Audio-visual multiple-speaker tracking for robot perception

Yutong Ban

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques et Informatique

Arrêté ministériel : 25 mai 2016

Présentée par

Yutong BAN

Thèse dirigée par **Radu HORAUD**, Directeur de recherche, INRIA
et codirigée par **Xavier ALAMEDA-PINEDA**, INRIA

préparée au sein du **Laboratoire Jean Kuntzmann** dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Suivi multi-locuteurs avec information audio-visuel pour la perception du robot

audio-visual multiple-speaker tracking for robot perception

Thèse soutenue publiquement le **10 mai 2019**,
devant le jury composé de :

Monsieur RADU HORAUD
DIRECTEUR DE RECHERCHE, INRIA CENTRE DE GRENOBLE RHÔNE-ALPES, Directeur de thèse
Monsieur ANDREA CAVALLARO
PROFESSEUR, UNIVERSITE QUEEN MARY DE LONDRES, Rapporteur
Madame LAURA LEAL-TAIXE
PROFESSEUR, UNIVERSITE TECHNIQUE DE MUNICH ALLEMAGNE, Rapporteur
Monsieur JEAN-LUC SCHWARTZ
DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES, Président
Monsieur SILEYE BA
INGENIEUR DE RECHERCHE, SOCIETE DAILYMOTION - PARIS, Examinateur
Monsieur XAVIER ALAMEDA-PINEDA
CHARGE DE RECHERCHE, INRIA CENTRE DE GRENOBLE RHÔNE-ALPES, Examinateur

Abstract

The 3D virtual representations of dressed humans appear in movies, video games and VR environments. To generate these representations, we usually perform 3D acquisitions or synthesize sequences with physics-based simulation or other computer graphics techniques such as rigging and skinning. These traditional methods generally require tedious manual intervention and generate new contents with low speed or low quality, due to the complexity of clothing motion. To deal with this problem, we propose in this work, a data-driven learning approach, which can take both captured and simulated sequences as learning data, and outputs unseen 3D shapes of dressed human with different body shape, body motion, clothing fit and clothing materials.

Due to the lack of temporal coherence and semantic information, raw captures can hardly be used directly for analysis and learning. Therefore, we first propose an automatic method to extract the human body under clothing from unstructured 3D sequences. It is achieved by exploiting a statistical human body model and optimizing the model parameters so that the body surface stays always within while as close as possible to the observed clothed surface throughout the sequence. We show that our method can achieve similar or better result compared with other state-of-the-art methods, and does not need any manual intervention.

After extracting the human body under clothing, we propose a method to register the clothing surface with the help of isometric patches. Some anatomical points on the human body model are first projected to the clothing surface in each frame of the sequence. Those projected points give the starting correspondence between clothing surfaces across a sequence. We isometrically grow patches around these points in order to propagate the

correspondences on the clothing surface. Subsequently, those dense correspondences are used to guide non-rigid registration so that we can deform the template mesh to obtain temporal coherence of the raw captures.

Based on processed captures and simulated data, we finally propose a comprehensive analysis of the statistics of the clothing layer with a simple two-component model. It is based on PCA subspace reduction of the layer information on one hand, and a generic parameter regression model using neural networks on the other hand, designed to regress from any semantic parameter whose variation is observed in a training set, to the layer parameterization space. We show that our model not only allows to reproduce previous re-targeting works, but generalizes the data synthesizing capabilities to other semantic parameters such as body motion, clothing fit, and physical material parameters, paving the way for many kinds of data-driven creation and augmentation applications.

Keywords. Clothing motion analysis • 3D garment capture • Body under clothing • Shape and motion estimation • Statistical shape space • Non-rigid registration •

Résumé

Les représentations virtuelles 3D de l'humain habillé apparaissent dans les films, les jeux vidéo, et depuis peu, dans les contenus VR. Ces représentations sont souvent générées par lacquisition 3D ou par la synthétisation des séquences avec les simulations basées sur la physique ou d'autres techniques d'infographie telles que le squelette et le skinning. Ces méthodes traditionnelles nécessitent généralement une intervention manuelle fastidieuse, elles génèrent à faible vitesse des contenus de mauvaise qualité, en raison de la complexité du mouvement des vêtements. Afin de résoudre ce problème, nous proposons dans ce travail une approche dapprentissage pilotée par les données, ce qui peut prendre à la fois des captures et des séquences simulées comme données dapprentissage, et produire sans les avoir vu des formes 3D de lhumain habillé ayant différentes formes et mouvements corporels, dans les vêtements de différentes adaptations et de matériaux variés.

En raison du manque de la cohérence temporelle et des informations sémantiques, il est difficile dutiliser directement les captures brutes dans l'analyse et l'apprentissage. Par conséquent, nous proposons dabord une méthode automatique pour extraire le corps humain sous des vêtements à partir de séquences 3D non structurées. Il est réalisé en exploitant un modèle de corps humain statistique et en optimisant les paramètres du modèle, de sorte que la surface du corps reste toujours à l'intérieur de la surface vêtue observée, et aussi près que possible de celle-ci. Nous montrons que notre méthode peut atteindre un résultat similaire ou meilleur que dautres méthodes de pointe et na pas besoin de lintervention manuelle.

Après avoir extrait le corps humain sous les vêtements, nous proposons une méthode pour enregistrer la surface du vêtement à l'aide de patchs isométriques. Certains points anatomiques du modèle du corps humain

sont d'abord projetés sur la surface du vêtement dans chaque cadre de la séquence. Ces points projetés donnent la correspondance de départ entre les surfaces de vêtement sur une séquence. Nous développons isométriquement des plaques autour de ces points afin de propager les correspondances sur la surface du vêtement. Par la suite, ces correspondances denses sont utilisées pour guider l'enregistrement non rigide afin que nous puissions déformer le maillage du modèle pour obtenir la cohérence temporelle des captures brutes.

Sur la base des captures traitées et des données simulées, nous proposons enfin une analyse complète des statistiques de la couche de vêtements avec un modèle simple à deux composants. Il est basé, d'une part, sur la réduction des sous-espaces PCA des informations de couche, et de lautre, sur un modèle de régression de paramètres génériques utilisant des réseaux neuronaux, conu pour régresser de tous les paramètres sémantiques dont la variation est observée dans lensemble des données dentranement. Nous montrons que notre modèle permet non seulement de reproduire des travaux précédents sur le ré-ciblage, mais aussi de généraliser les capacités de synthèse de données à dautres paramètres sémantiques tels que les mouvements corporels, ladaptation des vêtements et les matériaux physiques, ce qui ouvre la voie pour de nombreuses applications des créations et des augmentations axées sur les données.

Contents

List of Figures

Chapter 1

Introduction

Contents

1.1 Motivation

In the past decade, we have witnessed more and more frequent presence of clothed virtual characters in numerous films, games as well as emerging virtual reality applications such as virtual try-on.

Rather than showing completely rigid characters or using 3D captures directly, current applications show the continuously deforming geometry of dressed human. They mainly rely on two streams of techniques: one is based on rigging and skinning Kavan et al. [2007], Cordier and Magnenat-Thalmann [2004], Hess [2012] and the other is based on physics

simulations Baraff and Witkin [1998], Goldenthal et al. [2007], Long et al. [2011]. Rigging and skinning requires 3D animation artists to manually put a virtual skeleton inside the character and then attach the surface of the character to each bone of the skeleton with proper weight. And thus when the skeleton moves, the vertex will be transformed according to the weighted transformation of each bone. This kind of method is not only used for skin deformation (as suggested by the name), but can also be applied on clothes Cordier and Magnenat-Thalmann [2004]. The low algorithm complexity and fast running speed are the preeminent advantages. Although this technique can hardly generate fine detailed wrinkles, low spacial frequency deformation of clothes is enough for some applications. It is widely used in 3D computer games and animated cartoons. On the other hand, physics-based simulation treats dresses as objects with physical properties such as elasticity, viscosity, plasticity and resilience, and obtains their shape by solving a time-varying partial differential equation. This technique is capable of generating vivid and detailed clothing shapes. But due to the heavy computation of solving partial differential equations, a trade-off between speed and quality is usually required. As a consequence, those methods are mostly applied to the places where the requirement of visual quality outweighs the running speed, for example, typically in the movies.

To cope with the problem of the tedious manual work and heavy computation introduced by above approaches, a branch of data-driven methods has been proposed. Such methods generates new shapes by interpolating or mixing shape examples from a database. The database can be generated by using rigging and skinning approaches Lewis et al. [2000],

by applying physics-based simulators Xu et al. [2014b], or by performing 3D acquisition from real characters Neophytou and Hilton [2014]. These data-driven approaches show similar visual qualities w.r.t their dataset, and in some cases, possess a significant speed boost. However, these works are limited to interpolation between data points without further analysis on the data space.

Yet in the field of shape modeling, there are some applications of data-driven methods that analyze the data. They model the distribution of shape variation by performing statistics on given examples. Such approach has been proved to be a great success in statistical human body models Anguelov et al. [2005a], Jain et al. [2010], Loper et al. [2015], where principal component analysis is performed on body shape dataset and a lower dimensional Gaussian distribution is obtained to express plausible body shape variations.

Encouraged by these data-driven methods in human and clothing shape modeling, we would like to push the data-driven shape modeling of dressed moving human one step further by modeling the statistical shape space and learning the mapping from some controlling parameters to the shape rather than simple interpolation between raw examples. With potential high visual quality and fast running speed, such approach could be exploited in virtual try-on, computer games, movies and other applications.

1.2 Challenges

Modeling the shape space of dressed human with data-driven methods requires first to obtain a set of 3D sequences of dressed human performing various motion as the raw data. For dataset of real captures, since the unstructured surface points in the raw acquisition are not informative (e.g. no temporal coherence between frames and no meaningful corresponding anatomic point), we need to perform some preprocessing, which extracts useful information, such as clothing deformation and human motion parameterization for each frame, so that we can perform further analysis. When the shape is parameterized and the controlling variables are identified, we can analyze the functional relationship between them. In each part that is mentioned, challenges can be found. They involve but are not limited to

1. Dataset. Lack of public dataset for related researches
2. Preprocessing. Difficulty of data preprocessing, including
    a) Underlying body estimation;
    b) Clothing surface registration.
3. Modeling. Analyzing shape distribution and modeling the functional relationship between controlling variables to the shape is difficult due to high dimensionality and non-linearity.

In the following we will explain these challenges in details.

Dataset A key issue of data-driven modeling approaches is to obtain a rich dataset. One plausible source for obtaining large dataset is from

physics-based simulations. The simulator can provide clean shapes, registered surfaces and data with various well controlled and structured parameters, like different body shapes, motions, clothing materials and clothing fit. Such dataset contains no noise and almost all the useful information is at hand. However, it also suffer from the limitation of the implemented physics model itself as well as the oversimplification of clothing properties. On the contrary, real captures can deal with complicated clothes and produces true (although with noise) shapes. Thanks to the development in visual capture system and 3D reconstruction algorithms, it is now possible for us to acquire dataset of 3D clothing shape from real world by various means, such as using multi-view RGB camera system or Fusion4D Dou et al. [2016] with RGBD cameras. Despite of this, there are only very few publicly available 3D dataset for dressed human, such as Adobe data Vlasic et al. [2008] and BUFF Zhang et al. [2017], due to the tedious and time-consuming acquisition. Both existing synthetic dataset and real captured dataset suffer from small size and lack of coverage on controlling parameters.

Preprocessing Once the dataset is available, we can perform preprocessing to prepare the data for further analysis. The raw data contains unstructured shapes, without meaningful corresponding anatomic points and without temporal coherence in the sequence. It is necessary to obtain the information about the underlying body and to register the clothing surface through the sequence. The former can provide body shape and body motion, which are two of the key factors for clothing shape variation, while the latter computes trajectories of every vertex on the clothing surface, providing the possibility and convenience to compare two surfaces

and to extract relative deformation.

Although human body shape and the pose are generally known for synthetic data, the body are usually largely covered by the clothes in real captures. To analyze the relationship between the shape of human body and dressed human from real captures, it is necessary to first determine the body from dressed observation. For the cases where clothes are very tight, simply treating the dressed human as bare body wouldn't introduce much error. However, if the clothes is loose and wide, it is crucial to propose a method that can estimate both the body pose and the body shape for further analysis.

While human body is close to a piece wise rigid object, dressed human possesses a much higher degree of freedom. It is caused by the physical nature of cloth and the lack of constraints of the clothing to the body. Such high degree of freedom leads to difficulty of clothing surface registration through a sequence. Although some methods that registers dressed human sequences are available, such as Vlasic et al. [2008], Rosenhahn et al. [2007], Cagniart et al. [2010] and Allain et al. [2015], it remains an active research topic. We are interested in methods with higher registration accuracy and the ability of capturing dynamic details and reducing registration failures.

Modeling The shape of dressed human has a significant variation. It is not only caused by the non-rigidity of the cloth material itself, but also influenced by other numerous factors, which include but not limited to clothing style, clothing fit, human body shape, and human motion. And these factors do not have a simple linear relationship with the shape

variation. It is therefore difficult to find a mapping from these large number of controlling variables to the high dimensional shape space. How to build a unified framework that models these controlling variables at the same time and deals with the high dimensionality in an efficient way is a difficult yet valuable problem.

The above challenges are the major obstacles for study of the dressed human shape space. In this thesis, we will propose corresponding solutions and demonstrate their effectiveness.

1.3 Overview

In this thesis, we propose a data-driven approach to learn the shape space of dressed human in motion. The proposed solution contains three steps: human body estimation, clothing surface registration, and clothing deformation modeling respectively. The first two steps of our solution concentrate on extracting useful information from captured data, and the third step learns the shape space of dressed human from both synthetic and captured data. Figure 1.1 outlines the structure of our solution.

In Chapter 2, we examine the related work. We first look at the the work on 3D acquisition and human body shape space modeling. These are the fundamental works that we rely on. Then we discuss some 3D surface registration methods from the perspective of correspondence, deformation model, and tracking structure. Finally we study the existing works on modeling the dressed human shape space, pointing out their potentials as well as their limitation.

