6.

## 4 Learning Photoconsistency

Our reconstruction approach takes as input $N$ images $\{I_i\}_{i=1}^N$, along with their projection operators $\{\pi_i\}_{i=1}^N$, and computes depth maps, for the input images, which are subsequently fused into a 3D implicit form. This section explains how these maps are estimated.

Given a pixel $p$ in an input image $I_i$, the problem is therefore to find the depth $d$ at which its viewing ray intersects the observed surface. The point along the ray of pixel $p$ at depth $d$ is noted $r_i(p,d)$. Our approach searches along viewing rays using a likelihood function for a point to be on the surface given the input color pairs in the evaluation volume. In contrast to traditional methods that consider handcrafted photoconsistency measures, we learn this function from multiview datasets with ground-truth surfaces. To this purpose we build a convolutional neural network which, given a reference camera $i$ and a query point $x \in \mathbb{R}^3$, maps a local volume of color pair samples around $x$ to a scalar photoconsistency score $\rho_i(x) \in [0..1]$. The photoconsistency score accounts in practice for color information from camera $i$ at native resolution, and for other camera colors in addition to their relative orientations as implicitly encoded in the volume color pair construction. These important features allow our method to adapt to specific ray incidences. Its voluntarily asymmetric nature also allows subsequent inferences to automatically build visibility decisions, *e.g.* deciding for occlusion when the primary camera $i$'s color is not confirmed by other view's colors. This would not have been possible with a symmetric photoconsistency function such as [39].

We thus cast the photoconsistency estimation as a binary classification problem from these color pairs around the location $x$, with respect to the reference image $I_i$ and the other images. In the following, we first provide details about the 3D sampling regions before describing the CNN architecture used for the classification and its training. We then explain the volume sweeping strategy that is subsequently applied to find depths along rays.

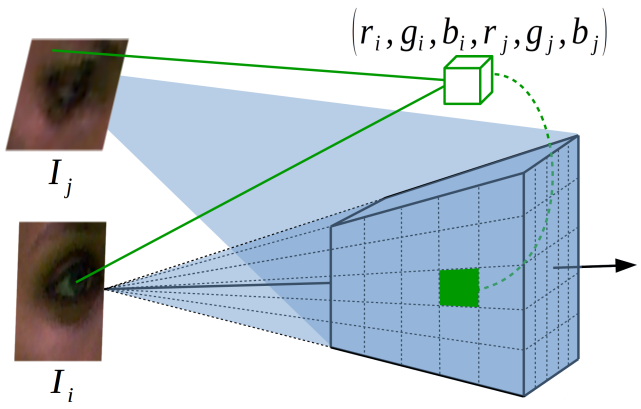



**Fig. 3** The 3D volume used to estimate photoconsistency along rays from the reference image $I_i$. $k^3$ samples within the volume are regularly distributed along viewing rays and contain color pairs as back-projected from images $I_i$ and $I_j$. At a given depth along a ray from $I_i$ any image $I_j \neq i$ can define such a pairwise comparison volume.

## 4.1 Volume Sampling

In order to estimate photoconsistency along a viewing ray, a 3D sampling region is moved along that ray at regular distances. Within this region, pairs of colors backprojected from the images are sampled. Each pair contains a color from the reference image $I_i$ and its corresponding color in another image $I_j$. Samples within the 3D region are taken at regular depths along viewing rays in the reference image (see Figure 3). The corresponding volume is a truncated pyramid that projects onto a 2D region of constant and given pixel dimension in the reference image. This allows the 3D sampling to adapt to the camera properties, *e.g.* pixel resolution and focal length.

More precisely, we denote $r_i(p,d)$ the 3D location at depth $d$ along the viewing line back-projected from pixel $p$ in the reference image $I_i$. The $k^3$ input sample grid used to compare pairs of colors from images $\{i,j\}_{j\neq i}$ is then the set of back-projected rays in a $k^2$ window centered on $p$, regularly sampled from depth $d - k\lambda/2$ to $d + k\lambda/2$ around $r_i(p,d)$, with $\lambda$ chosen such that spacing in the depth direction is equal to the inter-pixel distance from the reference camera at that depth. Every sample contains the reference color of the originating pixel in image $I_i$ and the color of the point projected in image $I_j$.

Volume sampling is always performed with the same orientation and ordering with respect to the reference camera. Convolutions are thus consistently oriented with respect to the camera depth direction.

*Volume Size* In practice, we choose $k = 8$. Our strategy is to learn pairwise photoconsistent configurations along rays. This way, decisions for the surface presence are conditioned to the observation viewpoints, which implicitly enforce visibility rules since only one 3D point per ray can be detected. This is in contrast to more global strategies where such per viewpoint visibility is less easy to impose, as with regular voxel grids, *e.g.* [6] with $32^3$ or $64^3$ grids. In addition, by considering the surface detection problem alone, and letting the subsequent step of fusion integrate depth in a robust and consistent way, we simplify the problem and require little spatial coherence, hence allowing for small grids. We provide a more detailed study of the performances of the classifiers with various depth values in section 7.1.

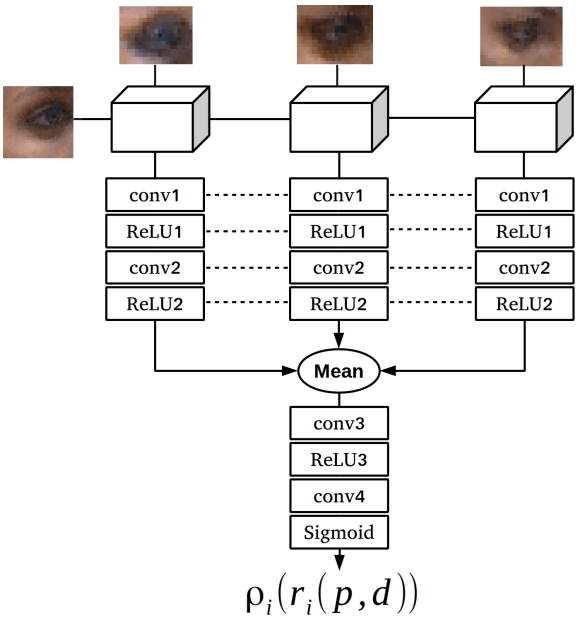



**Fig. 4** CNN architecture. Each cube is a pairwise comparison volume with $k^3$ samples that contain 6 valued vectors of RGB pairs and over which 3D convolutions are applied. The output score $\rho_i(r_i(p,d)) \in [0..1]$ encodes the photoconsistency measure at depth $d$ along the ray from pixel $p$ in image $I_i$.

### 4.2 Multi-View Neural Network

As explained in the previous section, at a given point $x$ along a viewing ray from image $I_i$ we can build $N-1$ color volumes with pairs of views $(I_i, I_{j\neq i})$. Each volume is composed of $k^3$ cells with pairs of RGB values. In order to detect whether the surface is going through $x$, we use siamese encoders similar in spirit to [39], with however 3D volumes instead of 2D patches. Each encoder considers as input a pairwise color volume and provides a feature. Features from all color volumes at $x$ are then averaged and fed into a final decision layer. Weight sharing and averaging are chosen to achieve camera order invariance.

The network is depicted in Figure 4. The inputs are the $N-1$ color volumes of size $k^3 \times 6$ where RGB pairs are concatenated at each sample within the volume. Convolutions are performed in 3D over the 6 valued vectors of RGB pairs. The first layers (encoders) of the network process every volume in parallel, with shared weights. Every encoder is a sequence of two convolutions followed by non-linearities, and max-pooling with stride. Both convolutional layers consist of respectively 16 and 32 filters of kernel $4 \times 4 \times 4$, followed by a Rectified Linear Unit (ReLU) and a max-pooling with kernel $2 \times 2 \times 2$ with stride 2. We then average the obtained $2 \times 2 \times 2 \times 32$ features and feed the result to a 128 filter $1 \times 1 \times 1$ convolutional layer, followed by a ReLU and a final $1 \times 1 \times 1$ decision layer, for a total of $72K$ parameters . The network provides a score $\rho_i(r_i(p,d)) \in [0..1]$ for the photoconsistency at depth $d$ along the ray from pixel $p$ in image $I_i$.

We experimented this network using different configurations. In particular, instead of averaging pairwise comparison features, we tried max-pooling which did not yield better results. Compared to the volumetric solution proposed by [6], the number of parameters is an order of magnitude less. As mentioned earlier, we believe that photoconsistency is a local property that requires less spatial coherence than shape properties.

### 4.3 Network Training

The network was implemented using TensorFlow [44] and trained from scratch using the DTU Robot Image Dataset [11], which provides multiview data equipped with *ground-truth* surfaces that present an accuracy of $0.5mm$. From this dataset 11 million $k^3$ color volumes were generated, from which we randomly chose 80 percent for training, and the remaining part for evaluation. Both positive and negative samples were equally generated by randomly sampling volumes up to $20cm$ away from ground truth points, where a volume is considered as positive when it contains at least $\mu$ ground truth points. In theory, the network could be trained with any number of camera pairs, however, in practice, we randomly choose from one up to 40 pairs. Training was performed with the binary cross entropy function as loss. Model weights are optimized by performing a Stochastic Gradient Descent, using Adaptive Moment Estimation [45] on $560,000$ iterations with batch size of 50 comparisons, and with a random number of compared cameras (from 2 up to 40). Since our sampling grids are relatively small and camera dependent, we are able to generate enough sample variability for training, without the need for data augmentation.

## 5 Depth Estimation

As previously noted, our main motivation is to reconstruct live dynamic scenes, typically humans in motion. In such cases, it is advantageous to focus on the foreground objects in the observed scene rather than modeling the full scene. To this purpose, we limit the search domain for depths along viewing rays to a region defined by image silhouettes. In the following we explain how such a region is defined and we detail then the volume sweeping we adopt to identify image depths.

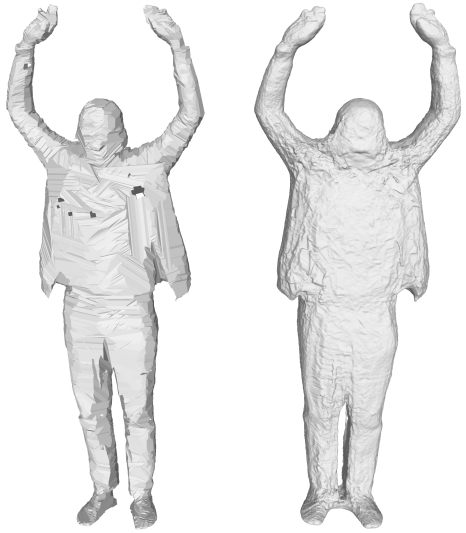


**Fig. 5** Left: the Confidence Volume with $\alpha = \beta = 54$, equivalent to the Visual hull with the 54 cameras that see the subject; Right: the Confidence Volume with $\alpha = \beta = 10$.

### 5.1 Confidence Volume

We assume we are given a set of $N$ images $\{I_i\}_{i=1}^N$ observed with a set $C$ of calibrated cameras with known projections $\{\pi_i\}_{i=1}^N$ and centers $\{c_i\}_{i=1}^N$. We assume we are also given a set of silhouettes $\{\Omega_i\}_{i=1}^N$, often available in multi-view scenarios by performing image segmentation, for instance background subtraction with constrained capture environments. The silhouettes are generally imprecise, as a result of multiple causes, including color ambiguities, that plague the segmentation. However, the redundant information they provide over several viewpoints can be used to restrict the search domains along viewing rays to segments that are likely to intersect the object surface. To this purpose we define the confidence volume $V$ as:

$$V = \{x \in \mathbb{R}^3 : \exists^{>\alpha} i \, (\pi_i(x) \in I_i) \\ \wedge \exists^{>\beta} i \, (\pi_i(x) \in \Omega_i)\}, \tag{1}$$

as the locus of points in $\mathbb{R}^3$ which project in $i > \alpha$ images and $\beta \leq i$ silhouettes to which they belong. When $\beta = i$, $V$ is simply the visual hull with $i$ images. $\alpha, \beta$ are two user defined constants that restrict weakly supported depth predictions with $\alpha$ and enable predictions away from the exact visual hull when $\beta < \alpha$. Intuitively, $V$ is a dilated version of the visual hull in the space region seen by at least $\alpha$ images, as shown in fig 5. As explained in the following section, the intersection of a viewing ray with $V$ defines the starting point of the depth search interval along that ray.

### 5.2 Volume Sweeping

In order to estimate pixel depths, the sampling volume introduced in section 4.1 is swept along their viewing rays while computing multi-view photoconsistency using the network detailed in section 4.2. For every camera, we sample therefore along viewing rays, test possible depth values, and choose the best candidate with respect to the network score. In practice, a reference view $I_i$ is only compared to the other views $I_j$ such that $\cos(\theta_{ij}) > 0.5$, where $\theta_{ij}$ is the angle between the optical axes of camera $i$ and $j$. Then, we sample rays from camera $i$ through every pixel $p$ and build colored volumes at every candidate depth, starting at the intersection with the confidence volume introduced in the previous section. Once the probability of surface presence is computed for every candidate, we define the estimated depth $d_i$ as:

$$d_i = \underset{d \in [d_{min}, d_{max}]}{\operatorname{argmax}} \; (\rho_i(r_i(p, d))), \tag{2}$$

where $\rho_i(r_i(p, d))$ is the consistency measure along the ray from $p$ in image $I_i$, as estimated by the network. $[d_{min}, d_{max}]$ is the search range with: $d_{min} = d_V(p)$ the intersection of the viewing ray at $p$ with the confidence volume; $d_{max}$ such that the search is stopped when the accumulated photoconsistency score reaches a given value $\rho_{max}$, in a *winner-takes-all* surface detection strategy.

$$\int_{x=d_{min}}^{d_{max}} \rho_i(r_i(p, x)) \mathrm{d}x \leq \rho_{max} \tag{3}$$

Depths for all pixels and from all images are further fused using a truncated signed distance function (TSDF) [46]. The following section explains how we define and extract the zero level-set of the TSDF.

## 6 Surface Extraction

We explained in the previous section how to compute depth maps for every viewpoint. We now have to fuse them into an implicit form, namely the TSDF [46] from which we can extract the zero-level set that corresponds to the reconstructed surface, which appears in *black* in Figure 6. Contrary to previous works [47, 24, 13, 22], we do not store TSDF values in a regular voxel grid but we rather devise a simple yet efficient sampling procedure derived from Voronoï Tesselation strategies, that specifically accommodates multi-view capture scenarios. It is worth mentioning that other works such as [27] also make use of irregular sampling strategies for MVS, but in a volumetric graph-cut framework.

### 6.1 Implicit Form Definition

For a point $x \in \mathbb{R}^3$, the truncated signed distance $TD(x) \in \mathbb{R}$ to the surface is defined as the weighted average of all camera contributions $F_i(x), i \in C$:

$$F_i(x) = \begin{cases} \min(\mu, \eta(x)) & iif\ \eta(x) \geq -\mu, \\ \emptyset & \text{otherwise}, \end{cases} \\ \eta(x) = d_i(\pi_i(x)) - \|c_i - x\|, \tag{4}$$

and:

$$TD(x) = \frac{\sum\limits_{i \in C_x} \rho_i(x) F_i(x)}{\sum\limits_{i \in C_x} \rho_i(x)}, \tag{5}$$

where $C_x = \{i \in C : F_i(x) \neq \emptyset\}$ and $\rho'_i(x)$ the photo-consistency measure (4) of the estimated depth along the ray from camera $i$ passing through $x$. If $d_i$ is undefined at $x$, *e.g.* $x$ is outside the camera visibility domain, then camera $i$ does not contribute to the TSDF. When no camera contributes at $x$ but $x$ is inside the confidence volume $V$ then it is considered as inside, *i.e.* $TD(x) < 0$. Note that contributions are weighted by the normalized photoconsistency measure which means that when cameras disagree about the photoconsistency at $x$, cameras with higher measures have an increased impact whereas cameras with low detection probability measures only marginally impact the reconstruction.

### 6.2 Extraction Procedure

From the previously defined TSDF, we extract the surface using a sampling strategy based on ray casting and Voronoï Tessellation. Figure 6 provides a 2D example of the main steps of the algorithm that are as follows:

1. (*orange*) Sample points inside the implicit form defined by the TSDF. This is achieved by randomly selecting pixels in all images and computing the point, along each pixel rays, inside but close to the surface according to the TSDF. The process is iterated until a user defined number of 3D points is reached.
2. (*blue*) Determine the Voronoï diagram: given the points inside the shape surface, a Voronoï diagram of this set of points is computed.
3. (*green*) Clip the Voronoï diagram with the zero level set of the TSDF. This operation extracts the intersection of the Voronoï cells with the surface to form an oriented mesh.

In the above strategy, sampling points close to the surface, and originating from image viewpoints, ensures that the 3D discretization is denser on the surface than inside the volume and also denser on surface regions observed by more images. The latter enables more precision to be given to surface regions for which more image observations are available.

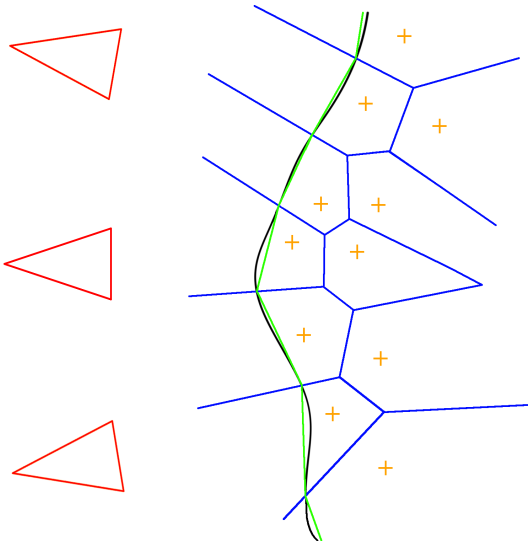


**Fig. 6** Our surface extraction procedure. The zero-level set of the implicit form (**black**) is observed by different cameras (**red**). They are used to provide the inside samples (**orange**) that will be used as the centroids for the Voronoï tessellation. This tessellation is finally clipped at the zero-level set and the final surface (**green**) can be extracted.

We visualize in figure 7 an example of extracted surface. We show 2 of the 40 input views in the top row, and our reconstruction in the middle. The bottom side of the bust is never seen by any camera. We show in the bottom row the difference in sampling resulting from the observation of the shape. The horizontal bottom side of the model is never observed, yet still correctly reconstructed. On the other hand, the triangles of the mesh in that area are much larger than the ones in the vertical upper part, which is observed by more cameras. This strategy allows for complete reconstructions of captured shapes with an adaptive sampling density depending on the observations of the object, focusing more samples in the regions where the details can be recovered.

**Runtime** The full pipeline allows us to reconstruct one time frame, *i.e.* 68 images ($2048 \times 2048$), in approximately 30 to 40 minutes using two NVIDIA Titan X GPUs, depending on the number of pixels that observe the shapes.

## 7 Experiments

Our main goals in this section are (i) to evaluate whether and how our learned photoconsistency contributes with respect to existing methods and (ii) to ver-

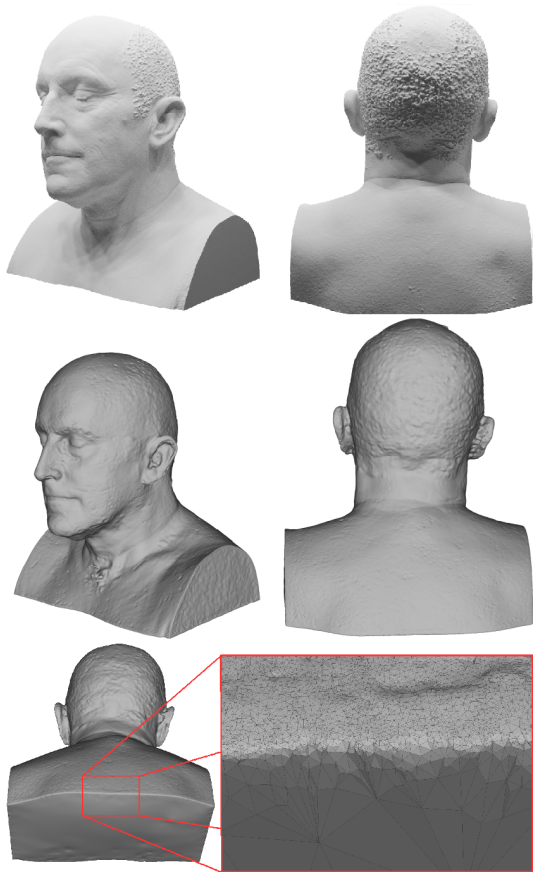


**Fig. 7** Two points of view of a synthetic model (*top*) and the result of our reconstruction (*middle*). A close-up of the extracted surface (*bottom*) at the limit between well-observed and unseen regions. The top part of the close-up is seen by many cameras whereas the bottom part is never observed.

ify whether these transfer to the more complex case of generic 3D capture scenes in practice, *e.g.* humans with complex clothing. To this aim, we perform various evaluations to verify and quantify the benefit of our learned multi-view similarity. We start by providing multiple validation experiments to justify the choices for the learning and reconstruction strategies in 7.1. Second, for comparison purposes, we apply in 7.2 our depth estimation approach in the static case using the [11] benchmark and compare it to state of the art MVS methods, both handcrafted and learning based. We make use of the standard *accuracy* and *completeness* metrics, both averaged and median, for which the evaluation code is provided by the authors.

Finally, we build experiments to test the main claim of improvement with production capture data in 7.3. To this goal we use several dynamic sequences captured on different platforms, which exhibit typical difficulties of such data. In particular, we mainly focus on the Kinovis acquisition platform [9], which consists of 68 RGB cameras, of resolution $2048 \times 2048$ with focal lengths varying from $8mm$ to $25mm$. We achieve very significant qualitative improvements compared to the state of the art approaches both learning-based [6,8], and handcrafted [10], without fine-tunning and despite the difference of capture setup used for training. We also compare to [23] on an example provided by the authors and achieve slightly better quality using only half the available information.

## 7.1 Validation

We previously formulated the problem of surface detection along viewing rays as a binary classification problem, as explained in section 4. In order to assess the benefit of our volumetric strategy, we first focus on different classifiers performances. We provide in 7.1.1 receptive field comparisons on the training dataset, this to enhance the advantage of casting and learning correlations in 3D. Additionally, section 7.1.2 provides a study of the depth hyperparameter of the receptive field of our network. Then, since preliminary results of [14] seemed to show a better robustness to a larger baseline, we design an experiment with cameras that are further apart to better quantify this improvement in section 7.1.3. We finally provide in 7.1.4 an ablation study of the accumulator described in 3 to validate its importance in the depth estimation procedure, in the performance capture scenario.

Section 7.1.2 shows that a volume size of $8 \times 8 \times 8$ is a preferred trade-off, thus will be used from now on, when not specified.

### *7.1.1 Classifiers Study*

In this paragraph, we compare performances of different classifiers based on various receptive fields:

1. Zero-Mean Normalized Cross Correlation (*ZNCC*): ZNCC is applied over the samples within the volumetric support region.
2. Learning (CNN) with a planar support: a planar equivalent of our volumetric solution, with the same architecture and number of weights, in a fronto-facing plane sweeping fashion.
3. Learning (CNN) with a volumetric support: our solution described in the previous sections.

Figure 8 shows, with the classifiers' ROC curves, that the most accurate results are obtained with a volumetric support and learning. Intuitively, a volumetric sampling region better accounts for the local non-planar geometry of the surface than planar sampling regions.

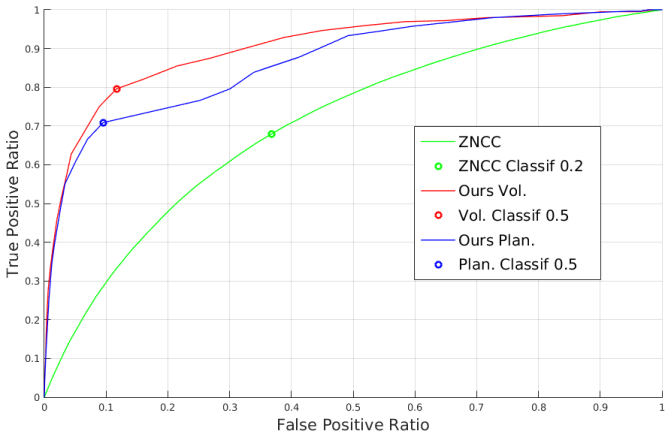



**Fig. 8** ROC Curves of three different classifiers, ZNCC, planar and volumetric supports, on the DTU Dataset [11]. Circles represent thresholds that optimize sensitivity + specificity with the values 0.2, 0.5 and 0.5 respectively.