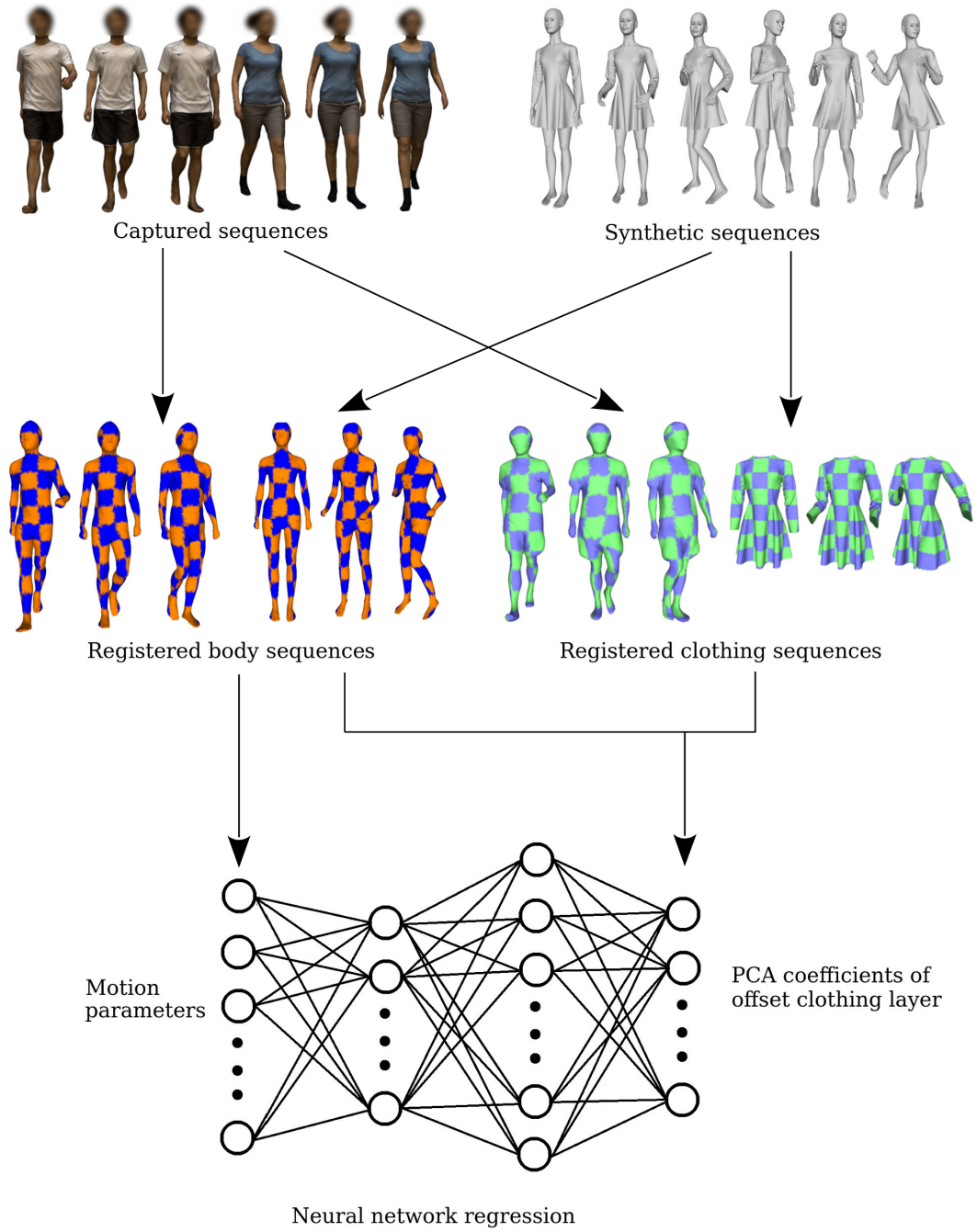


Figure 1.1 – Our proposed solution can take both captures and synthetic sequences as training data, extract registered body and clothing, and learn the mapping from the parameters of interest to the offset clothing layer.

In Chapter 3, we propose the first automatic method to estimate human body shape and pose from a raw captured 3D sequence of human with wide clothes. The problem is formulated as an optimization problem that solves for identity and posture parameters in a shape space of likely body shape variations. The automation is achieved by leveraging a recent robust pose detection method Zuffi and Black [2015]. 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.

Once we have the human body, we exploit it to align the clothing surface. In Chapter 4, we propose a surface registration framework where dense point-to-point correspondences between frames are obtained by growing isometric patches from a set of reliably obtained body landmarks. Compared to traditional geometric feature based correspondences, which are subject to instability of matches, this new method emphasizes the surface neighborhood coherence of matches, and matching density can be improved given sufficient body landmark coverage. We validate and verify the resulting improved registration performance in the evaluation section of Chapter 4.

Subsequently, when the body is obtained and the clothing surface is aligned, we propose a general framework to study the deformation of

the clothing layer from both synthetic or captured data in Chapter 5. We first propose a fuzzy correspondence to extract the offset clothing layer from the body surface. And then we use statistical analysis as a tool to analyze the geometric variability of the clothing layer to greatly reduce self-redundancies. Third, we show how to capture some of the underlying causal dynamics in relatively simple form, by modeling the variation of this clothing layer as a function of body motion as well as semantic variables using a statistical regression model.

Finally, we conclude and discuss the future work in Chapter 6.

Each of the three contributions summarized above has been published and presented in an international conference, respectively:

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

Computing temporal alignments of human motion sequences in wide clothing using geodesic patches
3DV 2016 - International Conference on 3D Vision.
Aurela Shehu*, Jinlong Yang[*], Jean-Sébastien Franco, Franck Hétroy-Wheeler, Stefanie Wuhrer.

∗. These authors contributed equally to this work

Analyzing Clothing Layer Deformation Statistics of 3D Human Motions

ECCV 2018 - European Conference on Computer Vision.

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

Chapter 2

Related work

Contents

In this chapter, we will review the related work in the following four aspects: 3D acquisition, human body shape space modeling, non-rigid surface registration methods and dressed human shape space modeling.

2.1 3D shape acquisition

3D shape acquisition has been an important research topic in computer vision for decades. Thanks to the development of acquisition hardware and reconstruction algorithms, now there are various methods and commercially available solutions.

The 3D shape acquisition pipeline can be generally split into two steps, 2D image acquisition and 3D reconstruction. 2D images can vary from stereo or multi-view RGB images, to depth images, obtained by structured light cameras or time-of-flight cameras. Meanwhile, the 3D reconstruction methods can vary from feature based stereo reconstruction Goesele et al. [2007], Furukawa and Ponce [2010], Geiger et al. [2011] to visual hull reconstruction using silhouettes Laurentini [1994], Grauman et al. [2003], Cheung et al. [2003] and to recent works on dynamic reconstruction Newcombe et al. [2015], Innmann et al. [2016], Dou et al. [2016], Yu12 et al. [2017], Yu et al. [2018]. In most of the 3D shape acquisition application, we would prefer high resolution, good accuracy with certain robustness, low cost and fast processing speed. In terms of dynamic dressed human shape acquisition, we will have additional requirement such as medium size of capture space and real time capturing performance. We will go through some popular 3D shape acquisition methods in the following part of this

section and introduce the system we used to capture some of the material in this thesis.

3D shape acquisition methods

Shape from stereo Stereo 3D reconstruction is a method that resembles human eyes. The capturing system is usually composed of two calibrated RGB cameras, which will acquire images of the object simultaneously. If the same point can be found on an image pairs, we can calculate the 3D position of that point by computing the intersection of two light rays, shooting from the focus of the camera to the corresponding points on the image, which is known as *triangulation*. The major effort on stereo reconstruction is to find the point pairs on the image pair that represent the same 3D point. Various methods can be applied. For example, feature based methods Lowe [2004], Bay et al. [2006], Goesele et al. [2007], Furukawa and Ponce [2010], Geiger et al. [2011] extract sparse features and matches them across images. Stereo system is usually of low cost, able of real-time performance, but normally suffers from low resolution and produces noisy data. To cover larger capturing volume and angle, a multi-view stereo system is usually applied, where multiple cameras are placed and calibrated for 2D image acquisition, while the stereo 3D reconstruction method remains the same. Such system will suffer from heavy computation and possibly sparse features.

Shape from multi-view silhouettes As mentioned in previous paragraph, a multi-view camera system is feasible for real-time capture, and compatible with large capturing volume. Since stereo 3D reconstruction

method requires a great amount computation power, this step can be replaced by visual hull reconstruction using silhouettes Laurentini [1994]. The silhouette images of the object defines what is inside and what is outside the object. Back projected to the 3D space, the 2D silhouette defines an infinite prism of inside and outside information. The intersection of multiple such prisms from different angles carves out a 3D visual hull of the shape acquired. Those silhouette images are usually extracted by exploiting chroma key or background subtraction. The Kinovis platform of Inria * is such a system. Some dataset in this thesis is captured in Kinovis. We will introduce it in details in Subsection 2.1.

Shape from structured light This method is based an active image acquisition set. A light source projects a set of known light patterns and a camera receives the reflected light at a know pose that differs from the light source. By analyzing the distortions of the received pattern, the acquisition system is able to extract the depth image. Typical examples of such system includes Kinect v1 and Breuckmann smart SCAN 3D-HE. Some of these acquisition systems can achieve real-time performance; some of them can achieve very high resolution but with lower frame rate Zanuttigh et al. [2016]. However, being an active system, these systems, on one hand, can hardly make use of multiple pattern projectors due to potential interference Zanuttigh et al. [2016]. Thus, it is difficult to capture all the sides of the 3D object, which makes it not suitable for our purpose. On the other hand, it is sensitive to ambient illumination. For example, Kinect v1 will generally fail, if the object is exposed directly under the sunshine Zanuttigh et al. [2016].

*. https://kinovis.inria.fr/

Shape from time-of-flight sensor Theses capturing systems are equipped with a light source that emits pulses of light to the surface of the scene. By calculating the round-trip travel time of the light pulses, the system is able to determine the distance between itself and the hit surface. In the last decade, some consumer level systems have appeared, such as Kinect v2 and Intel RealSense Depth Camera. Yet for these consumer level hardwares, the spatial resolution is usually too low for our purpose Li [2014]. But unlike structured light systems, time-of-flight sensors will hardly interfere with each other. Therefore, people can use multiple sensors to build a larger capturing system, which is able to capture almost the entire surface of the object with enhanced spatial resolution. Cooperated with the state-of-art fusion works such as Newcombe et al. [2015], Dou et al. [2016, 2017], which further reduce the noise and enhance both spatial resolution and temporal coherence, this 3D shape acquisition method could be an alternative to our acquisition method.

Kinovis

The Kinovis [†] is a 3D acquisition platform at Inria Grenoble. It is aimed to capture dynamic shapes, typically moving humans. The acquisition volume is approximately 7m $\times$ 5m $\times$ 3m, which is capable of containing 6 to 8 persons or allowing one person performing walking or running. A multi-view system of 68 RGB cameras (4M Pixels) is established to produce shape and appearance based on a visual hull reconstruction method Douze et al. [2015]. In addition, a standard Qualisys Mocap (Motion capture) system is also installed and calibrated to capture infrared

†. The Kinovis project is funded by France National Research grant ANR-11-EQPX-0024 KINOVIS.

reflective marker trajectories. We choose 30 FPS for RGB image acquisition and 120 FPS for Mocap during the acquisition.

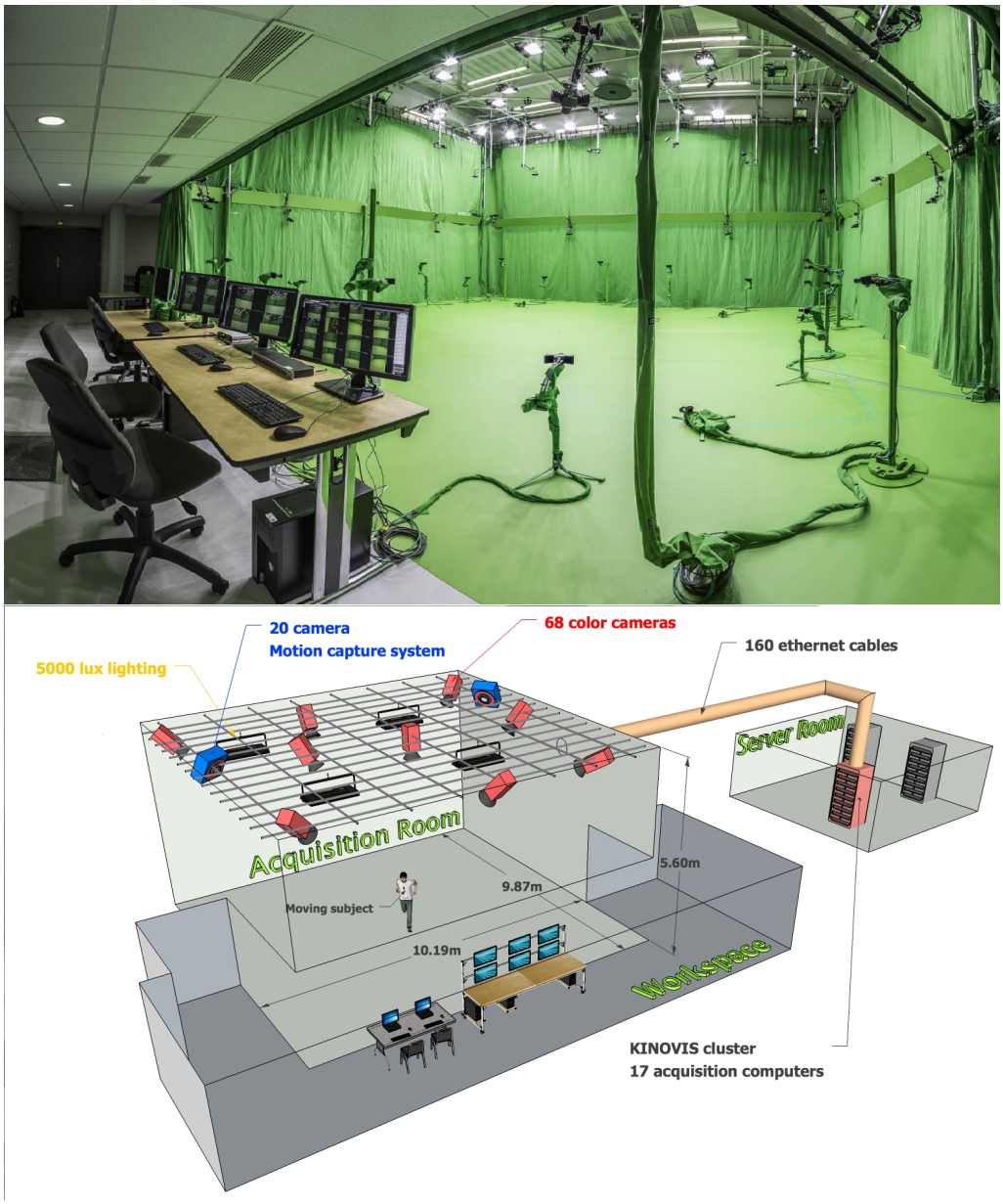


Figure 2.1 – Kinovis 4D modeling platform and it's setup. Images from https://kinovis.inria.fr/inria-platform/

2.2 Human body shape space

During the last two decades, 3D shape space modeling has greatly evolved in computer vision and computer graphics areas, thanks to the rapidly growing 3D dataset and the development of machine learning techniques. Among all the research in this field, statistical human body shape space is the most relevant to us, as the shape of a dressed human can be considered as an additional deformation on top of human body shape. Human body shape being a fundamental base for our work, we will first examine some popular and publicly available models and then review some works that extract the body shape from captures of dressed human.