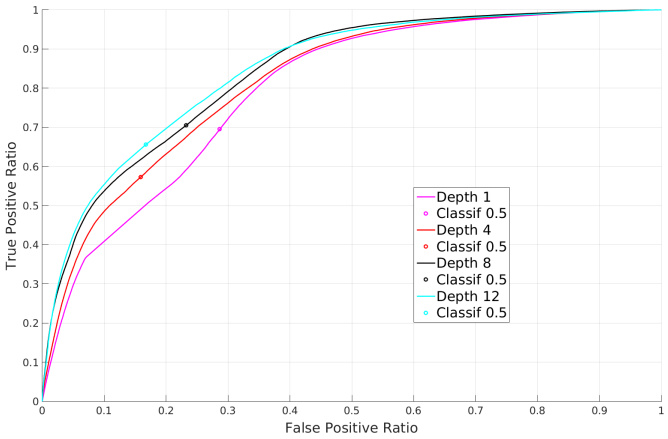



**Fig. 9** ROC Curves of four different classifiers using $8 \times 8$ receptive fields with various depths. Circles represent thresholds that optimize sensitivity + specificity.

This graph also emphasizes the significantly higher discriminative ability of learned correlations compared to deterministic ones.

### *7.1.2 Volume Sampling*

To further demonstrate this, we then proceed to a study on the impact of the depth parameter of the sampling volume. While keeping a $8 \times 8$ pixels reprojection on the images, we study the performances on classifiers with receptive fields varying in depth. Figure 9 shows classifiers performances with depth values ranging from 1 to 12. To perform this experiment, we had to diminish the networks number of parameters to fit the 12 depth training in memory and keep reasonable training and testing times, explaining the worse performances compared to previous ROC curves. This experiment demonstrates that the more information the network gathers along the ray the better the detection of the surface is. We choose a depth of 8 as it gives the best trade-off between computational complexity and performance.

### *7.1.3 Baseline Study*

We now evaluate the robustness to various baselines by accounting for a higher number of cameras and more distant cameras in the classification. Table 1 shows the accuracy of the classifiers with a varying number of cameras and for the optimal threshold values in Figure 8. As already noticed in the literature, *e.g.* [16, 19], a planar receptive field gives better results with a narrow baseline and the accuracy consistently decreases when the inter-camera space grows with additional cameras. In contrast the classifier based on a volumetric support exhibits more robustness to the variety in the camera baselines. This appears to be an advantage with large multi-camera setup as it enables more cameras to contribute and reduces hence occlusion issues.

| Camera # | 5 | 20 | 49 |
|---|---|---|---|
| ZNCC | 64.98 | 65.46 | 65.58 |
| Ours Plan. | 80.67 | 77.87 | 75.92 |
| Ours Vol. | **82.95** | **84.84** | **83.45** |

**Table 1** Classifier accuracy (%).

To push this experiment further, we design an experiment to test the robustness of our approach on a sparse capture platform, with lower scene coverage and wider baseline. Since no ground truth exists for this kind of performance capture scenario, we simulate it using of a realistic rendering engine to create a synthetic dataset. Similar to [9] in terms of camera parameters and capture volume, we chose to render only 10 randomly placed cameras, evenly distributed on an hemisphere around the capture volume. The average spacing between a camera and its 10 closest neighbors is $8.03m$ in this case, where it is $2.5m$ for the 68 POV kinovis platform and $0.188m$ in the 49 POV DTU case. For this experiment, we set the neighboring camera acceptation threshold $\cos(\theta_{ij})$ to 0.1, meaning that we accept almost orthogonal cameras. The synthetic cameras render the scene using Filmic Blender [48], a photorealistic configuration for Blenders Cycles ray-tracing engine. The images are generated with random parameters, *i.e.* the cameras parameters vary, in terms of position, orientation, focal length, and pixels number of samples, the latter directly affecting sensor noise. With this platform, we rendered a dozen of models such as procedu-

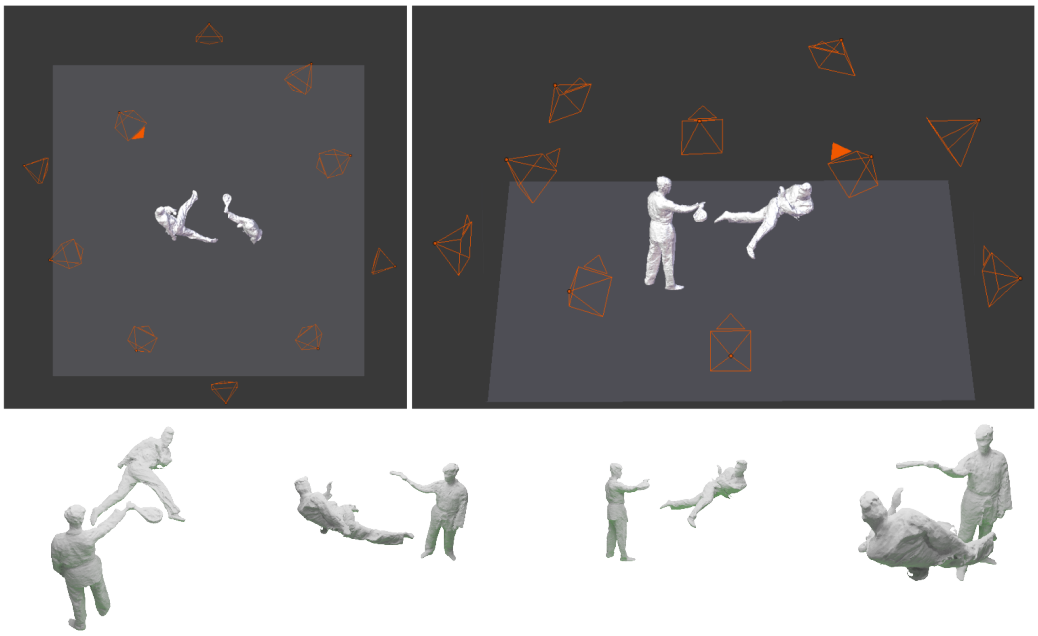


**Fig. 10** An example of sparse synthetic performance capture data generation. *(top)* Top and side view of the 10 cameras positioned around a surface. *(bottom)* Four examples of generated points of view.

rally generated geometric shapes, real life reconstructions or CAD models with various appearances. The multiview networks are trained from scratch on these synthetic examples, and evaluated on unseen synthetic data. Figure 10 shows an example of our synthetic platform as well as the generated synthetic data. We show in figure 11 the impact of a volumetric support: when the baseline between the cameras becomes extreme, it offers more robustness compared to a planar support, which appears very slanted in the compared view. Even though it is only a synthetic dataset, we believe that it gives interesting insights on the versatility of our volume sweeping strategy for the performance capture scenario. A qualitative result of this improved robustness is shown in figure 12. The area of the face is highly occluded, and the volumetric support helps recovering a smoother surface. Also note the details of the belt: the volume allows a sharp reconstruction of finer details, where a plane cannot handle finer geometry details.

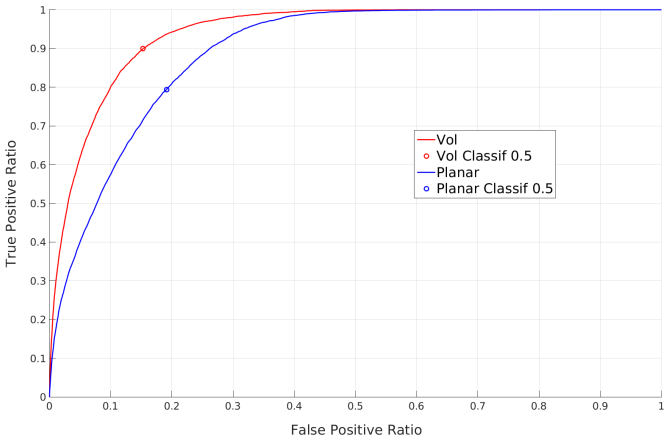



**Fig. 11** ROC Curves of two different classifiers using planar and volumetric receptive fields, on the sparse synthetic data. Circles represent thresholds that optimize sensitivity + specificity.

### *7.1.4 Accumulation Term*

We now provide a qualitative experiment to justify the use of the accumulation term in equation 3 in figure 13. This figure demonstrates the importance of the accumulation scheme in the performance capture scenario. The noisy photoconsistency in this case leads to a lot of false positives, creating extreme holes in the reconstructions when not using the accumulation scheme, *i.e.* $\rho_{max} \rightarrow \infty$. The addition of this term ($\rho_{max} = 1.6$) allows for smooth and faithful reconstructions, still containing most of the important geometric details.

## 7.2 Quantitative Comparisons

In this section, we compare our solution to various state-of-the-art methods using the DTU Robot Image Dataset [11]. We use the standard accuracy and completeness metrics to quantify the quality of the esti-

**Fig. 12** (*Left*) 3 input images, (*middle*) plane based classifier, (*right*) volumetric classifier. The face is highly occluded in many views (*left*) making the reconstruction noisy and inaccurate when using a planar support whereas the volume counterpart yields smoother and more accurate details.

**Fig. 13** (*top row*) Input images of captured subjects. (*middle row*) Reconstructions without probability accumulation along rays. (*bottom row*) Results with accumulation.

mated surface as described in [49], that is we define *accuracy* for a point of the reconstructed shape as the smallest Euclidean distance to the *ground-truth*, and the *completeness* of a point of the *ground-truth* as the smallest Euclidean distance to the reconstructed shape. For both metrics, we compare the average and median values over all the points of the shapes. To diminish the impact of far outliers in the metrics, we make use of the default thresholding parameter of [49]. We compare to Furukawa et al. [16], Campbell et al. [50] and Tola et al. [34], that are well-known handcrafted strategies, as well as to additional learning-based results from Ji et al. [6] and Hartmann et al. [39]. To conduct a fair comparison with [39], that is a patch based approach building a depthmap with a network comparable to ours, we use the result of our volume sweeping approach on the same depth map. When performing reconstructions on the DTU, we did not use the accumulation scheme in 3, *i.e.* $\rho_{max} \rightarrow \infty$. To speed up computations, we limit the search along a viewing ray to $5mm$ around a coarse depth estimation based on image descriptors [51]. Depths are sampled every $0.5mm$. As a post processing step, we simply add a soft bilateral filter, similarly to [39], accounting for color, spatial neighborhood, and probability of the detection.

Reconstruction results are depicted in table 2. We achieve quality on par with the best performing methods on this dataset, with a median accuracy and completeness in the range of the ground truth accuracy that we measured around $0.5mm$. It should be noticed that the best accuracy is obtained by Tola et al. [49] which tend to favor accuracy against completeness whereas Campbell et al. [50], in a symmetric manner, tend to favor completeness against accuracy. We obtain more balanced results on the 2 criteria, similarly to the widely used approach by Furukawa et al. [16], with however better performances. We also outperform the recent learning based method Surfacenet [6] on most measures in this experiment.

Compared to Hartmann et al. [39], and under similar experimental conditions, our approach give better results with 2 orders of magnitude less parameters, thereby confirming the benefit of volumetric supports over planar ones. Compared to Surfacenet. [6] (cube size $64 \times 64 \times 64$, sample step $0.4mm$) we obtain reconstructions of slightly better quality with an order of magnitude less parameters.

### 7.3 Qualitative Evaluation and Generalization

One of our main goals is to verify whether a learning based strategy generalizes to real life dynamic data and how it compares to state-of-the-art approaches in this

| Measure | Acc. | | Compl. | |
|---|---|---|---|---|
| | Mean | Med. | Mean | Med. |
| Tola et al. [49] | **0.448** | **0.205** | 0.754 | 0.425 |
| Furukawa et al. [16] | 0.678 | 0.325 | 0.597 | 0.375 |
| Campbell et al. [50] | 1.286 | 0.532 | **0.279** | **0.155** |
| Ji et al. [6] | 0.530 | 0.260 | 0.892 | 0.254 |
| Ours (*fused*) | 0.490 | 0.220 | 0.532 | 0.296 |
| Hartmann et al. [39] | 1.563 | 0.496 | 1.540 | 0.710 |
| Ours (*depthmap*) | **0.599** | **0.272** | **1.037** | **0.387** |

**Table 2** Reconstruction accuracy and completeness.

case. To this purpose, we focus our qualitative evaluation on two different dynamic capture datasets, both drastically different from the training one. We first perform, in section 7.3.1, reconstructions of dynamic RGB sequences captured by the Kinovis platform [9]. We then test, in section 7.3.2, our reconstruction method on a different real life dynamic dataset, captured with the active setup of [23] and compare to their results. It is important to note that the network previously trained on the DTU Dataset [11] was kept as such without any fine tuning at all times in this section.

### *7.3.1 Kinovis Data*

We first focus on data captured by [9], that is a hemispherical setup with 68 cameras of various focal lengths. In this scenario, standard MVS assumptions are often violated, *e.g.* wide baseline, specular surfaces, motion blur and occlusions, challenging therefore the reconstruction methods. A video demonstrating our results and providing comparisons on dynamic sequences is available online: `https://hal.archives-ouvertes.fr/hal-01849286`.

Most general purpose MVS methods we tested tend to fail in the performance capture scenario, either providing incomplete or low resolution results, or being extremely noisy. Figure 16 illustrates the reconstruction obtained using COLMAP [52], which is a handcrafted general purpose MVS pipeline based on *Patch-Match Stereo* [53]. Both methods perform overall correctly, as seen on the left side of the figure. However, [52] (*top-right*) struggles to recover fine-grained details while keeping the noise and artifacts level low, contrary to our approach (*bottom-right*). This results demonstrates the benefit of a dedicated method in the context of performance capture.

In order to assess the performances of our learned photoconsistency term, we compared in figure 1 to [10], which is a patch based sweeping method using traditional image features and specifically designed for this scenario. Both methods share a significant part of their pipeline, except for the photoconsistency evaluation, thus providing good insights about the benefits of the

**Fig. 14** (*top*) Input images, (*middle*) result with [10], (*bottom*) result with our method. Motion blur and low contrast are visible in the input images . Best viewed magnified.

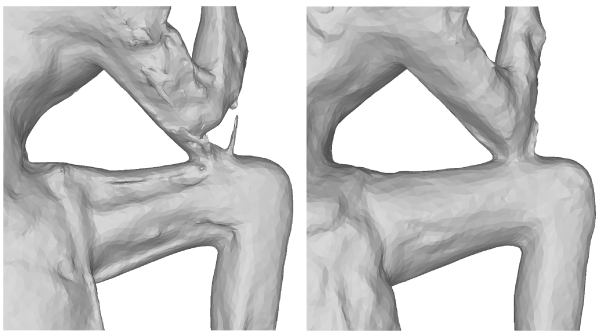


**Fig. 15** Close up view of the arm region in Figure 1. (*left*) Results from [10], (*right*) our reconstruction

**Fig. 16** (*top*) Reconstruction using COLMAP [52], (*bottom*) our result.

proposed learned term. Even though [10] performs well in contrasted regions, the patch based descriptors reach their limits in image regions with low contrast or low resolution. Figure 15 and 14 give such examples. They show that our strategy helps recovering finer surface details, while strongly decreasing noise in low contrast regions. The results obtained also demonstrate strong improvements in surface details, such as dress folds, that were undetected by the deterministic approach. In addition, they demonstrate lower levels of noise, particularly in self-occluded regions, and more robustness to motion blur as with the toes or tongue-in-cheek details that appear in Figure 14-bottom.

We then compare to the recent learning based approach [6] using the code available online (see Figure 17). Reconstructions with this approach were limited to a tight bounding box and different values for the volume sampling step were tested. The best results were obtained with a $2mm$ step. To conduct a fair comparison with our method, all points falling outside the visual hull were removed from the reconstruction. In this scenario, the point cloud obtained using [6] appeared to be very noisy and incomplete (see Figure 17-middle), plaguing the subsequent surface extraction step. Fig-

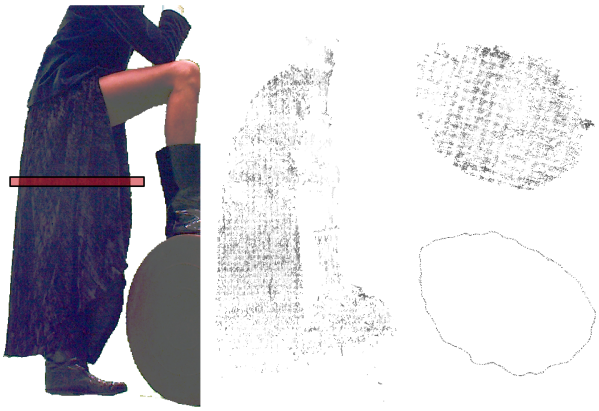


**Fig. 17** Qualitative comparison with [6]. (*Left*) input image with the horizontal section in *red*, (*middle*) point cloud with [6], (*right-top*) point cloud horizontal section with [6] (*right-bottom*) point cloud horizontal section with our approach.

ure 17-left also shows a horizontal section of the model in a poorly contrasted image region of the dress. The global strategy used in [6] wrongly reconstruct many surface points inside the shape volume (top figure), as a result of the ambiguous appearance of the dress. In contrast, our approach (bottom figure) correctly identify surface points by maximizing learned correlations along viewing rays.

In addition to this, we also compare to results of [8] provided by the authors in Figure 18. This method outputs a rather dense colored point cloud but similarly to results from [6], extracting a smooth surface from this point cloud remains a difficult task due to strong noise and missing data. Since the method uses custom and undocumented calibration parameters, it was not straightforward to remove points lying outside the visual hull. Moreover, the precision of the point cloud from [8] restricts its usage for performance capture and realistic reconstructions rendering. Figure 19 provides a close-up of the face of a subject. The level of detail of the result from [8] is not fine enough to correctly capture facial details, compared to the density of our output surface.

**Fig. 18** (*top*) Results provided by [8] on the kick 540 sequence. (*middle*) Poisson Reconstruction of the output point cloud of [8]. (*bottom*) Our result.

### *7.3.2 Active Capture Platform*

Finally, we compare our reconstructions of a scene captured with results from the active system of [23]. This setup consists of 52 RGB cameras mounted as stereo pairs but also differs from the previous dynamic capture scenario, as it also features an active system, projecting random infrared dots on the shape. 52 infrared cameras, also paired on stereo rigs then capture the reprojected spots on the shape, resulting in highly contrasted images, allowing to disambiguate the photoconsistency computation, especially in textureless regions without interfering with the visible appearance of the subject. In figure 20, we compare to results provided by the authors. While [23] make use of all the data available, we restrict our method to work with RGB images only. On the other hand, we allow cameras that are far apart to participate in the computation of the photoconsistency. Our results demonstrate the quality of our method's results, showing detailed reconstructions competitive with the results of [23] even though we only use the passive system, *i.e.* half of the available information. Figure 21 displays a close-up of the face of the subject. Our method allows to recover high-frequency facial details, such as the shape of the nostrils or the lips commissures, thus providing highly faithful reconstructions.

**Fig. 19** Point clouds density comparison between results provided by [8] (*left*) and our output (*right*). Best viewed magnified.

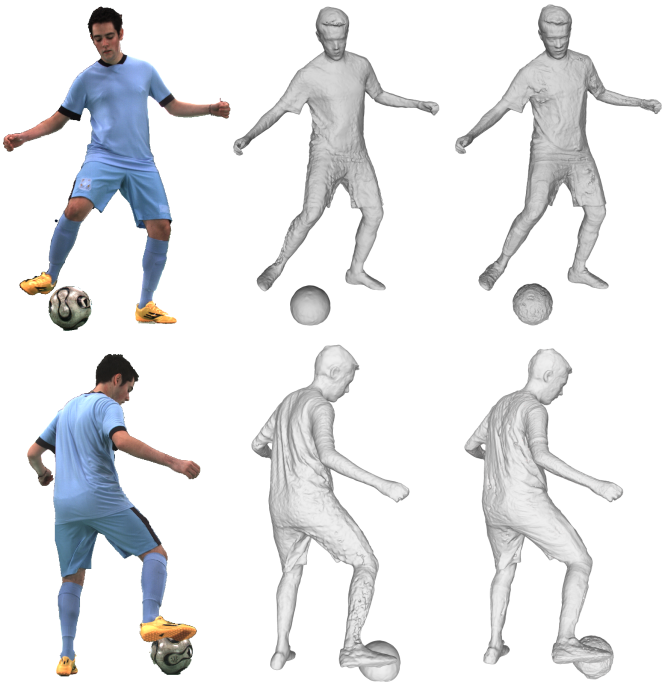


**Fig. 20** Two points of view of a subject from [23] (*left*). (*middle*) Reconstruction provided by the authors. (*right*) Results using our learning strategy.

**Fig. 21** Close up of the face of the subject from [23] (*left*). The reconstruction provided by the authors (*middle*) is very smooth compared to our result (*right*).

## 8 Conclusion and Future Works

We presented a learning framework for surface reconstruction in passive multi-view scenarios. Our solution consists in a $N$-view volume sweeping, trained on static scenes from a small scale dataset equipped with ground truth. Thanks to this new model, we validate the improvement of CNN-learned MVS photoconsistency in the case of complex and dynamic performance capture, with significant challenges typical of these datasets such as low light areas and low texture content and perceived resolution. This result is achieved with an order of magnitude less training parameters than previous comparable learned MVS works, showing significant network generalization from a training performed only on static DTU inputs, fully leveraging the high quality ground truth now available with these datasets. Thanks to our local strategy, our method achieved significantly improved detail recovery and noise reduction in complex real life scenarios, outperforming all existing approaches in this case.

The discretization of the volume around a query point involves a lot of redundancy and is a computationally expensive step for both training and inference. Moreover, even when optimized to process several neighboring depths in parallel, it remains rather memory inefficient. A possible future work could be to find a continuous representation for colored rays crossing the volume of interest, that could be used to infer surface presence probability in a similar manner with a much lighter computational cost.

Finally, we believe our approach is a first step towards a data-driven method to unify shape from silhouette and multi-view stereo inference, as made possible by the wide baseline robustness and general volumetric receptive field of our network, with the prospect of increased automation and quality.

## Acknowledgements

This work was conducted at the INRIA Grenoble. Funded by France National Research grant ANR-14-CE24-0030 ACHMOV. Images 1 - 2 - 15 - 17 of Anja Rubik courtesy of Ezra Petronio and Self Service Magazine. Geometric model in figure 7 courtesy of 3DScanStore [54].