Statistical human body models

Given raw dense captures of humans, historically, first approaches to process this data is to track human pose by fitting a 3D generic kinematic body-part model to these captured data. The *cylinder model* and the *stick-figure model* (see Figure 2.2) that represent human body are most often used Moeslund and Granum [2001].

To handle variation in body size, some of these models adapt the size of rigid components of the skeleton Sidenbladh et al. [2000], Sminchisescu and Triggs [2003], Sigal et al. [2010]. Taking this one step further, people build statistical human body shape models, so that the model can better explain captured observation by estimating a reasonable shape and pose parameters of the model Anguelov et al. [2005a], Loper et al. [2015], Neophytou and Hilton [2013]. We will briefly examine some of these models

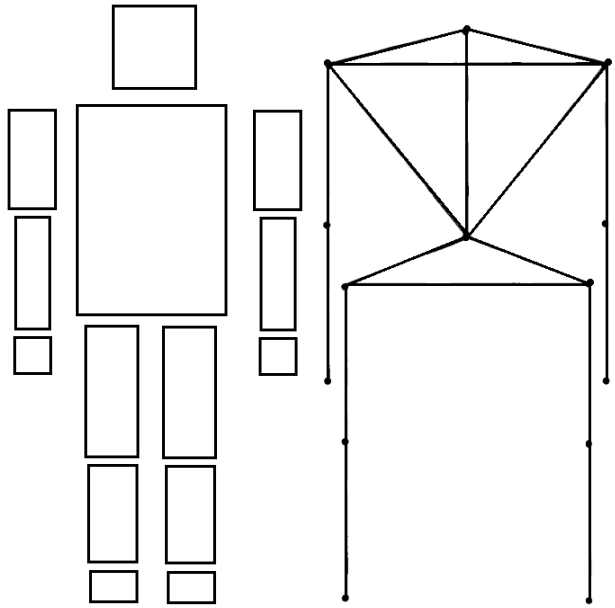


Figure 2.2 – Cylinder models and stick-figure models are most often used in the early stage to represent human body. Left: Cylinder model used in Rohr [1997]. Right: Stick-figure model used in Chen and Lee [1992].

in the following. More detailed and comprehensive information can be found in surveys such as Cheng et al. [2018].

SCAPE The Shape Completion and Animation of People (SCAPE) model Anguelov et al. [2005a] is the first data-driven statistical 3D human shape model that spans variation in both subject shape and pose. The pose dependent deformation is learned by registering the template model to a set of dense 3D scans of a single person in multiple poses. The registration is performed with the help of correspondences determined by Correlated Correspondence algorithm Anguelov et al. [2005b]. And the deformation is decoupled into rigid components, which is described in terms of skeleton, and non-rigid components, which models the remaining deformation such as flexing of the muscles. The identity dependent deformation is

learned on a set of 3D scans of multiple people in different poses. After registering the template model to all subject in the dataset, the identity dependent deformation space is compressed by using principal component analysis (PCA). In both cases, shape deformation is defined on each triangle of the template mesh. Since shape variations in pose and identity are accounted for separately, it produces a parametric model with two independent sets of parameters: joint rotation for pose dependent deformation and eigen-coefficient vector for identity dependent deformation, shown as

$$\boldsymbol{t}_i = \boldsymbol{T}_i(\boldsymbol{\theta}, J(\boldsymbol{\beta}))\boldsymbol{G}_i(\boldsymbol{\beta})\boldsymbol{F}_i(\boldsymbol{\theta})\hat{\boldsymbol{t}}_i \tag{2.1}$$

where $\boldsymbol{t}_i$ is the deformed triangle of $\hat{\boldsymbol{t}}_i$, $\boldsymbol{F}(\boldsymbol{\theta})$ represents shape deformation caused by pose $\boldsymbol{\theta}$, $\boldsymbol{G}(\boldsymbol{\beta})$ describes shape deformation introduced by different identity vector $\boldsymbol{\beta}$, and $\boldsymbol{T}(\boldsymbol{\theta}, J(\boldsymbol{\beta}))$ shows the rigid transformation determined by the pose based on rigging and skinning defined with skeleton joints $J(\boldsymbol{\beta})$.

Despite of the simplicity and the convenience brought by the SCAPE model, it still suffers from some disadvantages:

1. assumption that pose dependent deformation can be transfered from a particular subject to all other subjects;
2. triangle based deformation leads to computation complexity hence prevent further speeding up.

These disadvantages will be coped with in later works.

SPSD As we have mentioned, the SCAPE model assumes that pose dependent deformation of a particular subject can be transfered to any subject, therefore, it learns pose dependent deformation on a set of captures of only a single subject with multiple poses. To account for this, Neophytou *et al*. proposed Shape and Pose Space Deformation (SPSD) model Neophytou and Hilton [2013] that captures subject specific pose deformations. It is achieved by modeling the residual deformation (after eliminating the skinning deformation and identity deformation) as a function of both identity and pose:

$$\boldsymbol{t}_i = \boldsymbol{T}_i(\boldsymbol{\theta}, J(\boldsymbol{\beta}))\boldsymbol{G}_i(\boldsymbol{\beta})\boldsymbol{F}_i(\boldsymbol{\theta}, \boldsymbol{\beta})\hat{\boldsymbol{t}}_i \tag{2.2}$$

S-SCAPE Another drawback of SCAPE model is that the deformation is defined on triangles. Hence, an additional computationally demanding optimization step is needed to solve the vertex coordinates. To accelerate the performance, Jain *et al*. proposed a simplified version of SCAPE model (S-SCAPE) Jain et al. [2010], which models the shape variation directly on vertex displacement. It computes the body shape by performing PCA on the vertex coordinates of a training set in standard posture and combining this with a linear blend skinning (LBS) to model posture changes:

$$\boldsymbol{v}_i = \boldsymbol{T}_i(\boldsymbol{\theta}, J(\boldsymbol{\beta}))\boldsymbol{G}_i(\boldsymbol{\beta})\hat{\boldsymbol{v}}_i \tag{2.3}$$

where $\boldsymbol{v}_i$ is the deformed vertex position of neutral template vertex $\hat{\boldsymbol{v}}_i$, and $\boldsymbol{T}(\boldsymbol{\theta})$ is the LBS transformation for vertex i. Although this simplification will not be able to model pose dependent deformation such as

muscle bulging, it provides much simpler control and faster speed. To account for the known problem that counter-intuitive deformations will appear if any posture variations present in the training dataset is modeled in identity space, Pishchulin et al. [2017] starts by normalizing the posture of the training data before performing statistics. Due to its high efficiency and reasonable accuracy, we use S-SCAPE model in this thesis for human body. In theory, any other statistical human model could also be used in the pipeline.

SMPL Unlike S-SCAPE, Loper *et al*. proposed a Skinned Multi-Person Linear (SMPL) model Loper et al. [2015] that considers artifact of LBS and the influence of muscle bulging. The shape deformation is define directly on vertex coordinates. The model is deformed with pose and identity parameter vectors according to:

$$\boldsymbol{v}_i = \boldsymbol{T}_i(\boldsymbol{\theta}, J(\boldsymbol{\beta}))\boldsymbol{G}_i(\boldsymbol{\beta})\boldsymbol{F}_i(\boldsymbol{\theta})\hat{\boldsymbol{v}}_i \tag{2.4}$$

where $\boldsymbol{F}_i(\boldsymbol{\theta})$ compensates the artifacts of LBS and models pose dependent deformation such as flexion of muscles. The model is trained on two dataset, "multi-pose" dataset and "multi-shape" dataset Robinette et al. [2002]. Though compared with SPSD, the pose dependent offsets of SMPL are not dependent on body shape, it still produces high-quality visual synthesis and numerically low reconstruction error. Meanwhile, SMPL requires a much less runtime since it doesn't involve the optimization to reconstruct vertices for triangle based methods.

Dyna All the above mentioned models study the static shape space of human. The dynamic behavior of the soft-tissue is neglected and not modeled, leading to over-rigid visual result in generated motion sequence. Pons-Moll *et al*. proposed a model of Dynamic Human Shape in Motion (Dyna) Pons-Moll et al. [2015] to learn the mesh triangle deformation relative to a base body model caused my soft-tissue motion. Similar to the SCAPE model, Dyna defines shape deformation on triangles. But instead of modeling the identity shape with normalized pose as a function to only identity parameter, Dyna models it as a function of both identity and dynamic parameters $\boldsymbol{\delta}$ as follows:

$$\boldsymbol{t}_i = \boldsymbol{T}_i(\boldsymbol{\theta}, J(\boldsymbol{\beta}))(\boldsymbol{G}_i(\boldsymbol{\beta}) + \boldsymbol{D}_i(\boldsymbol{\delta}))\boldsymbol{F}_i(\boldsymbol{\theta})\hat{\boldsymbol{t}}_i \tag{2.5}$$

where $\boldsymbol{\delta}$ is a function of dynamic parameters for previous two frames and the joint angle $\dot{\boldsymbol{\theta}}$, joint acceleration $\ddot{\boldsymbol{\theta}}$, global speed $\boldsymbol{v}_k$, and global acceleration $\boldsymbol{a}_k$ for current frame k, as well as identity parameter:

$$\boldsymbol{\delta}_k = f(\boldsymbol{\delta}_{k-1}, \boldsymbol{\delta}_{k-2}, \dot{\boldsymbol{\theta}}, \ddot{\boldsymbol{\theta}}, \boldsymbol{v}_k, \boldsymbol{a}_k, \boldsymbol{\beta}) \tag{2.6}$$

This mapping function f that models dynamic shape deformation is learned on a dataset of 40,000 scans of ten subjects. This model successfully captures dynamics such as damping of the soft tissue and belly or chest deformation caused by breathing. Later, Loper *et al*. proposed a vertex-based version of Dyna, called Dynamic SMPL (DMPL) Loper et al. [2015].

Other human models The models introduced above are the most frequently used human models in current research. There are also other human models. For example, Tensor-based human model Chen et al. [2013] uses multi-linear model and considers identity and pose jointly; *Lie Bodies* model Freifeld and Black [2012] represents body shapes on a manifold and performs Principal Geodesic Analysis on the manifold to compress the shape space; relative rotation encoded model Hasler et al. [2009b] formulates the triangle deformation in a rotational-invariant way to better preserve identity shape from the influence of different poses. These human models extended the SCAPE model in different directions. But because they are not as frequently used as the models introduced above in the recent years, and they are not very closely related to the thesis, we will not go into details about these models.

State-of-the-art Very recent advances in human body models lie in two directions:

1. combination of body, hand and face models;
2. deep learning on the shape space.

All of the previously mentioned models considers the shape variation of human body, including torso and limbs. Some of them have certain variation on the size of the head, and on the size of the entire hand. But face shape variation caused by both identity and expression is not modeled. Identity dependent hand variation is not modeled either. MANO (hand Model with Articulated and Non-rigid deformations) Romero et al. [2017] creates a generative hand model from a dataset containing around 1000 high-resolution 3D scans of hands of 31 subjects under various poses.

It shares very similar idea with SMPL in terms of shape modeling. But instead of using all joint angles independently in the pose space, MANO uses a PCA subspace for hand poses. Combining MANO with SMPL, the new model is able to produce realistic shape variation of people performing natural movements. To also incorporate the ability of capturing face variation, the Frank model Joo et al. [2018] is built from a seamless combination of body, hand, and face models, where except the hand is an artist-rigged model with scaling parameters, both face and body models are trained from 3D dataset. Based on the Frank model, the Adam model Joo et al. [2018] is proposed, so that a common parameterization for all shape degrees of freedom is available. Apart from this advantage, the Adam model is also able to parameterize the shape of hair and clothes due to the modeling of an extra vertex displacement along surface normal.

Apart from combining body, hand and face models together, another trend of state-of-the-art work tries to model the shape space using different methods than linear compression techniques such as PCA. Most traditional generative body models assume there is a linear mapping between body vertex coordinates to a low dimentional parameterized body shape representation. They either utilize PCA or linear transform to reduce the dimensionality of the parameter space in the modeling and then synthesis new shapes in the compact subspace. Instead of this, some work admits the nonlinearity of the mapping and trys to learn the shape space directly using artificial neural networks Laine et al. [2017], Ranjan et al. [2018]. They claim that the linearity in PCA can not efficiently explain the data distribution, and thus the traditional models can not capture extreme deformations. And with other techniques, such as spectral convolutional

autoencoder, the new model can achieve much lower reconstruction error with fewer parameters. Although these works are focused on human face, the claims on PCA is also likely to be true for human body. Instead of replacing PCA with deep learning, Kanazawa et al. [2018] try to use an adversarial neural network to discriminate whether a give PCA coefficient vector of human body along with the pose vector represents a real human. This adversarial neural network is used to improve the 3D human surface recovery from 2D images. But it is also potentially useful to refine the representation of human body shape space.