## References

1. Žbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Res. **17**(1) (January 2016)

2. Luo, W., Schwing, A.G., Urtasun, R.: Efficient deep learning for stereo matching. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. (2016) 5695–5703
3. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. (2015)
4. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. (2017) 5622–5631
5. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: ECCV. (2016)
6. Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L.: Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In: The IEEE International Conference on Computer Vision (ICCV). (2017)
7. Kar, A., Häne, C., Malik, J.: Learning a multi-view stereo machine. In: Proc. Neural Information Processing Systems (NIPS). (2017)
8. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. ECCV (2018)
9. : Kinovis inria platform. `https://kinovis.inria.fr/inria-platform/`
10. Leroy, V., Franco, J.S., Boyer, E.: Multi-View Dynamic Shape Refinement Using Local Temporal Integration. In: IEEE, International Conference on Computer Vision 2017, Venice, Italy (2017)
11. Jensen, R.R., Dahl, A.L., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. (2014) 406–413
12. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015)
13. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR) 2015, Boston, MA, USA, June 7-12, 2015. (2015) 343–352
14. Leroy, V., Franco, J., Boyer, E.: Shape reconstruction using volume sweeping and learned photoconsistency. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX. (2018)
15. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA. (2006) 519–528
16. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. In: 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA. (2007)
17. Strecha, C., von Hansen, W., Gool, L.V., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. 2008 IEEE Conference on Computer Vision and Pattern Recognition (2008)
18. Gall, J., Stoll, C., Aguiar, E.D., Theobalt, C., Rosenhahn, B., peter Seidel, H.: Motion capture using joint skeleton tracking and surface estimation. In: CVPR. (2009)
19. Oswald, M.R., Cremers, D.: A convex relaxation approach to space time multi-view 3d reconstruction. In: ICCV Workshop on Dynamic Shape Capture and Analysis (4DMOD). (2013)
20. Mustafa, A., Kim, H., Guillemaut, J., Hilton, A.: Temporally coherent 4d reconstruction of complex dynamic scenes. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. (2016) 4660–4669
21. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2017)
22. Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: Volumedeform: Real-time volumetric non-rigid reconstruction. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII. (2016)
23. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A.G., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Trans. Graph. (2015)
24. Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S.R., Kowdle, A., Escolano, S.O., Rhemann, C., Kim, D., Taylor, J., Kohli, P., Tankovich, V., Izadi, S.: Fusion4d: Real-time performance capture of challenging scenes. ACM Trans. Graph. (2016)
25. Merrell, P., Akbarzadeh, A., Wang, L., michael Frahm, J., Nistr, R.Y.D.: Real-time visibility-based fusion of depth maps. In: Int. Conf. on Computer Vision and Pattern Recognition. (2007)
26. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE Comput. Graph. Appl. **27**(3) (May 2007)
27. Labatut, P., Pons, J., Keriven, R.: Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007. (2007) 1–8
28. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. International Journal of Computer Vision (2000)
29. Cremers, D., Kolev, K.: Multiview stereo and silhouette consistency via convex functionals over convex domains. IEEE Trans. Pattern Anal. Mach. Intell. **33**(6) (June 2011)
30. Ulusoy, A.O., Geiger, A., Black, M.J.: Towards probabilistic volumetric reconstruction using ray potentials. In: 3D Vision (3DV), 2015 3rd International Conference on. (2015)
31. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
32. Bay, H., Tuytelaars, T., Gool, L.J.V.: SURF: speeded up robust features. In: Computer Vision - ECCV 2006, 9th

European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I. (2006) 404–417
33. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA. (2003)
34. Tola, E., Lepetit, V., Fua, P.: DAISY: an efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5) (2010) 815–830
35. Pons, J.P., Keriven, R., Faugeras, O.: Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. IJCV **72**(2) (2007)
36. Collins, R.T.: A space-sweep approach to true multi-image matching. In: CVPR. (1996)
37. Furukawa, Y., Hernández, C.: Multi-view stereo: A tutorial. Foundations and Trends in Computer Graphics and Vision (2015)
38. Gallup, D., Frahm, J., Mordohai, P., Yang, Q., Pollefeys, M.: Real-time plane-sweeping stereo with multiple sweeping directions. In: 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA. (2007)
39. Hartmann, W., Galliani, S., Havlena, M., Van Gool, L., Schindler, K.: Learned multi-patch similarity. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
40. Huang, Z., Li, T., Chen, W., Zhao, Y., Xing, J., LeGendre, C., Ma, C., Luo, L., Li, H.: Deep volumetric video from very sparse multi-view performance capture. European Conference on Computer Vision (ECCV) (2018)
41. Gilbert, A., Volino, M., Collomosse, J., Hilton, A.: Volumetric performance capture from minimal camera viewpoints. European Conference on Computer Vision (ECCV) (2018)
42. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the International Conference on Computer Vision (ICCV). (2017)
43. Flynn, J., Neulander, I., Philbin, J., Snavely, N.: Deepstereo: Learning to predict new views from the world's imagery. Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
44. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015) Software available from tensorflow.org.
45. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In Bengio, Y., LeCun, Y., eds.: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. (2015)
46. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996. (1996)
47. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R.A., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A.J., Fitzgibbon, A.W.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011. (2011) 559–568
48. : Filmic blender. `https://sobotka.github.io/filmic-blender/`
49. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. Mach. Vis. Appl. (2012)
50. Campbell, N.D.F., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In: Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I. (2008) 766–779
51. Tola, E., Lepetit, V., Fua, P.: A fast local descriptor for dense matching. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA. (2008)
52. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV). (2016)
53. Bleyer, M., Rhemann, C., Rother, C.: Patchmatch stereo - stereo matching with slanted support windows. In Hoey, J., McKenna, S.J., Trucco, E., eds.: British Machine Vision Conference (BMVC). (2011)
54. : 3d scanstore. `https://www.3dscanstore.com`

# ANNEXE B

# List of publications

— 45 international peer reviewed, among which :
  — 37 international conferences (with 20 ICCV/CVPR/ECCV, high impact computer vision conferences).
  — 7 international peer reviewed journals (PAMI, IJCV, TVCG, IJMDB)
  — 1 book chapter
— 3 international workshops & demos (VRST, AMDO)
— 5 French peer reviewed conferences (RFIA)
— 3 research reports

## B.1 International Peer Reviewed Conferences

— **Reconstructing Human Body Mesh from Point Clouds by Adversarial GP Network**
Boyao Zhou, Jean-Sébastien Franco, Federica Bogo, Bugra Tekin, Edmond Boyer. ACCV - Asian Conference on Computer Vision, Nov 2020, Kyoto, Japan

— **Learning to Measure the Static Friction Coefficient in Cloth Contact**
Abdullah-Haroon Rasheed, Victor Romero, Florence Bertails-Descoubes, Stefanie Wuhrer, Jean-Sébastien Franco, Arnaud Lazarus. CVPR 2020 - IEEE Conference on Computer Vision and Pattern Recognition, June 2020 (**Oral Presentation**, acceptance rate **<5%**)

— **Moulding Humans : Non-parametric 3D Human Shape Estimation from Single Images**
Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, Gregory Rogez. ICCV 2019 - International Conference on Computer Vision, Oct 2019, Seoul, South Korea. pp.1-10

— **Adaptive Mesh Texture for Multi-View Appearance Modeling**
Matthieu Armando, Jean-Sébastien Franco, Edmond Boyer 3DV 2019 - 7th International Conference on 3D Vision, Sep 2019, Quebec City, Canada. pp.1-9
— **Analyzing Clothing Layer Deformation Statistics of 3D Human Motions**
Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, Stefanie Wuhrer ECCV 2018 - European Conference on Computer Vision, Sep 2018, Munich, Germany. pp.1-17
— **Shape Reconstruction Using Volume Sweeping and Learned Photoconsistency**
Vincent Leroy, Jean-Sébastien Franco, Edmond Boyer European Conference on Computer Vision, Sep 2018, Munich, Germany
— **Multi-View Dynamic Shape Refinement Using Local Temporal Integration**
Vincent Leroy, Jean-Sébastien Franco, Edmond Boyer International Conference on Computer Vision 2017, Oct 2017, Venice, Italy.
— **Surface Motion Capture Transfer with Gaussian Process Regression**
Adnane Boukhayma, Jean-Sébastien Franco, Edmond Boyer CVPR 2017 - IEEE Conference on Computer Vision and Pattern Recognition, Jul 2017, Honolulu, United States. pp.9
— **Cotemporal Multi-View Video Segmentation**
Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, Patrick Pérez, George Drettakis International Conference on 3D Vision, Oct 2016, Stanford, United States
— **Computing temporal alignments of human motion sequences in wide clothing using geodesic patches**
Aurela Shehu, Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, Stefanie Wuhrer 3DV 2016 - International Conference on 3D Vision 2016, Oct 2016, Stanford, United States
— **Estimation of Human Body Shape in Motion with Wide Clothing**
Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, Stefanie Wuhrer European Conference on Computer Vision 2016, Oct 2016, Amsterdam, Netherlands
— **Eigen Appearance Maps of Dynamic Shapes**
Adnane Boukhayma, Vagia Tsiminaki, Jean-Sébastien Franco, Edmond Boyer ECCV 2016 - European Conference on Computer Vision, Oct 2016, Amsterdam, Netherlands.
— **Volumetric 3D Tracking by Detection**
Chun-Hao Huang, Benjamin Allain, Jean-Sébastien Franco, Nassir Navab, Slobodan Ilic, Edmond Boyer IEEE. CVPR 2016 - IEEE Conference on Computer Vision and Pattern Recognition, Jun 2016, Las Vegas, United States. (**Spotlight Oral Presentation**, acceptance rate **<10%**)
— **An Efficient Volumetric Framework for Shape Tracking**
Benjamin Allain, Jean-Sébastien Franco, Edmond Boyer CVPR 2015 - IEEE International Conference on Computer Vision and Pattern Recognition, Jun 2015, Boston, United States. (**Oral Presentation**, acceptance rate **<3%**)
— **High Resolution 3D Shape Texture from Multiple Videos**
Vagia Tsiminaki, Jean-Sébastien Franco, Edmond Boyer, CVPR 2014 - IEEE International Conference on Computer Vision and Pattern Recognition, Jun 2014, Columbus, OH, United States
— **On Mean Pose and Variability of 3D Deformable Models**
Benjamin Allain, Jean-Sébastien Franco, Edmond Boyer, Tony Tung, ECCV 2014 - European Conference on Computer Vision, Sep 2014, Zurich, Switzerland. Springer

— **Multi-View Object Segmentation in Space and Time**
Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, Francois Le Clerc, Patrick Pérez, ICCV 2013 - International conference on computer vision, Dec 2013, Syndey, Australia.

— **3D Shape Cropping**
Jean-Sébastien Franco, Benjamin Petit, Edmond Boyer, Vision, Modeling and Visualization, Sep 2013, Lugano, Switzerland. Eurographics Association, p. 65-72.

— **N-Tuple Color Segmentation for Multi-View Silhouette Extraction**
Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, François Leclerc, Patrick Pérez, ECCV 2012 - 12th European Conference on Computer Vision, Oct 2012, Firenze, Italy.

— **Surface Flow from Visual Cues**
Benjamin Petit, Antoine Letouzey, Edmond Boyer, Jean-Sébastien Franco, Vision, Modeling and Visualization Workshop, Oct. 2011, Berlin, Germany.

— **Learning Temporally Consistent Rigidities**
Jean-Sébastien Franco, Edmond Boyer, CVPR 2011- IEEE Computer Vision and Pattern Recognition, Jun. 2011, Colorado Springs, United States. p. 1241-1248 (acceptance rate **30%**).

— **Probabilistic 3D Occupancy Flow with Latent Silhouette Cues**
Li Guan, Jean-Sébastien Franco, Edmond Boyer, Marc Pollefeys, CVPR 2010 - Conference on Computer Vision and Pattern Recognition, San Francisco, USA, June 2010 (acceptance rate **25%**).

— **A 3D Data Intensive Tele-immersive Grid**
Benjamin Petit, Thomas Dupeux, Benoît Bossavit, Joeffrey Legaux, Bruno Raffin, Emmanuel Melin, Jean-Sébastien Franco, Ingo Assenmacher, Edmond Boyer, ACM Multimedia, Oct 2010, Firenze, Italy. p. 1315-1318

— **Conversion of Performance Mesh Animation into Cage-based Animation**
Yann Savoye, Jean-Sébastien Franco, SIGGRAPH ASIA 2010 - ACM SIGGRAPH ASIA 2010 (Sketches and Posters Program), Dec 2010, Seoul, South Korea.

— **Cage-based Tracking for Performance Animation**
Yann Savoye, Jean-Sébastien Franco, ACCV 2010 - 10th Asian Conference on Computer Vision, Nov 2010, Queenstown, New Zealand.

— **Opportunistic Music**
Martin Hachet, Arash Kian, Florent Berthaut, Jean-Sébastien Franco, Myriam Desainte-Catherine, Joint Virtual Reality Conference JVRC 2009 (EGVE - ICAT - EuroVR), Lyon, France, Dec 2009

— **Remote and Collaborative 3D Interactions**
Benjamin Petit, Jean-Denis Lesage, Edmond Boyer, Jean-Sébastien Franco, Bruno Raffin, IEEE 3DTV Conference, Potsdam, Germany, May 2009

— **Multi-object shape estimation and tracking from silhouette cues**
Li Guan, Jean-Sébastien Franco, Marc Pollefeys, CVPR 2008 - Conference on Computer Vision and Pattern Recognition, Anchorage, USA, Jun 2008. (acceptance rate **30%**)

— **3D Object Reconstruction with Heterogeneous Sensor Data**
Li Guan, Jean-Sébastien Franco, Marc Pollefeys, 3DPVT 2008 - International Symposium on 3D Data Processing, Visualization and Transmission, Atlanta, USA, Jun. 2008. **Oral Presentation**.