Estimation of body shape under clothes

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. Bălan and Black [2008] 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. Inspired by this work and observing the richer information contained in dynamic sequences, we will evaluate our work on 3D motion sequences of different subjects captured in different motions and clothing styles (see Chapter 3). Hasler et al. [2009a] 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 are semi-automatic. They require manual input for

posture initialization.

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 Weiss et al. [2011], Helten et al. [2013]. These methods are not designed to work with clothing, and it is assumed that only tight clothing is present.

Existing work that designed to cope with the presence of clothing can be divided into two categories depending on the input data: 2D inputs and 3D inputs. State-of-the-art work that extract both articulated pose and body shape using 2D inputs include Bogo et al. [2016], Varol et al. [2018] and those using 3D inputs include Wuhrer et al. [2014], Neophytou and Hilton [2014], Zhang et al. [2017]. Compared with 2D inputs, 3D inputs are generally more difficult to acquire but contain compact and much richer information. Since we use Kinovis platform to obtain 3D data directly, we will concentrate only on 3D input processing methods.

The key idea of excluding the influence of clothing 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. [2014] use a shape space that learns local information around each vertex to estimate human body shape for a 3D motion se-

quence. 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-normalize the training shapes with the help of localized information.

Neophytou and Hilton [2014] 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 surface registration is required by Neophytou and Hilton. Computing surface registration through a 3D sequence is a difficult problem, and manual annotation is tedious when considering larger sets of motion sequences.

Zhang et al. [2017] estimate the body shape and pose under clothing from an unstructured 3D sequence. They make use of the SMPL model, and enhance the face by allowing certain free vertex displacement so that the method can also capture the subject's face.

2.3 Non-rigid registration of dressed human

Registration is a common problem in computer vision. In a raw captured 3D sequence, each frame is an independent surface, in the sense that no temporal correspondence of the surface point is given. Non-rigid registration of 3D surface aligns two surfaces, possibly two surfaces in adjacent frames, giving each point on one surface a corresponding point on the other surface, creating a deformation vector field between two surfaces. In this thesis, we are interested in registering the surface through an entire sequence to obtain complete point trajectories. Therefore, if not specified, we use *surface registration* to refer to *surface registration through the entire sequence*. Unlike a raw sequence of 3D captures that we take as input, a registered sequence possesses the information about global motion and non-rigid deformation.

3D surface registration is a general topic, in this thesis, we will mainly concentrate on the methods that are applied on scenarios of dressed human. Most of these methods exploit a template mesh or so-called a reference shape. This template is deformed to fit the observed shapes in each frame from the sequence. Although some works don't need the template mesh, but instead, treating 3D registration as a 4D space-time optimization problem, they usually assume small deformations between adjacent frames Zollhöfer et al. [2018], which may not the case for human with clothes. The template can be obtained from various sources. For example, it can be a manually selected shape from a certain frame of the sequence Zollhöfer et al. [2014], or a laser scan of the subject with the same dresses Vlasic et al. [2008], or a fused model from dynamic 3D recon-

struction Yu et al. [2018]. Once the template is available, those registration methods try to compute correspondences between frames or between observed frame and the template. Then the template is deformed according to a certain deformation model constrained by the correspondences to match the observation. The sequence can be tracked in the chronologically sequential order or in other structures.

Correspondences

Sparse correspondences Sparse correspondences are usually computed by detecting and matching color features or geometric features from 2D or 3D input, such as SIFT Lowe [2004], SURF Bay et al. [2006], FPFH Rusu et al. [2009], WKS Aubry et al. [2011] and CSHOT Tombari et al. [2010]. Sparse correspondences are often used to guide to optimization of the deformation, achieving faster convergence speed. In addition to that, the color features can help significantly to stabilize the alignment in the tangent plane of the model, prevent from surface sliding.

Dense correspondences Sparse correspondences lack of coverage on the surface. Therefore, they are not able to provide complete registration. Dense correspondences are computed to ensure the capture of fine level deformation. They can be computed by using functional maps Ovsjanikov et al. [2012], retrieving point correspondences in a function space. The dense correspondences can also be computed directly in Cartesian space using point-to-point Li et al. [2008], Newcombe et al. [2015], Guo et al. [2015] or point-to-plane Li et al. [2009], Zollhöfer et al. [2014], Dou et al. [2016] geometric criteria or in photometric domain along with illumination

estimation Guo et al. [2017].

3D correspondences is a key basis for shape registration. As there is a great amount of research in computing 3D correspondences, we would like to refer to some survey papers for more information: Van Kaick et al. [2011], Tam et al. [2013], Biasotti et al. [2016].

Deformation model

The correspondences alone are generally not enough to register two surfaces due to two reasons: they can not cover the entire surface and they may contain many false correspondences. Due to these defects, it is necessary to have a proper deformation model with effective regularization on the deformation field. Additionally, many deformation models reduce the dimensionality of the shape space, facilitating optimization computation. Most commonly used deformation models in 3D dressed human registration can be categorized into the following three classes.

Skeletal deformation model The kinematics of human skeleton is a natural and intuitive regularization for shape deformation. Some works Gall et al. [2009], Vlasic et al. [2008], Rosenhahn et al. [2007] rig a skeleton onto the template mesh, and skin the template surface to each bones with manually defined or learned skinning weights, so that the deformation of the template is controlled by the joint angles. Such a low dimensional shape space facilitates the optimization. But this model is only used for coarse registration, since fine deformation of clothes does not follow the regularization of skeleton. Therefore, a surface-based deformation model is also incorporated to achieve fine-level registration.

Surface-based deformation model To deal with free-form surface deformation, it is necessary to allow more flexibility of deformation but meanwhile maintain local consistency on the deformation field. Some works exploit local intrinsic surface properties, such as isometric deformations Bronstein et al. [2006], Sahillioglu and Yemez [2010] and as-rigid-as-possible deformations Sorkine and Alexa [2007]. Instead of parameterizing the deformation on every surface vertices, some other work Cagniart et al. [2010], Klaudiny and Hilton [2011] use embedded deformation model Sumner et al. [2007] or patch-based approaches Cagniart et al. [2010]. Those two methods are very similar in the sense that both of them define vertex transformation based on the sparsely distributed deformation controlling handles: surface patches or embedded deformation nodes. The number of those deformation handles is much smaller than the number of surface vertices. Thus the deformation degree of freedom of the surface is reduced to the level of deformation handle number. Connectivities among the nodes or the patches are defined so that deformation regularization such as as-rigid-as-possible energy can be applied to constrain the handles. Many state-of-the-art works on non-rigid surface alignment is done using embedded deformation model Zollhöfer et al. [2018].

Volumetric deformation model A major disadvantage of surface-based deformation methods is that they often lead to shrink of the volume. The shortest way to achieve volume preserving property is to treat the object as a volumetric entity, instead of defining deformation on the surface, computing the deformation field for each volume unity. This volume unity could be obtained from volumetric tetrahedralization Zollhöfer et al.

[2014], centroidal Voronoi tessellation Allain et al. [2015] or from regular volumetric grids Sederberg and Parry [1986], Zollhöfer et al. [2012], Slavcheva et al. [2017]. Compared with surface-based deformation model, volumetric deformation model usually needs more deformation parameters and performs less satisfying when the volume of the observation is either changeable (a deflating balloon) or even not true (e.g. a skirt that is considered as a solid object).

Registration order in multi-frame registration

It is often not robust to register two surfaces with large non-rigid deformation between them. Therefore, it is preferred to register a surface to another similar one. This can usually be achieved by registering an sequence in a chronological order: we deform the current surface to register the surface in the next frame and then we use the deformed surface to register the surface in further frame, assuming that adjacent frames always contains similar shapes. For real-time applications, chronological order is a hard constraint. For off-line processing, besides chronological order, surface registration of an sequence can also be applied in a global manner. Non-sequential registration can make use of the structure in the shape space to facilitate the registration. Budd et al. [2013] build a shape similarity tree structure according to volumetric histogram of spatial occupancy and then deform the template to fit a certain frame that is selected following the tree structure. Mustafa *et al*. Mustafa et al. [2016] proposed another tree structure of shape similarity that is measured from 2D image features. This non-sequential registration can minimize the total non-rigid deformation, making the method more robust to both large

coverage of shape space and large frame to frame deformation.

2.4 Clothing shape space modeling

Simulation-based modeling

To model the deformation of the clothing layer, a possible solution is direct physics-based simulation, for example with mass-spring systems Baraff and Witkin [1998], Bridson et al. [2003], continuum mechanics Volino et al. [2009], or individual yarn structures Kaldor et al. [2010]. The physical simulation models are complex and rely on numerous control parameters. Those parameters can be tuned manually, estimated from captures Stoll et al. [2010] or learned from perceptual experiments Sigal et al. [2015]. One line of works trains models on physics-based simulations using machine learning techniques, which subsequently allow for more efficient synthesis of novel 3D motion sequences of dressed humans de Aguiar et al. [2010], Guan et al. [2012], Xu et al. [2014a]. In particular, these methods learn a regression from the clothing deformation to low-dimensional parameters representing body motion. These methods allow to modify the body shape, motion, and to alter the clothing. But the main disadvantage is that the methods are limited by the quality of the simulated synthetic training data. Since the simulation of complex clothing with multiple layers remains a challenging problem, this limitation restricts the model to relatively simple clothing. Our work addresses this problem by allowing to train from both simulated and captured sequences.

Capture-based modeling

Thanks to laser scanners, depth cameras, and multi-camera system, now it's possible to capture and reconstruct 3D human motion sequences as raw mesh sequences Allain et al. [2015], Collet et al. [2015], Newcombe et al. [2015], and recent processing algorithms Neophytou and Hilton [2014], Zhang et al. [2017] allow to extract semantic information from the raw data. A recent line of work leverages this rich source of data by using captured sequences to learn the deformation of the clothing layer. Neophytou and Hilton [2014], Pons-Moll et al. [2017] (see Figure 2.3). Neophytou *et al*. Neophytou and Hilton [2014] propose a method that trains from a single subject in fixed clothing and allows to change the body shape and motion after training by using radial basis function for interpolation between given examples. Pons-Moll *et al*. Pons-Moll et al. [2017] extract the body shape and individual pieces of clothing from a raw capture sequence and use this information to transfer the captured clothing to new body shapes. Lack of analysis on the influence of human motion, the method can not transfer the captured clothing to a different pose. These methods allow to learn from complex deformations without requiring a physical model of the observations. But a common disadvantage is that the model does not allow the modification of the clothing itself, such as the fit of the clothing and the cloth material. Our work addresses this problem by exploiting self-redundancies of the deformation to build a regression model from semantic sizing or material parameters to the clothing layer.

Figure 2.3 – Top: Neophytou and Hilton [2014] can change body shape as well as pose. Bottom: Pons-Moll et al. [2017] can change body shape.

Chapter 3

Body estimation under clothing

Contents

3.1 Introduction

To make use of captured 3D data, we need to extract semantic information such as body pose and register the raw sequence to get temporal coherence. This chapter is focused on extracting human body information from raw 3D captured sequences.

It is one of the fundamental computer vision tasks to extract body information in human related application. Traditionally this information contains human position, articulated pose and more recently it also involves body shape. While the body information may be relatively easy to extract from captures of people in minimal clothes, such as swimming suit or underwear, it is not trivial when the body is covered by wide or loose clothes. And to study clothing deformation, wide clothing is inevitable. Therefore we would like to find solutions in such scenarios.

Given an input motion sequence of raw 3D meshes or oriented point clouds (with unknown correspondence information) showing a dressed person, our goal 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 Weiss et al. [2011], Helten et al. [2013] or require manual initialization for the pose Wuhrer et al. [2014], Neophytou and Hilton [2014], 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. In this chapter, 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 Wuhrer et al. [2014], Neophytou and Hilton [2014]. 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 Weiss et al. [2011], Helten et al. [2013]. 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 chapter, 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 Pishchulin et al. [2017]. 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 Zuffi and Black [2015] 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 Neophytou and Hilton [2014], Bălan and Black [2008].

In short, this chapter proposes:

- An automatic approach to estimate 3D human body shape and pose in motion with 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.

3.2 Dataset acquisition

The 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 acquired the first dataset consisting 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. This data set is used for the evaluation in this Chapter. Later, we expanded the dataset to involve 13 subjects (7 female, 6 male) with 3 motions and 4 clothing styles each. The expanded dataset is used in Chapter 5. 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 2.1. The basic motions that were captured are walk, body rotation, and knees pull up. The captured clothing styles are very tight, layered (long-sleeved layered clothing on upper body), wide (wide pants for men and dress for women) and T-shirt with shorts for the expanded dataset. Each subject has 14 markers placed on anatomically significant locations: forehead, shoulders, elbows, wrists, belly, knees, feet and heels. If these locations are covered by cloth while the subject is at resting pose then markers are placed on the clothes. The body shapes of all subjects vary significantly. Fig. 3.1 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 can be done per vertex on the two body shapes under the same normalized posture. In this chapter, we use cumulative plots to show the results.

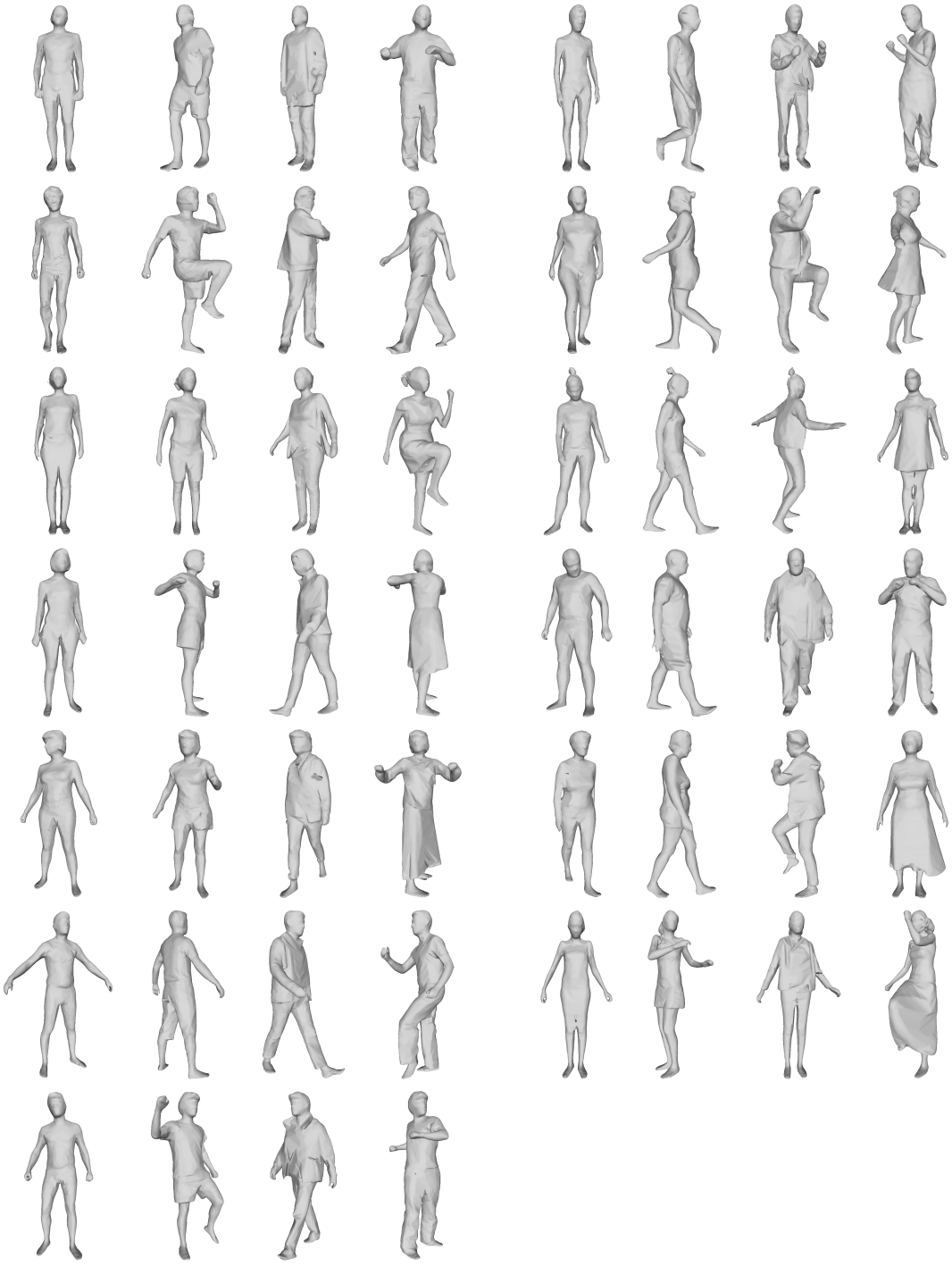


Figure 3.1 – Representative examples of our motion database. Each row shows two subject dressed in tight clothing, T-shirt and shorts, layered clothing and wide clothing respectively.

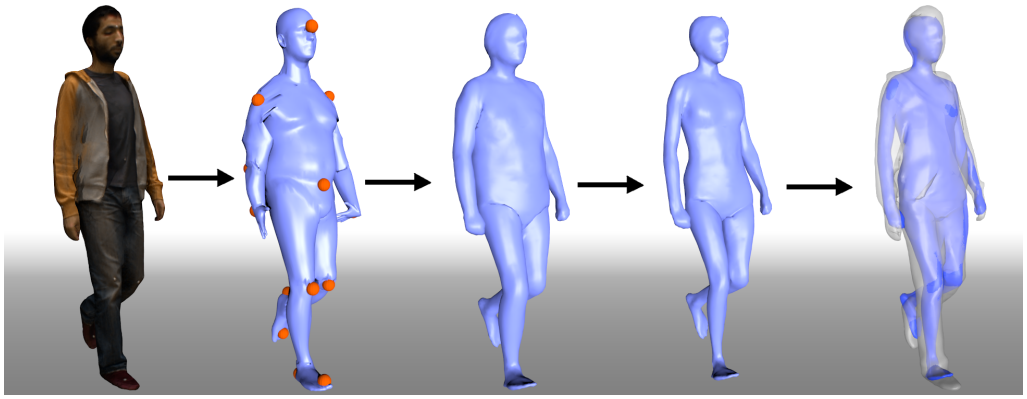


Figure 3.2 – Overview of the proposed pipeline. From left to right: input frame, result of Stitched Puppet Zuffi and Black [2015] with annotated landmarks, result after estimation of initial identity and posture, final result, and overlay of input and final result.

3.3 Estimating body model parameters for a motion sequence

We start by providing an overview of the proposed method. Fig. 3.2 shows the different parts of the algorithm visually. Given as input a trained S-SCAPE model and a motion sequence consisting of N_f frames $\mathcal{S}_i$ 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 $\mathcal{S}_i$.

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

$$E(\mathcal{S}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \omega_{lnd} E_{lnd}(\mathcal{S}, \boldsymbol{\beta}, \boldsymbol{\theta}) + \omega_{data} E_{data}(\mathcal{S}, \boldsymbol{\beta}, \boldsymbol{\theta}) + \omega_{cloth} E_{cloth}(\mathcal{S}, \boldsymbol{\beta}, \boldsymbol{\theta}) \tag{3.1}$$

w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ subject to constraints that keep $\boldsymbol{\beta}$ in the learned probability

distribution of parameter values. Here, ω_{lnd}, ω_{data}, and ω_{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 $\mathcal{S}$ 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 $\mathcal{S}$.

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

$$E\left(\mathcal{S}_{1:N_f}, \boldsymbol{\beta}, \boldsymbol{\theta}_{1:N_f}\right) = \sum_{i=1}^{N_f} E\left(\mathcal{S}_i, \boldsymbol{\beta}, \boldsymbol{\theta}_i\right) \tag{3.2}$$

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, $\mathcal{S}_{1:N_f} = \{\mathcal{S}_1, \ldots, \mathcal{S}_{N_f}\}$ 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. 3.2 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.

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 σ_i along the i-th principal component. Similarly to previous work Bălan and Black [2008], 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.

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

$$E_{lnd}\left(\mathcal{S}, \boldsymbol{\beta}, \boldsymbol{\theta}\right) = \sum_{i=1}^{N_{lnd}} \left\| \boldsymbol{s}_i\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right) - \boldsymbol{l}_i(\mathcal{S}) \right\|^2, \tag{3.3}$$

where $\boldsymbol{l}_i(\mathcal{S})$ denotes the i-th landmark of frame $\mathcal{S}$, $\boldsymbol{s}_i\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right)$ denotes the vertex corresponding to the i-th landmark of $\boldsymbol{s}\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right)$, and $\|\cdot\|$ denotes the

ℓ^2 norm.

The landmarks $\boldsymbol{l}_i(\mathcal{S})$ are computed automatically with the help of the state of the art Stitched Puppet Zuffi and Black [2015], 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 $\mathcal{S}$, and report the 3D positions of the pre-selected indices after fitting as landmarks $\boldsymbol{l}_i(\mathcal{S})$. While the Stitched Puppet aims to fit the body shape and posture of $\mathcal{S}$, only the coordinates $\boldsymbol{l}_i(\mathcal{S})$ are used by our framework. Note that our method does not require accurate $\boldsymbol{l}_i(\mathcal{S})$, since $\boldsymbol{l}_i(\mathcal{S})$ 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. 3.3. 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.

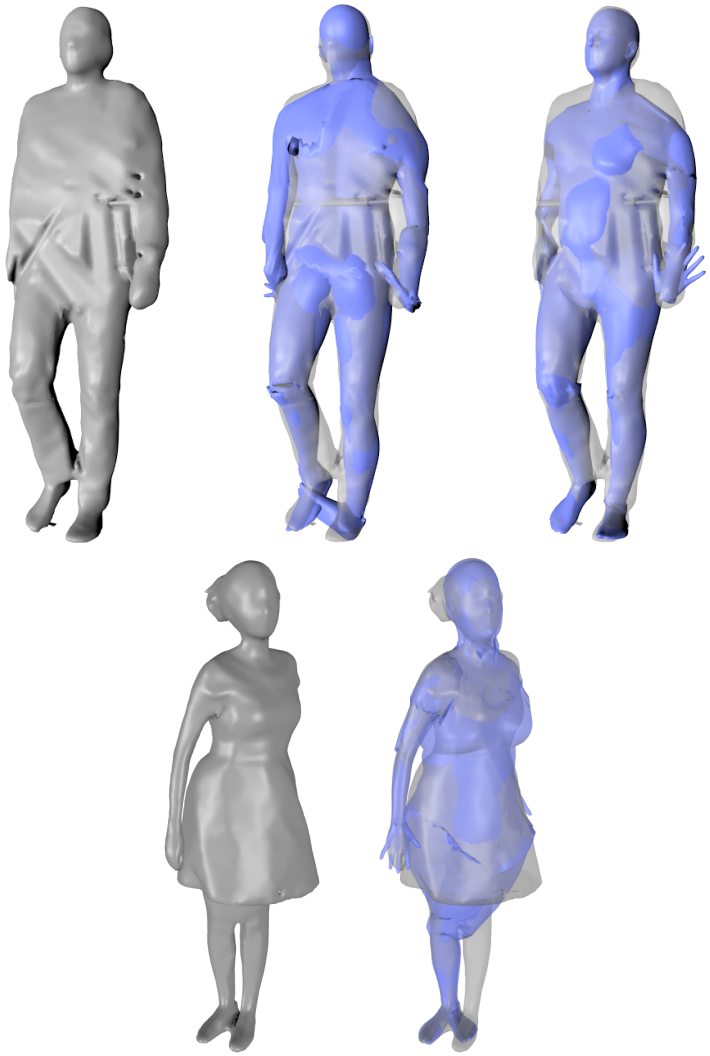


Figure 3.3 – Top: overfitting problem of Stitched Puppet in the presence of clothing. Input frame, Stitched Puppet result with 160 particles, and Stitched Puppet result with 30 particles are shown in order. Bottom: the failure case from our database caused by mismatching of Stitched Puppet.

Data energy

The data energy pulls the S-SCAPE model towards the observation $\mathcal{S}$ 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 $\mathcal{S}$ as

$$E_{data}\left(\mathcal{S},\boldsymbol{\beta},\boldsymbol{\theta}\right)=\sum_{i=1}^{N_v}\delta_{\boldsymbol{NN}}\left\|\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\theta}\right)-\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\theta}\right),\mathcal{S}\right)\right\|^2, \tag{3.4}$$

where N_v denotes the number of vertices of $\boldsymbol{s}$ and $\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\theta}\right),\mathcal{S}\right)$ denotes the nearest neighbour of vertex $\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\theta}\right)$ on $\mathcal{S}$. 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 $\mathcal{S}$ is below $200mm$ and the angle between their outer normal vectors is below $60°$, and to zero otherwise.

Clothing energy

The clothing energy is designed to encourage the predicted body shape $\boldsymbol{s}$ to be located entirely inside the observation $\mathcal{S}$. 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 $\mathcal{S}$ rather than to fit inside of $\mathcal{S}$, see Fig. 3.4. To remedy this, we define the clothing energy as

$$\begin{aligned}E_{cloth}\left(\mathcal{S},\boldsymbol{\beta},\boldsymbol{\theta}\right)=&\sum_{i=1}^{N_v}\delta_{out}\delta_{\boldsymbol{NN}}\left\|\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\theta}\right)-\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\theta}\right),\mathcal{S}\right)\right\|^2\\&+\omega_r\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\right\|^2,\end{aligned} \tag{3.5}$$

where δ_{out} is used to identify vertices of $\boldsymbol{s}$ located outside of $\mathcal{S}$. This is achieved by setting δ_{out} to one if the angle between the outer normal of $\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\theta}\right),\mathcal{S}\right)$ and the vector $\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\theta}\right)-\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\theta}\right),\mathcal{S}\right)$ is below $90°$, and to zero otherwise. Furthermore, ω_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 $\mathcal{S}_i$. This allows to account for clothing using a simple optimization without the need to find skin and non-skin regions as in previous work Bălan and Black [2008], Stoll et al. [2010], Neophytou and Hilton [2014]. The regularization $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2$ used in Eq. 3.5 is required to avoid excessive thinning of limbs due to small mis-alignments in posture.

Fig. 3.4 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. 3.4 is visualized in the middle and the right of Fig. 3.4, and shows that E_{cloth} leads to a significant improvement of the accuracy of $\boldsymbol{\beta}$.