— **3D Occlusion Inference from Silhouette Cues**
Li Guan, Jean-Sébastien Franco, Marc Pollefeys, CVPR 2007 - Computer Vision and Pattern Recognition, Minneapolis, USA, Jun. 2007. (**Oral Presentation**, acceptance rate **<6%**)
— **Visual Hull Construction in the Presence of Partial Occlusion**
Li Guan, Sudipta Sinha, Jean-Sébastien Franco, Marc Pollefeys, 3DPVT 2006 - 3D Processing, Visualization and Transmission, Chapel Hill, USA, Jun. 2006
— **Visual Shapes of Silhouette Sets**
Jean-Sébastien Franco, Marc Lapierre, Edmond Boyer, 3DPVT 2006 - 3D Processing, Visualization and Transmission 2006, Chapel Hill, USA, Jun. 2006. **Oral Presentation**.
— **The GrImage Platform : A Mixed Reality Environment for Interactions**
Clément Ménier, Jean-Sébastien Franco, Edmond Boyer, Bruno Raffin, ICVS 2006 - International Conference on Vision Systems, New York (USA), Jan. 2006.
— **Fusion of Multi-View Silhouettes Cues using a Space Occupancy Grid**
Jean-Sébastien Franco, Edmond Boyer, ICCV 2005 - International Conference on Computer Visio, Beijing (China), Oct. 2005. (acceptance rate **20.4%**)
— **Marker-less Real Time 3D Modeling for Virtual Reality**
Jérémie Allard, Edmond Boyer, Jean-Sébastien Franco, Clément Ménier, Bruno Raffin, Immersive Projection Technology (IPT'2004), Iowa State University (USA), May 2004.
— **Exact Polyhedral Visual Hulls**
Jean-Sébastien Franco, Edmond Boyer, BMVC 2003 - British Machine Vision Conference, Vol. I, p. 329-338, Norwich (UK), Sept. 2003. **Oral presentation**.
— **A Hybrid Approach for Computing Visual Hulls of Complex Objects**
Edmond Boyer, Jean-Sébastien Franco, CVPR 2003 - Computer Vision and Pattern Recognition. Vol. I, p. 695-701, Madison (USA), Jun 2003. (acceptance rate **23.1%**)

## B.2 INTERNATIONAL PEER REVIEWED JOURNALS

— **Volume Sweeping : Learning Photoconsistency for Multi-View Shape Reconstruction**
Vincent Leroy, Jean-Sébastien Franco, Edmond Boyer. International Journal of Computer Vision, 2020
— **Mesh Denoising with Facet Graph Convolutions**
Matthieu Armando, Jean-Sébastien Franco, Edmond Boyer. IEEE Transactions on Visualization and Computer Graphics 2020
— **Tracking-by-Detection of 3D Human Shapes : from Surfaces to Volumes**
Chun-Hao Huang, Benjamin Allain, Jean-Sébastien Franco, Nassir Navab, Slobodan Ilic, Edmond Boyer. IEEE Transactions on Pattern Analysis and Machine Intelligence, Institute of Electrical and Electronics Engineers, 2017,
— **Sparse Multi-View Consistency for Object Segmentation**
Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, François Le Clerc, Patrick Perez. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, pp.1 - 14.

— **Probabilistic Multi-view Dynamic Scene Reconstruction and Occlusion Reasoning from Silhouette Cues**
Li Guan, Jean-Sébastien Franco, Marc Pollefeys, International Journal of Computer Vision, Springer Verlag, 2010, 90 (3), pp. 283-303
— **Efficient Polyhedral Modeling from Silhouettes**
Jean-Sébastien Franco, Edmond Boyer, IEEE Transactions on Pattern Analysis and Machine Intelligence, Institute of Electrical and Electronics Engineers (IEEE), vol. 31 no. 3, pp. 414-427, Mars 2009.
— **Multicamera Real-Time 3D Modeling for Telepresence and Remote Collaboration**
Benjamin Petit, Jean-Denis Lesage, Ménier Clément, Jérémie Allard, Jean-Sébastien Franco, Bruno Raffin, Edmond Boyer, François Faure. International journal of digital multimedia broadcasting Volume 2010 (2010), Article ID 247108, 12 pages.

## B.3 BOOK CHAPTER

— **Multiview Calibration Synchronization and Dynamic Scene Reconstruction**
Marc Pollefeys, Sudipta Sinha, Li Guan, Jean-Sébastien Franco. Dans Multi-Camera Networks : Concepts and Applications (2009)

## B.4 INTERNATIONAL WORKSHOPS & DEMOS

— **CageIK : Dual-Laplacian Cage-Based Inverse Kinematics**
Yann Savoye, Jean-Sébastien Franco, AMDO 2010 : Articulated Motion and Deformable Objects, Jul 2010, Majorque, Spain. Springer
— **Grimage : 3D modeling for remote collaboration and telepresence** Benjamin Petit, Jean-Denis Lesage, Jean-Sébastien Franco, Edmond Boyer, Bruno Raffin. ACM symposium on Virtual reality software and technology (VRST'08), Bordeaux, France, Oct. 2008.
— **A Distributed Approach for Real-Time 3D Modeling**
Jean-Sébastien Franco, Clément Ménier, Edmond Boyer, Bruno Raffin, CVPR Workshop on Real-Time 3D Sensors and their Applications, Washington DC (USA), Jullet 2004.

## B.5 FRENCH PEER REVIEWED CONFERENCES

— **Apprentissage de la Cohérence Photométrique pour la Reconstruction de Formes Multi-Vues** Vincent Leroy, Jean-Sébastien Franco, Edmond Boyer RFIAP 2018 - Reconnaissance des Formes, Image, Apprentissage et Perception, Jun 2018, Marne la Vallée, France. pp.1-10, 2018
— **Segmentation multi-vues par coupure de graphes**
Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, François Le Clerc, Patrick Pérez RFIA 2014 - Reconnaissance de Formes et Intelligence Artificielle, June 2014, Rouen, France
— **Modélisation probabiliste pour la segmentation multi-vues**

Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, Francois Le Clerc, Patrick Pérez ORASIS - Congrès des jeunes chercheurs en vision par ordinateur, June 2013, Cluny, France.

— **Flot d'occupation 3D à partir de silhouettes latentes**
Jean-Sébastien Franco, Li Guan, Edmond Boyer, Marc Pollefeys, Reconnaissance de Forme et Intelligence Artificielle (RFIA'10), Caen, France, January 2010.

— **Fusion d'information de silhouettes à l'aide d'une grille d'occupation 3D**
Jean-Sébastien Franco, Edmond Boyer, Reconnaissance des Formes et Intelligence Artificielle (RFIA'06), Tours (France), January 2006.

— **Modélisation 3D temps-réelle distribuée**
Clément Ménier, Jean-Sébastien Franco, Edmond Boyer, Bruno Raffin, Orasis (Orasis'05), Clermont-Ferrand (France), Mai 2005.

— **Une approche hybride pour calculer l'enveloppe visuelle d'objets complexes**
Jean-Sébastien Franco, Edmond Boyer, Orasis (Orasis'03) p. 67-74, Gérardmer (France), Mai 2003.

## B.6 RESEARCH REPORTS

— **QuickCSG : Fast Arbitrary Boolean Combinations of N Solids** Matthijs Douze, Jean-Sébastien Franco and Bruno Raffin Research Report arXiv :1706.01558, ArXiv. June 2017.

— **Shape Animation with Combined Captured and Simulated Dynamics**
Benjamin Allain, Li Wang, Jean-Sébastien Franco, Franck Hetroy-Wheeler, Edmond Boyer Research Report arXiv :1601.01232, ArXiv. January 2016.

— **QuickCSG : Arbitrary and Faster Boolean Combinations of N Solids**
Matthijs Douze, Jean-Sébastien Franco, Bruno Raffin Research Report Inria RR-8687, March 2015

— **Fusion of Multi-View Silhouettes Cues using a Space Occupancy Grid**
Jean-Sébastien Franco, Edmond Boyer. Research Report Inria RR-5551, April 2005.

# Bibliographie


[ACOL00] Marc Alexa, Daniel Cohen-Or, and David Levin, *As-rigid-as-possible shape interpolation*, Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '00, 2000, pp. 157–164.

[AFB15] Benjamin Allain, Jean-Sébastien Franco, and Edmond Boyer, *An efficient volumetric framework for shape tracking*, Computer Vision and Pattern Recognition, IEEE, 2015.

[AFB19] Matthieu Armando, Jean-Sébastien Franco, and Edmond Boyer, *Adaptive Mesh Texture for Multi-View Appearance Modeling*, 3DV 2019 - 7th International Conference on 3D Vision (Quebec City, Canada), IEEE, September 2019, pp. 700–708.

[AFB20] Matthieu Armando, Jean-Sébastien Franco, and Edmond Boyer, *Mesh Denoising with Facet Graph Convolutions*, IEEE Transactions on Visualization and Computer Graphics (2020), code available at the following address :https ://gitlab.inria.fr/marmando/deep-mesh-denoizing.

[AFBT14a] Benjamin Allain, Jean-Sebastien Franco, Edmond Boyer, and Tony Tung, *On mean pose and variability of 3d deformable models*, ECCV, 2014.

[AFBT14b] Benjamin Allain, Jean-Sébastien Franco, Edmond Boyer, and Tony Tung, *On mean pose and variability of 3d deformable models*, European Conference on Computer Vision, 2014.

[AKSK08] Ijaz Akhter, Sohaib Khan, Yaser Sheikh, and Takeo Kanade, *Nonrigid structure from motion in trajectory space*, In NIPS, 2008.

[APSK07] Ehsan Aganj, Jean-Philippe Pons, Florent SÃ©gonne, and Renaud Keriven, *Spatio-temporal shape from silhouette using four-dimensional delaunay meshing*, 01 2007, pp. 1–8.

[ASK+05] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis, *SCAPE : shape completion and*

*animation of people*, ACM Transactions on Graphics **24** (2005), no. 3, 408–416, Proceedings of SIGGRAPH.

[AT06] A. Agarwal and B. Triggs, *Recovering 3d human pose from monocular images*, IEEE Transactions on PAMI **28** (2006), no. 1.

[BBK06] Alexander M. Bronstein, Michael M. Bronstein, and Ron Kimmel, *Efficient computation of isometry-invariant distances between surfaces*, SIAM Journal on Scientific Computing **28** (2006).

[BC08] L. Ballan and G. M. Cortelazzo, *Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes*, 3DPVT, 2008.

[BGC+15] Adrien Bartoli, Yan Gérard, François Chadebecq, Toby Collins, and Daniel Pizarro, *Shape-from-template*, IEEE Trans. Pattern Anal. Mach. Intell. **37** (2015), no. 10, 2099–2118.

[BH10] Christopher Budd and Adrian Hilton, *Temporal alignment of 3d video sequences using shape and appearance*, Conference on Visual Media Production, 2010, pp. 114–122.

[BHH11] Chris Budd, Peng Huang, and Adrian Hilton, *Hierarchical shape matching for temporally consistent 3d video*, 05 2011, pp. 172–179.

[BHKH13] Chris Budd, Peng Huang, Martin Klaudiny, and Adrian Hilton, *Global non-rigid alignment of surface sequences*, International Journal of Computer Vision **102** (2013), 256–270.

[BK02] S. Baker and T. Kanade, *Limits on super-resolution and how to break them*, IEEE PAMI **24** (2002), no. 9, 1167–1183.

[BM92] P. J. Besl and Neil D. McKay, *A method for registration of 3-d shapes*, IEEE Transactions on PAMI (1992).

[BP07] Ilya Baran and Jovan Popović, *Automatic rigging and animation of 3D characters*, ACM Transactions on Graphics **26** (2007), no. 3, 72 :1–72 :8.

[BPC13] A. Bartoli, D. Pizarro, and T. Collins, *A robust analytical solution to isometric shape-from-template with focal length calibration*, 2013 IEEE International Conference on Computer Vision, 2013, pp. 961–968.

[BPWG07] Mario Botsch, Mark Pauly, Martin Wicke, and Markus Gross, *Adaptive space deformations based on rigid cells.*, Comput. Graph. Forum **26** (2007), no. 3, 339–347.