Optimization schedule

Minimizing $E\left(\mathcal{S}_{1:N_f}, \boldsymbol{\beta}, \boldsymbol{\theta}_{1:N_f}\right)$ defined in Eq. 3.2 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

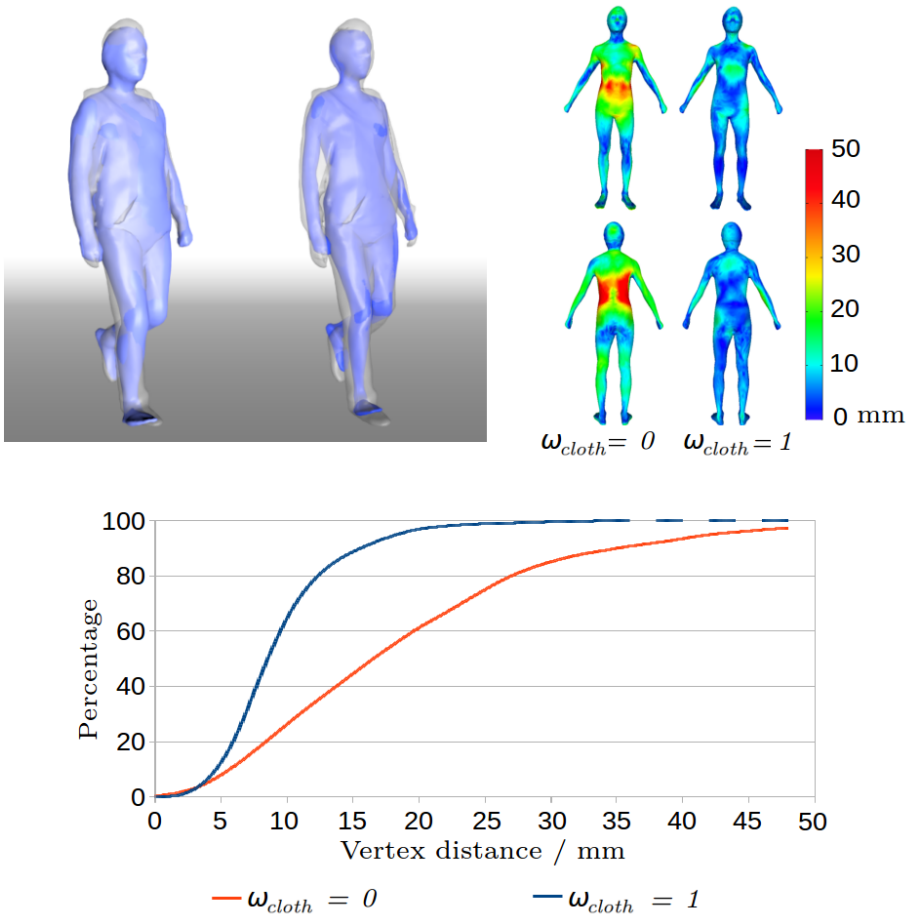


Figure 3.4 – Influence of E_{cloth} on walking sequence. Top left: input data overlayed with result with $\omega_{cloth} = 0$ (left) and $\omega_{cloth} = 1$ (right). Top right: color-coded per-vertex error with $\omega_{cloth} = 0$ (left) and $\omega_{cloth} = 1$ (right). Bottom: cumulative per-vertex error of estimated body shape with $\omega_{cloth} = 0$ and $\omega_{cloth} = 1$.

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 $E\left(\mathcal{S}_{1:N_k}, \boldsymbol{\beta}, \boldsymbol{\theta}_{1:N_k}\right)$ 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. 3.1 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 ω_{lnd} to one. We linearly reduce ω_{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 $E\left(\mathcal{S}_{1:N_k}, \boldsymbol{\beta}, \boldsymbol{\theta}_{1:N_k}\right)$ 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. 3.1 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 ω_{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 $\mathcal{S}_{1:N_f}$. To this end, we initialize the identity parameters to $\boldsymbol{\beta}_0$, fix all posture parameters to the computed $\boldsymbol{\theta}_i$, and minimize $E\left(\mathcal{S}_{1:N_f}, \boldsymbol{\beta}, \boldsymbol{\theta}_{1:N_f}\right)$ 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.

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 Pishchulin et al. [2017].

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 Wuhrer et al. [2014], and are located at forehead, shoulders, elbows, wrists, knees, toes, heels, and abdomen. Fig. 3.2 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 ω_r depends on the clothing style. The looser the clothing, the smaller ω_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.

3.4 Evaluation

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. 3.3). We exclude this sequence from the following evaluation.

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. 3.5 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.

To evaluate the accuracy of the identity parameters $\boldsymbol{\beta}$, we use for each

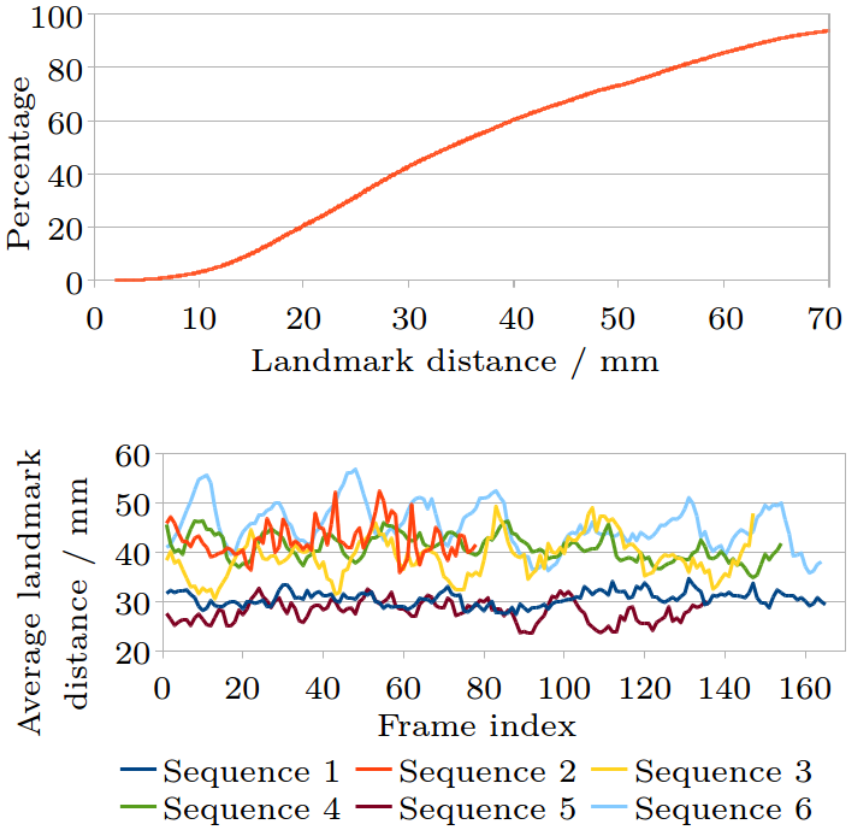


Figure 3.5 – Accuracy of posture estimation over the walking sequences in tight clothing. Top: cumulative landmark errors. Bottom: average landmark error for each sequence.

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. 3.6 summarizes the per-vertex errors over all motion sequences captured in layered and wide clothing, respectively. The left side shows

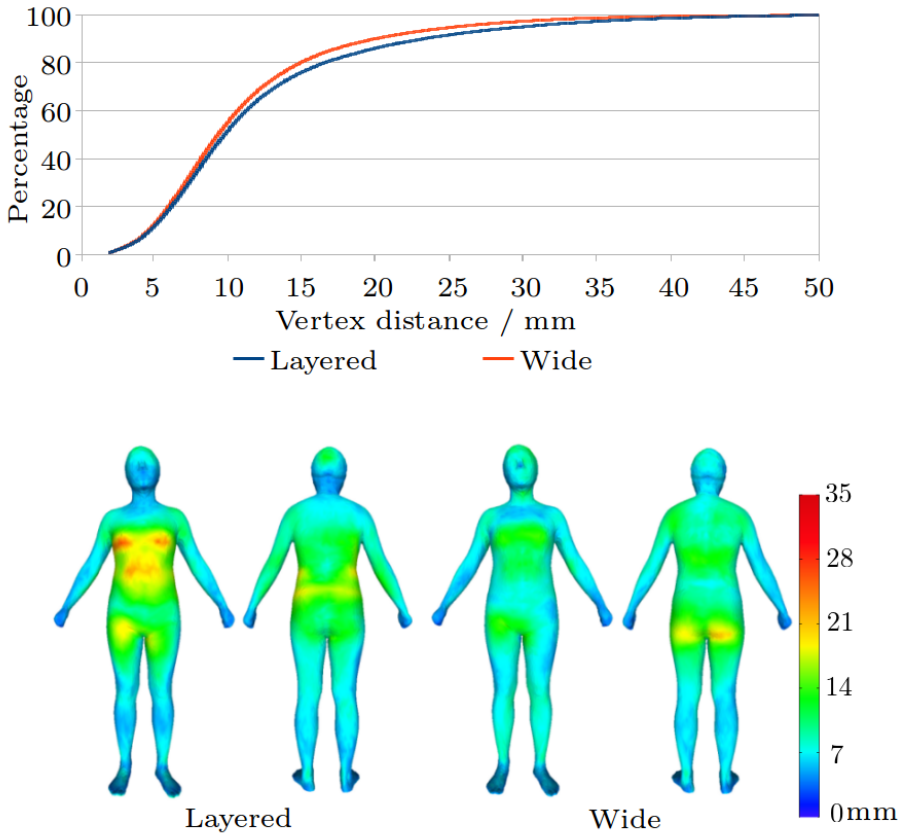


Figure 3.6 – Summary of shape accuracy computed over the frames of all motion sequences of all subjects captured in layered and wide clothing. Top: cumulative plots showing the per-vertex error. Bottom: mean per-vertex error color-coded from blue to red.

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.

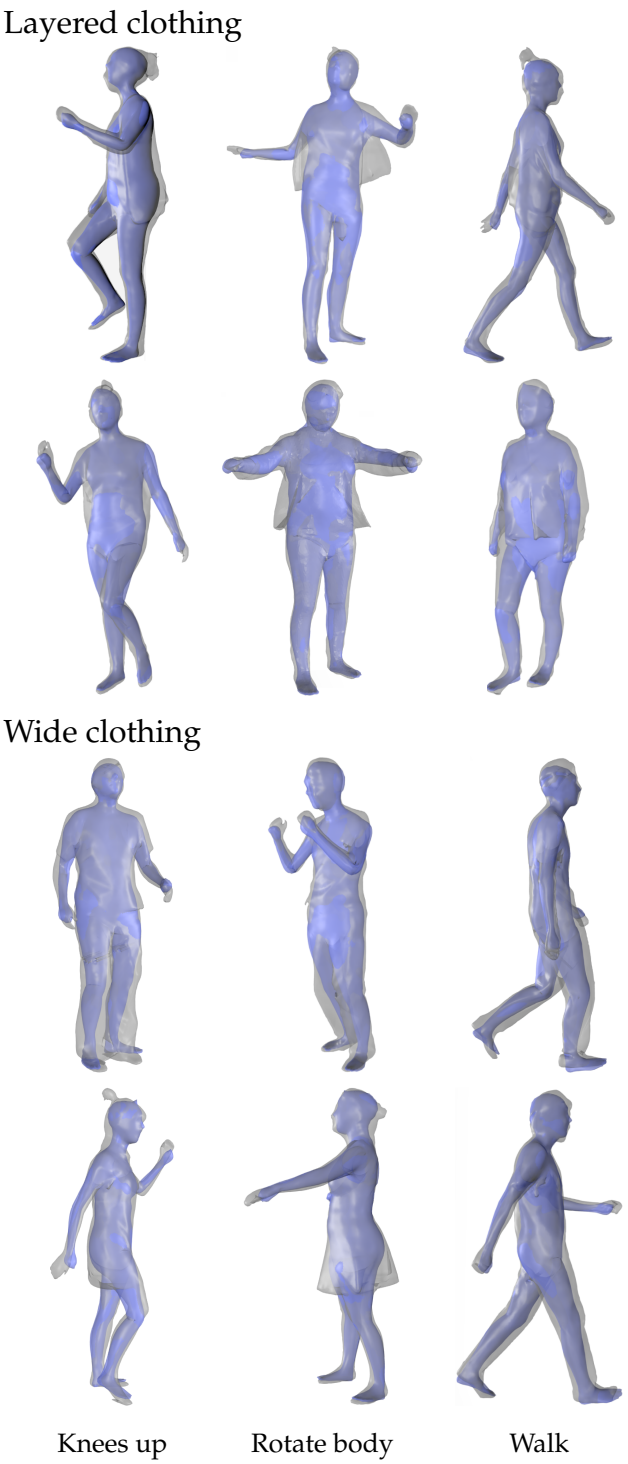
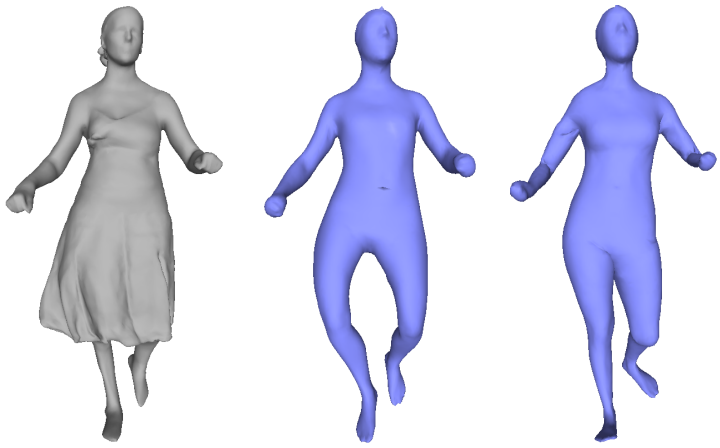


Figure 3.7 – Overlay of input data and our result.

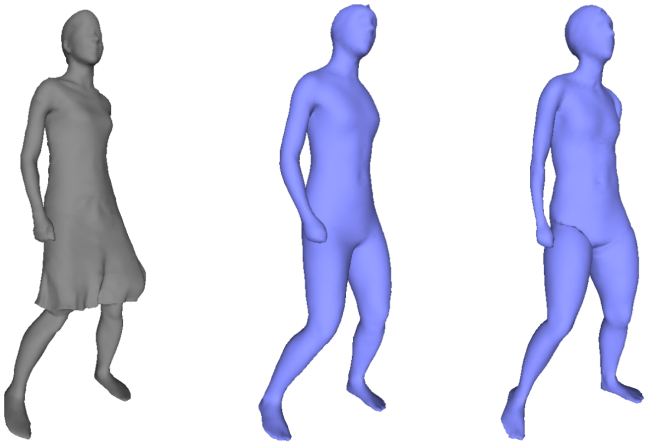
Fig. 3.7 shows some qualitative results for all three types of motions and two clothing styles. Note that accurate body shape estimates are obtained for all frames. Consider the frame that shows a female subject performing a rotating motion in layered clothing. Computing a posture or shape estimate based 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.