[BTF+16] Adnane Boukhayma, Vagia Tsiminaki, Jean-sébastien Franco, Edmond Boyer, Adnane Boukhayma, Vagia Tsiminaki, Jean-sébastien Franco, Edmond Boyer, Eigen Appearance, Adnane Boukhayma, and Vagia Tsiminaki, *Eigen Appearance Maps of Dynamic Shapes To cite this version : HAL Id : hal-01348837 Eigen Appearance Maps of Dynamic Shapes*.

[BTG06] Herbert Bay, Tinne Tuytelaars, and Luc J. Van Gool, *SURF : speeded up robust features*, Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I, 2006, pp. 404–417.

[CAPM20] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll, *Implicit functions in feature space for 3d shape reconstruction and completion*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, jun 2020.

[CBI10a] Cedric Cagniart, Edmond Boyer, and Slobodan Ilic, *Free-from mesh tracking : a patch-based approach*, Computer Vision and Pattern Recognition, 2010.

[CBI10b] _____, *Probabilistic deformable surface tracking from multiple videos*, European Conference on Computer Vision, 2010.

[CCS+15] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan, *High-quality streamable free-viewpoint video*, ACM Transactions on Graphics **34** (2015).

[CK11] Daniel Cremers and Kalin Kolev, *Multiview stereo and silhouette consistency via convex functionals over convex domains*, IEEE Trans. Pattern Anal. Mach. Intell. **33** (2011), no. 6.

[Col96] Robert T. Collins, *A space-sweep approach to true multi-image matching.*, CVPR, 1996.

[CVHC08] Neill D. F. Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla, *Using multiple hypotheses to improve depth-maps for multi-view stereo*, Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I, 2008, pp. 766–779.

[CWL+16] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen, *Synthesizing training images for boosting human 3D pose estimation*, 3DV, 2016.

[CXG+16] Christopher Bongsoo Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese, *3d-r2n2 : A unified approach for single and multi-view 3d object reconstruction*, ECCV, 2016.

[dAST+08] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, *Performance capture from sparse multi-view video*, ACM Transactions on Graphics **27** (2008), no. 3.

[dATSS07] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel, *Marker-less deformable mesh tracking for human shape and motion capture*, Computer Vision and Pattern Recognition, 2007.

[DB11] A. Del Bue and A. Bartoli, *Multiview 3d warps*, 13th Internationl Conference on Computer Vision (ICCV 2011), Barcelona, Spain, November 2011.

[DFB+12] Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, François Le Clerc, and Patrick Pérez, *N-tuple color segmentation for multi-view silhouette extraction*, ECCV, 2012.

[DFB+13] Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, Francois Le Clerc, and Patrick Pérez, *Multi-View Object Segmentation in Space and Time*, ICCV'13 - International Conference On Computer Vision, December 2013.

[DFB+15] Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, François Le Clerc, and Patrick Pérez, *Sparse Multi-View Consistency for Object Segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **37** (2015), no. 9, 1890–1903.

[DFB+16] Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, Patrick Pérez, and George Drettakis, *Cotemporal Multi-View Video Segmentation*, 3DV 2016 -International Conference on 3D Vision (Stanford, United States), IEEE, October 2016, pp. 360–369.

[DFG99] Qiang Du, Vance Faber, and Max Gunzburger, *Centroidal voronoi tessellations : Applications and algorithms*, SIAM review **41** (1999), 637–676.

[DKD+16] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi, *Fusion4d : Real-time performance capture of challenging scenes*, ACM Trans. Graph. (2016).

[EDM+08] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. de Aguiar, N. Ahmed, C. Theobalt, and A. Sellent, *Floating Textures*, Comp. Graph. Forum **27** (2008), no. 2, 409–418.

[FB03] Jean-Sébastien Franco and Edmond Boyer, *Exact polyhedral visual hulls*, British Machine Vision Conference (BMVC'03) (Norwich, Royaume-Uni), vol. 1, 2003, pp. 329–338 (Anglais).

[FB05] ———, *Fusion of Multi-View Silhouette Cues Using a Space Occupancy Grid*, ICCV , 2005.

[FB09] ———, *Efficient polyhedral modeling from silhouettes*, IEEE Transactions on PAMI **31** (2009), no. 3.

[FB11] J.S. Franco and E. Boyer, *Learning temporally consistent rigidities*, Computer Vision and Pattern Recognition, 2011.

[FK98] Olivier Faugeras and Renaud Keriven, *Variational principles, surface evolution, pde's, level set methods, and the stereo problem*, IEEE Transactions on Image Processing **7** (1998), no. 3, 336–344 (English).

[FMBR04] J.S. Franco, C. Menier, E. Boyer, and B. Raffin, *A distributed approach for real-time 3d modeling*, CVPR Workshop, 2004.

[FP10] Yasutaka Furukawa and Jean Ponce, *Accurate, dense, and robust multi-view stereopsis*, IEEE Transactions on PAMI **32** (2010), no. 8.

[FSG07] Rik Fransens, Christoph Strecha, and Luc Van Gool, *Optical flow based super-resolution : A probabilistic approach.*, CVIU **106** (2007), no. 1, 106–115.

[GAKC13] B. Goldluecke, M. Aubry, K. Kolev, and D. Cremers, *A super-resolution framework for high-accuracy multiview reconstruction*, International Journal of Computer Vision (2013), To appear.

[GC09] Bastian Goldluecke and Daniel Cremers, *Superresolution texture maps for multiview reconstruction*, ICCV (2009), 1677–1684.

[GFBP10] Li Guan, Jean-Sébastien Franco, Edmond Boyer, and Marc Pollefeys, *Probabilistic 3d occupancy flow with latent silhouette cues.*, Computer Vision and Pattern Recognition, 2010.

[GFM+07] David Gallup, Jan-Michael Frahm, Philippos Mordohai, Qingxiong Yang, and Marc Pollefeys, *Real-time plane-sweeping stereo with multiple sweeping directions*, 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA, 2007.

[GFM+19] Valentin Gabeur, Jean-Sébastien Franco, Xavier MARTIN, Cordelia Schmid, and Gregory Rogez, *Moulding Humans : Non-parametric 3D Human Shape Estimation from Single Images*, ICCV 2019 - International Conference on Computer Vision (Seoul, South Korea), IEEE, October 2019, pp. 1–10.

[GFP07] Li Guan, Jean-Sébastien Franco, and Marc Pollefeys, *3D Occlusion Inference from Silhouette Cues*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (United States), June 2007, pp. 1–8.

[GFP08a] ———, *3D Object Reconstruction with Heterogeneous Sensor Data*, 3DPVT, 2008.

[GFP08b] ———, *Multi-object shape estimation and tracking from silhouette cues*, IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. (Anchorage, United States), June 2008, pp. 1–8.

[GFP10] ———, *Multi-view occlusion reasoning for probabilistic silhouette-based dynamic scene reconstruction*, International Journal of Computer Vision **90** (2010), no. 3, 283–303, PDF pre-print title different from final published version.

[GM04] B. Goldlücke and M. Magnor, *Space-time isosurface evolution for temporally coherent 3D reconstruction*, Computer Vision and Pattern Recognition, 2004.

[GNK18] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos, *Densepose : Dense human pose estimation in the wild*, arXiv (2018).

[GP02] Sebastien Granger and Xavier Pennec, *Multi-scale EM-ICP : A fast and robust approach for surface registration*, European Conference on Computer Vision, vol. 4, 2002, pp. 69–73.

[GSA+09] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans peter Seidel, *Motion capture using joint skeleton tracking and surface estimation*, CVPR, 2009.

[GSFP06] Li Guan, Sudipa Sinha, Jean-Sébastien Franco, and Marc Pollefeys, *Visual Hull Construction in the Presence of Partial Occlusion*, Third International Symposium on 3D Data Processing, Visualization, and Transmission (United States), June 2006, pp. 413–420.

[GVCH18] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton, *Volumetric performance capture from minimal camera viewpoints*, European Conference on Computer Vision (ECCV) (2018).

[GXW+15] Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai, *Robust non-rigid motion tracking and surface reconstruction using l0 regularization*, Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2015.

[HAB+18] Chun Hao Huang, Benjamin Allain, Edmond Boyer, Jean-Sébastien Franco, Federico Tombari, Nassir Navab, and Slobodan Ilic, *Tracking-by-Detection of 3D Human Shapes : from Surfaces to Volumes*, IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2018), no. 8, 1994–2008.

[HAF+16] Chun-Hao Huang, Benjamin Allain, Jean-Sébastien Franco, Nassir Navab, Slobodan Ilic, and Edmond Boyer, *Volumetric 3d tracking by detection*, Computer Vision and Pattern Recognition, 2016.

[HAR+10] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel, *Multilinear pose and body shape estimation of dressed subjects from image sets*, Computer Vision and Pattern Recognition (San Francisco, USA), IEEE, 2010.

[HGH+17] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler, *Learned multi-patch similarity*, The IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[HLC+18] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Chongyang Ma, Linjie Luo, and Hao Li, *Deep volumetric video from very sparse multi-view performance capture*, European Conference on Computer Vision (ECCV) (2018).

[HZ00] R.I. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge University Press, June 2000.

[IZN+16] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger, *Volumedeform : Real-time volumetric*

*non-rigid reconstruction*, Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII, 2016.

[JDV+14] Rasmus Ramsbol Jensen, Anders Lindbjerg Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs, *Large scale multi-view stereopsis evaluation*, 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, 2014, pp. 406–413.

[JGZ+17] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang, *Surfacenet : An end-to-end 3d neural network for multiview stereopsis*, The IEEE International Conference on Computer Vision (ICCV), 2017.

[JLT+15] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh, *Panoptic studio : A massively multiview system for social motion capture*, ICCV, 2015.

[JP09] Zsolt Janko and Jean-Philippe Pons, *Spatio-temporal image-based texture atlases for dynamic 3-D models*, IEEE 3DIM (Kyoto, Japon), IEEE, October 2009, pp. 1646–1653.

[KBJM18] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik, *End-to-end recovery of human shape and pose*, CVPR, 2018.

[KH13] Michael Kazhdan and Hugues Hoppe, *Screened poisson surface reconstruction*, ACM Trans. Graph. **32** (2013), no. 3, 29 :1–29 :13.

[KHM17] Abhishek Kar, Christian Häne, and Jitendra Malik, *Learning a multi-view stereo machine*, Proc. Neural Information Processing Systems (NIPS), 2017.

[kin] *Kinovis inria platform*, https://kinovis.inria.fr/inria-platform/.

[KPZK17] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun, *Tanks and temples : Benchmarking large-scale scene reconstruction*, ACM Transactions on Graphics **36** (2017), no. 4.

[KS00] K. Kutulakos and S. Seitz, *A Theory of Shape by Space Carving*, International Journal of Computer Vision **38** (2000), no. 3, 199–218.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in Neural Information Processing Systems 25 (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), Curran Associates, Inc., 2012, pp. 1097–1105.

[LB12] Antoine Letouzey and Edmond Boyer, *Progressive shape models*, Computer Vision and Pattern Recognition, IEEE, June 2012.

[LDX10] Yebin Liu, Qionghai Dai, and Wenli Xu, *A point-cloud-based multiview stereo algorithm for free-viewpoint video*, IEEE Trans. Vis. Comput. Graph. **16** (2010), no. 3, 407–418.

[LFB17] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer, *Multi-View Dynamic Shape Refinement Using Local Temporal Integration*, IEEE, International Conference on Computer Vision 2017 (Venice, Italy), 2017.

[LFB18] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer, *Shape reconstruction using volume sweeping and learned photoconsistency*, Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX, 2018.

[LFB21] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer, *Volume Sweeping : Learning Photoconsistency for Multi-View Shape Reconstruction*, International Journal of Computer Vision **129** (2021), 284–299.

[LGS+13] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt, *Markerless motion capture of multiple characters using multi-view image segmentation*, IEEE Transactions on PAMI (2013).

[LHS01] Hendrik P. A. Lensch, Wolfgang Heidrich, and Hans-Peter Seidel, *A silhouette-based algorithm for texture registration and stitching.*, Graphical Models **63** (2001), no. 4, 245–262.

[LI07] Victor S. Lempitsky and Denis V. Ivanov, *Seamless mosaicing of image-based texture maps.*, CVPR, IEEE Computer Society, 2007.

[Low04] David G. Lowe, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision **60** (2004), no. 2, 91–110.

[LPK07] Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven, *Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts*, IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007, 2007, pp. 1–8.