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. 3.8. We compare to Wuhrer et al. Wuhrer et al. [2014] on the dancer sequence De Aguiar et al. [2008] 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 Neophytou and Hilton [2014] on the swing sequence Vlasic et al. [2008] 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.



Comparison to Wuhrer et al. [2014]



Comparison to Neophytou and Hilton [2014]

Figure 3.8 – Per comparison from left to right: input, result of prior works, our result.

3.5 Conclusion

In this chapter, we have 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. It contains 6 different subjects performing 3 motions in 3 different styles each and

is expanded later to involve another 7 subjects and one more clothing styles for each. We have shown that, although being fully automatic, our posture and shape estimation achieves state-of-the-art performance. With this method, we can extract body information from 3D clothed human shape, which provides fundamental variables for further clothing shape space analysis.

The proposed method also suffers from certain limitations. The most significant one is that the automation is essentially supported by the Stitched Puppet algorithm for pose initialization. However, as shown in Figure 3.3 the Stitched Puppet is not designed for clothed human, therefore it has certain probability to fail in practice. A more robust pose detector is preferred. Some deep learning approaches can be possible candidates, including some of the state-of-the-art work such as Cao et al. [2016] and Zhou et al. [2016]. Although instead of using 3D input directly, these deep learning work usually work on 2D images, it is not difficult to project 3D input to 2D images and then retrieve the 3D pose information afterwards. However, the robustness of these candidates against loose clothes still needs to be tested. Apart from this, the body model we choose also simplifies the head, the hands and the feet. A more elaborate model, for example, the Frank or the Adam model Joo et al. [2018] can potentially capture more details.

In the end, we should also mention that Zhang et al. [2017] solve very similar problems as in this chapter. They make use of another body model, the SMPL model, and enhance the face by allowing certain free vertex displacements so that the method can also capture the subject's face.

Chapter 4

Non-rigid registration of clothing

Contents

4.1 Introduction

In the previous chapter, we have proposed a method to extract human body information from raw 3D acquisitions. In this chapter, we will

continue focusing on preprocessing of raw 3D acquisitions, aiming at temporal coherence.

Captured raw 3D sequences contain individual 3D reconstructions at each frame, with no temporal coherence between adjacent frames. An important and challenging task in 3D computer vision is to perform non-rigid registration of the object surface through the sequence, i.e. expressing the entire 3D sequence as the deformation of a single deformable template mesh. This representation has large benefits as it allows for efficient storage, transmission, reverse scene analysis and semantic characterization of the scene as one moving object. Particularly in this thesis, such process will provide us with the trajectory of each point on the clothing surface and subsequently make it possible to analyze the clothing deformation and compress the shape space of dressed humans (in Chapter 5).

How to obtain this non-rigid registration has been widely studied, especially in the context of humans with relatively tight or rigid clothing, where only the motion of the body needs to be characterized. This yields a large family of template-based surface registration methods, which follow a similar resolution canvas. First a surface template is chosen and obtained, that can be *e.g.* a generic human model, a particular reconstruction among those observed, or a pre-obtained scan of the human subject. Second, a point-to-point correspondence scheme is devised, using point proximity in Euclidean space, in a geometric or appearance feature space, or on a learned manifold. Third, because the correspondences so obtained are often insufficiently dense, non-uniformly distributed over the body, and erroneous, a deformation model with a reduced control parameter set

is used to constrain the estimation of the full body alignment and reduce the search space of deformation, typically based on human kinematics, piece-wise rigid or affine motion.

In case of loose clothing acquisitions, tight clothing alignment method assumptions generally fail. In particular the correspondence becomes much more challenging since such sequences exhibit stronger geometric and topological noise, more occlusions, and much larger and more widely spread non-rigid deformations, due to the inherent geometric variability of clothing. The sensitivity of correspondences to the representativeness of template topology and geometry is also drastically increased.

Existing work on surface registration of dressed human, as reviewed in details in Section 2.3, mainly suffer from two aspects. On the one hand, some state-of-the-art work on non-rigid registration, such as Dou et al. [2016], Newcombe et al. [2015] and Yu et al. [2018], can follow the detailed geometric deformation very well, but instead of maintaining a complete surface template through the sequence, they need to constantly refresh the template. On the other hand, the existing work on 3D human performance acquisition and surface registration, such as De Aguiar et al. [2008], Vlasic et al. [2008], Allain et al. [2014] and Allain et al. [2015], are able to produce a single deformable shape through the sequence, but does not follow the surface geometry very well.

In this chapter, we will address this problem with a full registration solution, centered on a correspondence model suitable for clothes. We ground this in two key assumptions. First, we assume that the topology of the human in clothing is fixed over time. This assumption holds for

the human body itself as well as many clothing styles such as t-shirts, trousers and skirts. Second, we assume that the geometry of the model deforms in a near-isometric way. This is true when considering local motions of the human and clothing that are not very elastic. In practice, the acquired object violates both assumptions due to aforementioned acquisition noise. That is, the acquired topology may change due to the merging of close-by body parts, which in turn completely changes the intrinsic geometry of the model. To design a method that is robust to this type of acquisition noise, we use a deformation model based on *partial* near-isometric patches. These patches are grown from a set of selected landmarks that can be robustly found in each frame, and ensure consistent densification of correspondences over the surface. To ease correspondence over all frames, we also provide an automatic template selection method among raw 3D models of the input sequence, maximizing topological adequacy.

The experiments demonstrate the success of the method for non-rigid registration of a variety of clothing sequences. In particular we favorably compare our method against two state of the art temporal alignment methods Allain et al. [2014, 2015] based on more restrictive locally rigid deformation assumptions, on both pre-registered and raw captured datasets. Results show that our method outperforms previous work for human characters with loosing clothing.

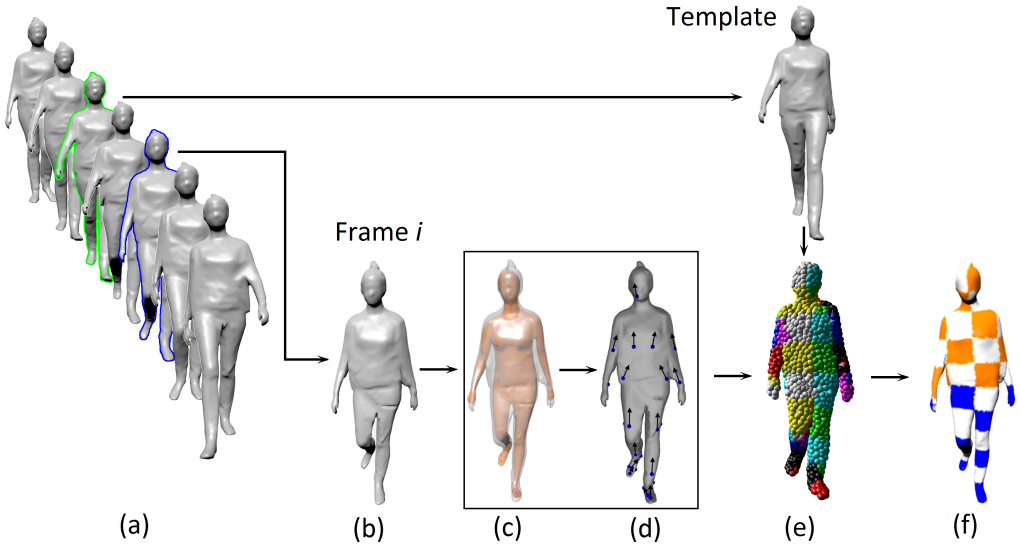


Figure 4.1 – Method overview. Given an input sequence of 3D meshes shown in (a), each frame is processed in three steps, which are shown for the frame indicated in blue and shown in (b). First, a statistical model of undressed body shape is fitted to the frame (c). Pre-marked anatomical points on the fitted human body model are mapped to the input frame (d). Second, these anatomical markers are used to guide a partial near-isometric correspondence computation between an automatically selected template (green model in (a), and shown in top row) and the frame (e). Third, the resulting correspondences are used as assignments to deform the template to the input frame (f).

4.2 Method overview

Given an unstructured temporal sequence showing a clothed human performing a motion, our method produces a consistently deforming model. To be robust to acquisition and topological noise our method is divided into three steps as shown in Figure 4.1: a human body model is used to find good initial starting points for the partial near-isometries; point trajectories are computed using a partial near-isometric deformation model; and the point trajectories are used as assignments in a template-

based deformation to find a coherent model that deforms over time. We now provide more detail for each step.

First, we estimate the undressed human body shape in motion for the given sequence. This step reduces significantly the search space of the problem by providing a good initialization for the computation of near-isometric partial matches. This step takes advantage of recent robust methods that use statistical human body models to estimate the naked human body shape under clothing Jain et al. [2010], Wuhrer et al. [2014], Neophytou and Hilton [2014], Zhang et al. [2017]. In our implementation, we use the method proposed in previous chapter. For the specific frame that is shown in blue in Figure 4.1(a) and enlarged in (b), the estimated human body shape is shown in Figure 4.1(c). The input frame is shown in grey and the estimated human body under clothing in brown. Next, we transition from the naked human body to surface of the clothing. We once manually select a small set of anatomical points on the statistical model and automatically transfer these points on the clothed sequence as shown in Figure 4.1(d).

Second, we use a partial near-isometric deformation model that is robust with respect to topological acquisition noise to compute point trajectories on the input sequence. Following Letouzey and Boyer [2012], we proceed by first taking advantage of the assumption that the real topology stays fixed over time to automatically select a template frame from the sequence based on topological and geometric criteria. The automatically selected template is shown in green in Figure 4.1(a) and enlarged in the first row. Second, for each frame of the sequence, we indepen-

dently compute correspondence information between the template and the frame. This computation finds partial near-isometric matches between two frames Brunton et al. [2014]. To increase robustness and make the algorithm efficient, we use the previously computed anatomical points on each frame for initialization. The partial matches are then merged into a global correspondence. The computed correspondence between template and a frame is shown in Figure 4.1(e). This correspondence information between the template and every frame of the sequence results in point trajectories over time.

Third, we find a consistent topology that deforms over time by extending a standard template fitting technique Allen et al. [2003] to take advantage of the previously computed point trajectories and operate on motion sequences. Since the previous step computes the correspondence information between the template and every other frame independently, the resulting point trajectories may be interrupted and the correspondence of a particular point on the template may not be known on every frame. To remedy this we add this final step that deforms the template to the entire motion sequence using previously computed point trajectories as assignments. The output of this step is shown in Figure 4.1(f).

This work is a joint work with another PhD student, Aurela Shehu. The template selection and the near-isometric matches were mostly achieved by Aurela Shehu; the mapping of anatomical points and the template deformation were mostly implemented by myself.

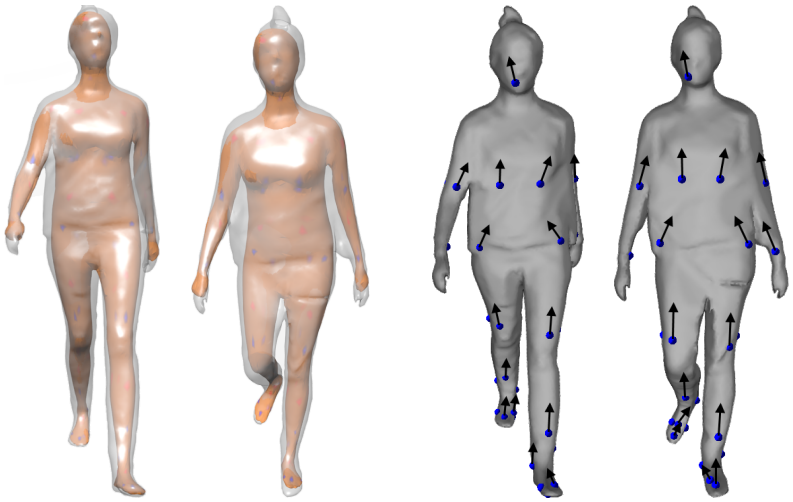


Figure 4.2 – Left: estimated human body $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}_i)$ shown in brown overlaid with two input frames. Right: oriented anatomical points on two meshes of a sequence. The points are shown in blue, and their orientations in black.

4.3 Non-rigid registration of clothed human

This section provides details on the three steps of our proposed method.

Human body estimation and mapping anatomical points to clothed human

We start by estimating the human body shape for the given input sequence. The result of this step allows to map a sparse set of anatomical points to the input frames of a human in wide clothing.

Given a human motion sequence in wide clothing, we start by estimating the undressed human body shape for the entire sequence using the method proposed in previous chapter.

Figure 4.2 shows the estimated human body shape and posture for

two frames of a sequence. Note that all meshes $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ generated this way share the same vertex ordering and mesh topology.

We manually select a small set of oriented anatomical points on the statistical model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$, where an oriented point is a 3D location on the surface along with a direction in its tangent plane. An oriented point is in practice selected by choosing two points: a starting point and a close-by neighbor that defines the direction in the tangent plane. Note that these points only need to be chosen once for any $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, as all body models share the same mesh structure.

To automatically map an oriented anatomical point $\mathbf{t}$ on $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}_i)$ to the clothing surface of the i-th frame $\mathcal{S}_i$, we intersect the line through $\mathbf{t}$ along the normal direction of $\mathbf{t}$ with $\mathcal{S}_i$. If there is any intersection outside of $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}_i)$, and the distance of the one closest to $\mathbf{t}$ is within a threshold τ_o, this intersection point is considered a valid mapping. Otherwise we look for the intersection that lies inside $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}_i)$, and chose the closest one to be a valid mapping if the distance to $\mathbf{t}$ is smaller than τ_i. In case no valid mapping is found, $\mathbf{t}$ is removed from consideration for $\mathcal{S}_i$.

An example of all valid oriented anatomical points on two frames of a sequence is shown in Figure 4.2, where points are shown in blue and orientations in black.