[LPMG17] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler, *A generative model for people in clothing*, Proceedings of the IEEE International Conference on Computer Vision, 2017.

[LS11] Ce Liu and Deqing Sun, *A Bayesian approach to adaptive video super resolution*, CVPR (2011), 209–216.

[LSU16] Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun, *Efficient deep learning for stereo matching*, 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 5695–5703.

[LWL+09] Yang Liu, Wenping Wang, Bruno Lévy, Feng Sun, Dong-Ming Yan, Lin Liu, and Chenlei Yang, *On centroidal voronoi tessellation - energy smoothness and fast computation*, ACM Transactions on Graphics **28** (2009), no. 101.

[MAW+07] Paul Merrell, Amir Akbarzadeh, Liang Wang, Jan michael Frahm, and Ruigang Yang David Nistér, *Real-time visibility-based fusion of depth maps*, Int. Conf. on Computer Vision and Pattern Recognition, 2007.

[MBM+17] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodolà, Jan Svoboda, and Michael M. Bronstein, *Geometric deep learning on graphs and manifolds using mixture model CNNs*, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 **2017-Janua** (2017), 5425–5434.

[MCA07] L. Mundermann, S. Corazza, and T.-P. Andriacchi, *Accurately measuring human movement using articulated icp with soft-joint constraints and a repository of articulated models*, Computer Vision and Pattern Recognition, 2007.

[MKGH16] Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton, *Temporally coherent 4d reconstruction of complex dynamic scenes*, 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 4660–4669.

[MS03] Krystian Mikolajczyk and Cordelia Schmid, *A performance evaluation of local descriptors*, 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA, 2003.

[MS04] Facundo Mémoli and Guillermo Sapiro, *Comparing point clouds*, SGP (2004).

[MST+20] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng, *NeRF : Representing scenes as neural radiance fields for view synthesis*, The European Conference on Computer Vision (ECCV), 2020.

[MT02] T. Matsuyama and T. Takai, *Generation, visualization, and editing of 3d video*, Proc. of 3DPVT, 2002.

[MVK+20] Armin Mustafa, Marco Volino, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton, *Temporally coherent general dynamic scene reconstruction*, International Journal of Computer Vision **129** (2020), no. 1, 123–141.

[NFS15] Richard A Newcombe, Dieter Fox, and Steven M Seitz, *DynamicFusion : Reconstruction and tracking of non-rigid scenes in real-time*, Computer Vision and Pattern Recognition, 2015.

[NH13] A. Neophytou and A. Hilton, *Shape and pose space deformation for subject specific animation*, 3DV, 2013.

[NIH+11] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon, *KinectFusion : Real-time dense surface mapping and tracking*, ISMAR, IEEE, 2011.

[NMOG19] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger, *Occupancy flow : 4d reconstruction by learning particle dynamics*, International Conference on Computer Vision, October 2019.

[NSI01] Ko Nishino, Yoichi Sato, and Katsushi Ikeuchi, *Eigen-texture method : Appearance compression and synthesis based on a 3d model*, IEEE Trans. Pattern Anal. Mach. Intell. **23** (2001), no. 11, 1257–1265.

[NYD16] Alejandro Newell, Kaiyu Yang, and Jia Deng, *Stacked hourglass networks for human pose estimation*, ECCV, 2016.

[OMMG10] Maks Ovsjanikov, Quentin Mérigot, Facundo Mémoli, and Leonidas J. Guibas, *One point isometric matching with the heat kernel*, Comput. Graph. Forum **29** (2010), no. 5.

[PMPHB17] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black, *ClothCap : Seamless 4D clothing capture and retargeting*, ACM Transactions on Graphics, (Proc. SIGGRAPH) **36** (2017), no. 4, Two first authors contributed equally.

[PWH+17] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele, *Building Statistical Shape Spaces for 3D Human Modeling*, Pattern Recognition **67** (2017), 276–286.

[RCO+19] Audrey Richard, Ian Cherabier, Martin R. Oswald, Vagia Tsiminaki, Marc Pollefeys, and Konrad Schindler, *Learned Multi-View Texture Super-Resolution*, Proceedings - 2019 International Conference on 3D Vision, 3DV 2019 (2019), 533–543.

[RKP+08] Bodo Rosenhahn, Uwe G. Kersting, Katie Powell, Thomas Brox, and Hans-Peter Seidel, *Tracking clothed people*, pp. 295–317, Springer Netherlands, Dordrecht, 2008.

[RRBD+20] Abdullah-Haroon Rasheed, Victor Romero, Florence Bertails-Descoubes, Stefanie Wuhrer, Jean-Sébastien Franco, and Arnaud Lazarus, *Learning to Measure the Static Friction Coefficient in Cloth Contact*, CVPR 2020 - IEEE Conference on Computer Vision and Pattern Recognition (Seattle, United States), IEEE, June 2020, pp. 9909–9918.

[RWS19] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid, *LCR-Net++ : Multi-person 2D and 3D Pose Detection in Natural Images*, IEEE Trans. PAMI (2019).

[SCD+06] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski, *A comparison and evaluation of multi-view stereo reconstruction algorithms*, 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, 2006, pp. 519–528.

[SdATS07] Carsten Stoll, Edilson de Aguiar, Christian Theobalt, and Hans-Peter Seidel, *A volumetric approach to interactive shape editing*, Research

Report MPI-I-2007-4-004, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, June 2007.

[SFC+11] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake, *Real-time human pose recognition in parts from single depth images*, Cvpr 2011 (2011), 1297–1304.

[SH03] J. Starck and A. Hilton, *Model-based multiple view reconstruction of people*, International Conference on Computer Vision, 2003.

[SH07] Jonathan Starck and Adrian Hilton, *Surface capture for performance-based animation*, IEEE Comput. Graph. Appl. **27** (2007), no. 3.

[SMNLF08] Mathieu Salzmann, Francesc Moreno-Noguer, Vincent Lepetit, and Pascal Fua, *Closed-form solution to non-rigid 3d surface registration.*, European Conference on Computer Vision, 2008.

[SSG+17] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger, *A multi-view stereo benchmark with high-resolution images and multi-camera videos*, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017.

[SvHG+08] Christoph Strecha, Wolfgang von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen, *On benchmarking camera calibration and multi-view stereo for high resolution imagery*, 2008 IEEE Conference on Computer Vision and Pattern Recognition (2008).

[SVMS14] Tomas Simon, Jack Valmadre, Iain Matthews, and Yaser Sheikh, *Separable spatiotemporal priors for convex reconstruction of time-varying 3d point clouds*, European Conference on Computer Vision (2014).

[SY10] Y. Sahillioğlu and Y. Yemez, *3d shape correspondence by isometry-driven greedy optimization*, Computer Vision and Pattern Recognition, 2010.

[SYLK18] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz, *Pwc-net : Cnns for optical flow using pyramid, warping, and cost volume*, 2018.

[TAL+07] Christian Theobalt, Naveed Ahmed, Hendrik P. A. Lensch, Marcus A. Magnor, and Hans-Peter Seidel, *Seeing people in different light-joint shape, motion, and reflectance capture.*, IEEE Trans. Vis. Comput. Graph. **13** (2007), no. 4, 663–674.

[TFB14] Vagia Tsiminaki, Jean-Sebastien Franco, and Edmond Boyer, *High Resolution 3D Shape Texture from Multiple Videos*, Proceedings ofIEEE Conference on Computer Vision and Pattern Recognition (2014), 1502–1509.

[TLF10] Engin Tola, Vincent Lepetit, and Pascal Fua, *DAISY : an efficient dense descriptor applied to wide-baseline stereo*, IEEE Trans. Pattern Anal. Mach. Intell. **32** (2010), no. 5, 815–830.

[TSF12] Engin Tola, Christoph Strecha, and Pascal Fua, *Efficient large-scale multi-view stereo for ultra high-resolution image sets*, Mach. Vis. Appl. (2012).

[TSSF12] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, *The Vitruvian manifold : Inferring dense correspondences for one-shot human pose estimation*, Computer Vision and Pattern Recognition, IEEE, 2012.

[Tun08] Tony Tung, *Simultaneous super-resolution and 3D video using graph-cuts*, CVPR (2008), 1–8.

[UGB15] Ali Osman Ulusoy, Andreas Geiger, and Michael J. Black, *Towards probabilistic volumetric reconstruction using ray potentials*, 3D Vision (3DV), 2015 3rd International Conference on, 2015.

[UZU+17] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox, *Demon : Depth and motion network for learning monocular stereo*, 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 5622–5631.

[VBMP08] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popovic, *Articulated mesh animation from multi-view silhouettes*, ACM Transactions on Graphics **27** (2008), no. 3.

[VBV18] Nitika Verma, Edmond Boyer, and Jakob Verbeek, *FeaStNet : Feature-Steered Graph Convolutions for 3D Shape Analysis*, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, jun 2018, pp. 2598–2606.

[VCR+18] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid, *Bodynet : Volumetric inference of 3D human body shapes*, ECCV, 2018.

[vL16] Jure Žbontar and Yann LeCun, *Stereo matching by training a convolutional neural network to compare image patches*, J. Mach. Learn. Res. **17** (2016), no. 1.

[VRM+17] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid, *Learning from Synthetic Humans*, CVPR, 2017.

[WBSS04] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli, *Image quality assessment : From error visibility to structural similarity*, IEEE Transactions on Image Processing **13** (2004), no. 4, 600–612.

[WHC+16] Lingyu Wei, Qixing Huang, Duygu Ceylan, Etienne Vouga, and Hao Li, *Dense human body correspondences using convolutional networks*, Computer Vision and Pattern Recognition, 2016.

[WSSC11] T. Windheuser, U. Schlickewei, F.R. Schmidt, and D. Cremers, *Geometrically consistent elastic matching of 3d shapes : A linear programming solution*, International Conference on Computer Vision, 2011.

[YFHWW16] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer, *Estimation of Human Body Shape in Motion with Wide Clothing*, ECCV 2016 - European Conference on Computer Vision 2016 (Amsterdam, Netherlands) (Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, eds.), vol. 9908, Lecture Notes in Computer Science, no. Part IV, Springer, October 2016, pp. 439–454.

[YFHWW18] Jinlong Yang, Jean-Sébastien Franco, Franck Hetroy-Wheeler, and Stefanie Wuhrer, *Analyzing Clothing Layer Deformation Statistics of 3D Human Motions*, ECCV 2018 - European Conference on Computer Vision (Munich, Germany), Lecture Notes in Computer Science, vol. 11211, Springer, September 2018, pp. 245–261.

[YLL+18] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan, *Mvsnet : Depth inference for unstructured multi-view stereo*, ECCV (2018).

[YMAL20] Jiayu Yang, Wei Mao, Jose M. Alvarez, and Miaomiao Liu, *Cost volume pyramid based depth inference for multi-view stereo*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[ZB15] Silvia Zuffi and Michael Black, *The stitched puppet : A graphical model of 3d human shape and pose*, CVPR, 2015.

[ZBH12] Andrei Zaharescu, Edmond Boyer, and Radu Horaud, *Keypoints and local descriptors of scalar functions on 2d manifolds*, International Journal of Computer Vision (2012).

[ZFB+20] Boyao Zhou, Jean-Sébastien Franco, Federica Bogo, Bugra Tekin, and Edmond Boyer, *Reconstructing Human Body Mesh from Point Clouds by Adversarial GP Network*, ACCV - Asian Conference on Computer Vision (Kyoto, Japan), Lecture Notes in Computer Science, vol. 12622, Springer, November 2020, pp. 123–139.

[ZHS+05] Kun Zhou, Jin Huang, John Snyder, Xinguo Liu, Hujun Bao, Baining Guo, and Heung-Yeung Shum, *Large mesh deformation using the volumetric graph laplacian*, ACM Trans. Graph. **24** (2005), no. 3, 496–503.

[ZK15] Sergey Zagoruyko and Nikos Komodakis, *Learning to compare image patches via convolutional neural networks*, IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015.

[ZPBPM17] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll, *Detailed, accurate, human shape estimation from clothed 3d scan sequences*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[ZWG+13] Y. Zeng, C. Wang, X. Gu, D. Samaras, and N. Paragios, *A generic deformation model for dense non-rigid surface registration : a higher-order MRF-based approach*, International Conference on Computer Vision, 2013.