Computation of point trajectories

Here we describe how to compute point trajectories given the input sequence and oriented anatomical points. First, we automatically select a

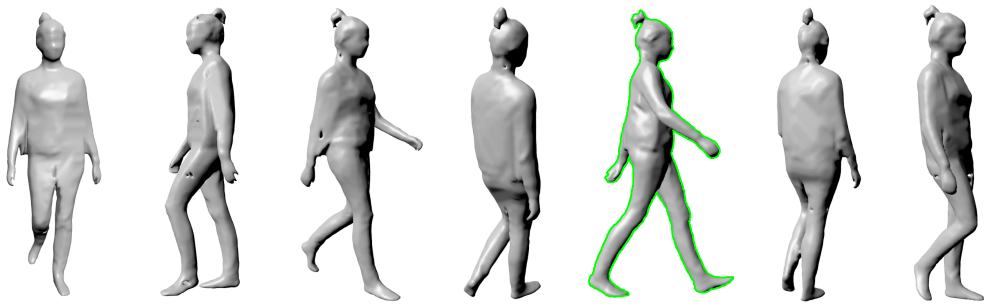


Figure 4.3 – Some frames of a sequence and the computed template $\mathcal{T}$ highlighted in green.

template from all frames of the sequence. Second, we establish a dense correspondence between the template and each frame of the sequence with the help of a partial near-isometric deformation model.

Automatic template selection Often in an acquired sequence the topology changes over time due to acquisition noise. In particular, tiny holes frequently appear and large contacts appear when limbs are close-by, see Figure 4.3 for an example. To remedy this problem during alignment, we automatically detect in the sequence a template $\mathcal{T}$ and register $\mathcal{T}$ to every other frame $\mathcal{S}$ of the sequence.

The template selection is based on topological and geometric criteria. Since we assume that the scene has a fixed topology and that observed topology changes come from acquisition artifacts, it follows that observed topological properties of the shape can only grow Letouzey and Boyer [2012]. That is, observed splits are accepted as changes of the real shape while observed merges are ignored from consideration as the real shape cannot merge. To account for acquisition artifacts, splits are only considered if they are persistent over time and small components are filtered.

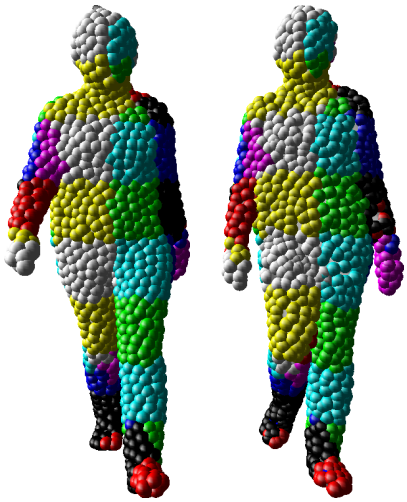


Figure 4.4 – Color-coded correspondence information computed between $\mathcal{T}$ (right) and $\mathcal{S}$ (left) using a partial near-isometric deformation model.

As observed splits are accepted, we first select as candidate templates all frames that have a maximum number of components. As observed splits and merges are directly linked to the genus number, we further select from the candidate templates the frames with minimum genus of the largest component.

The aforementioned topological criteria might provide several candidate templates. The final selection is based on a geometric quality criterion. From the candidate templates, we select the one with minimum area ratio (maximum point area/minimum point area, where a point area is the area of the Voronoi region around a vertex) of the largest component, as the rest of our pipeline benefits from a template whose vertices are as uniform in area as possible. Figure 4.3 shows the selected template $\mathcal{T}$ for one of our test sequences.

Dense correspondence computation To register $\mathcal{T}$ to a frame $\mathcal{S}$, we use a partial near-isometric deformation model that is robust to topological acquisition noise. Near-isometric models have previously been used successfully for cloth modeling Popa et al. [2009]. Ideally, $\mathcal{T}$ and $\mathcal{S}$ should be mapped by a global near-isometric mapping. In practice, due to acquisition noise, such a global mapping may not exist.

To remedy this, we use a partial near-isometric model to account for topological acquisition noise as in Brunton et al. [2014]. To this end, we consider every frame $\mathcal{S}$ to be a set of smooth, orientable 2-manifolds embedded in three-dimensional space. In practice, $\mathcal{S}$ is discretized by a set of points that are connected by a neighborhood graph. We denote the geodesic distance between two points $\mathbf{s}_i, \mathbf{s}_j \in \mathcal{S}$ by $d_{\mathcal{S}}(\mathbf{s}_i, \mathbf{s}_j)$.

Consider a mapping $f : \mathcal{U} \rightarrow \mathcal{S}$, where $\mathcal{U}$ is a subset of $\mathcal{T}$. The mapping function f is near-isometric if:

$$|d_{\mathcal{U}}(\mathbf{t}_i, \mathbf{t}_j) - d_{\mathcal{S}}(f(\mathbf{t}_i), f(\mathbf{t}_j))| \leq \epsilon \tag{4.1}$$

where $\mathbf{t}_i, \mathbf{t}_j$ are vertices in $\mathcal{U}$ and ϵ refers to the allowed stretching threshold. Since $\mathcal{U} \subset \mathcal{T}$ refers to a shape part, we seek parts of $\mathcal{T}$ that can be mapped to parts of $\mathcal{S}$ without much stretching.

Brunton et al. [2014] use the partial near-isometry model for pairwise frame alignment. They show that a correspondence between an oriented point on $\mathcal{T}$ and an oriented point on $\mathcal{S}$ is sufficient to recover an isometric mapping. In the following, we use $\mathbf{t}$ and $\mathbf{s}$ to denote both points and oriented points on $\mathcal{T}$ and $\mathcal{S}$, respectively. We use the previously computed oriented anatomical points, whose correspondence information

is known across frames. We use them as starting points for the partial near-isometric correspondence computation between $\mathcal{T}$ and $\mathcal{S}$. To speed up the computation time, we minimize point correspondence information coming from many overlapping partial near-isometries that is discovered from nearby starting points. For this, we stop when the discovered near-isometric part has an intrinsic radius bigger than a distance threshold τ_d. To increase robustness, we ignore from consideration near-isometric parts that have an intrinsic radius smaller than a threshold τ_s.

After finding multiple near-isometric parts between $\mathcal{T}$ and $\mathcal{S}$, we merge the parts into a global alignment by assigning to each $\mathbf{t}$ in $\mathcal{T}$ that is mapped to at least one point on $\mathcal{S}$ the geodesic average of all computed assignments on $\mathcal{S}$. Figure 4.4 shows a color-coded alignment computed between $\mathcal{T}$ and $\mathcal{S}$ for one of the test sequences.

After this step, we have computed pairwise correspondence information between $\mathcal{T}$ and every frame $\mathcal{S}$ of the sequence. This allows us to follow the trajectory of a particular point on $\mathcal{T}$ over time.

Deform template to register the surface

In the previous step, we computed frame correspondences between $\mathcal{T}$ and every other frame $\mathcal{S}$. To find a consistent topology that deforms over time we extend a standard template fitting technique Allen et al. [2003] to take advantage of the previously computed point trajectories. That is, the previously computed frame correspondence serves as assignment information to guide the template deformation to allow for fast motion and reduce drift.

In particular, let $\mathcal{T}$ contain $N_{\mathcal{T}}$ vertices, and let the motion sequence contain N_f frames $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{N_f}$. Without loss of generality, let $\mathcal{T}$ be frame $\mathcal{S}_j$. We fit the frames of the sequence starting at $\mathcal{T}$ by processing adjacent frames in (backward and forward temporal) order as $\mathcal{S}_{j-1}, \ldots, \mathcal{S}_1$ and $\mathcal{S}_{j+1}, \ldots, \mathcal{S}_{N_f}$. After $\mathcal{S}_i$ has been fitted, we update $\mathcal{T}$ to the fitting result of $\mathcal{S}_i$. This initializes the location and shape of the template to be close to the next frame to be fitted, thereby allowing for a simple template fitting scheme.

The template fitting follows a standard technique Allen et al. [2003], and fits $\mathcal{T}$ to $\mathcal{S}$ by optimizing the following energy

$$E_{template} = w_c E_c + w_d E_d + w_s E_s + w_t E_t, \tag{4.2}$$

which consists of the weighted linear combination of four simple energy terms: correspondence energy E_c, nearest neighbor energy E_d, deformation smoothness energy E_s, and temporal smoothness energy E_t, which are weighted by w_c, w_d, w_s, and w_t, respectively. To model the deformation, each vertex $\mathbf{t}_i$ of $\mathcal{T}$ is expressed in homogeneous coordinates and transformed using a 4×4 matrix $\boldsymbol{A}_i$ that represents an affine transformation in $\mathbb{R}^3$.

The first energy term modifies the marker energy proposed by Allen *et al*. Allen et al. [2003] to use the previously computed correspondences from $\mathcal{T}$ to $\mathcal{S}$ as assignments. The energy is expressed as

$$E_c = \frac{1}{\sum_{i=1}^{N_{\mathcal{T}}} w_{corr,i}} \sum_{i=1}^{N_{\mathcal{T}}} w_{corr,i} \| \boldsymbol{A}_i \mathbf{t}_i - \mathbf{s}_c(\mathbf{t}_i) \|_2^2, \tag{4.3}$$

where $w_{corr,i}$ is a weight equal to 1 if $\mathbf{t}_i$ has a correspondence on $\mathcal{S}$ and equal to 0 otherwise, $\mathbf{s}_c(\mathbf{t}_i)$ is the point of $\mathcal{S}$ that corresponds to $\mathbf{t}_i$ and

$\|\cdot\|_2$ is the Euclidean distance. Using this energy guides the deformation in case of large motion between adjacent frames and reduces drift. Note that unlike the marker term, our energy does not rely on any manually provided information on $\mathcal{S}$, but takes advantage of the correspondences found using the partial near-isometric deformation model.

The second energy term is a standard data term that pulls each $\mathbf{t}_i$ to its nearest neighbor on $\mathcal{S}$ as

$$E_d = \frac{1}{\sum_{i=1}^{N_\mathcal{T}} w_{NN,i}} \sum_{i=1}^{N_\mathcal{T}} w_{NN,i} \|\boldsymbol{A}_i \mathbf{t}_i - \mathbf{s}_n(\mathbf{t}_i)\|_2^2, \tag{4.4}$$

where $w_{NN,i}$ is a weight equal to 1 if the nearest neighbor is valid and 0 otherwise, and $\mathbf{s}_n(\mathbf{t}_i)$ is the point on $\mathcal{S}$ that is the nearest neighbor of $\mathbf{t}_i$. We consider the nearest neighbor valid if the surface normals at each point are less than 90° apart and the distance between two points is at most $0.2m$. When used without the correspondence energy (Equation (4.3)), this nearest neighbor energy is known to suffer from drift Budd et al. [2013]. To avoid this problem, we only activate this energy once the deformed template is close to $\mathcal{S}$, thereby reducing drift.

The third energy term is the standard deformation smoothness energy

$$E_s = \frac{1}{\sum_{i=1}^{N_\mathcal{T}} |\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|\boldsymbol{A}_i - \boldsymbol{A}_j\|_F^2, \tag{4.5}$$

where $\mathcal{N}(i)$ is the 1-ring neighborhood of $\mathbf{t}_i$ and $\|\cdot\|_F$ is the Frobenius norm. This term encourages a smooth deformation field across $\mathcal{T}$.

As the fourth term, we add a simple temporal smoothness energy to

prevent very large deformations between adjacent frames as

$$E_t = \frac{1}{N_\mathcal{T}} \sum_{i=1}^{N_\mathcal{T}} \|\boldsymbol{A}_i - \boldsymbol{I}\|_2^2, \tag{4.6}$$

where $\boldsymbol{I}$ is the identity matrix. This energy discourages large displacements of individual vertices between adjacent frames and helps to prevent jittering in case of slightly inaccurate correspondences between $\mathcal{T}$ and $\mathcal{S}$.

Figure 4.5 shows an example of a template-fitted frame for one of our sequences, and Figure 4.6 shows the template fitting with and without E_c, for an example where the target frame is only four frames away from $\mathcal{T}$.

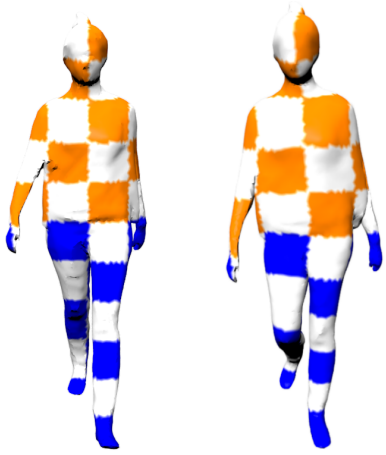


Figure 4.5 – Template $\mathcal{T}$ (left) and template-fitted frame $\mathcal{S}$ (right). Correspondence is color-coded.

Implementation Details

For the human body estimation under clothing, we use the method proposed in the previous chapter. We manually select 40 oriented anatomical points on the statistical model and map these points to the clothed sequence. To map points from human body to clothed human we set $\tau_i = 0.2m$ and $\tau_o = 0.05m$.

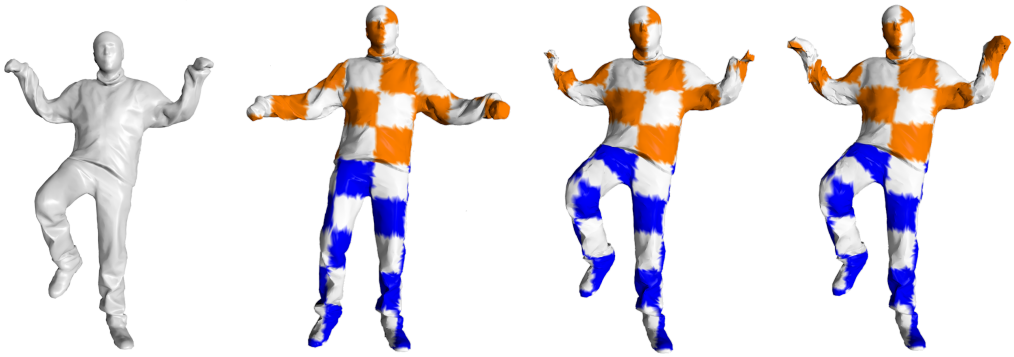


Figure 4.6 – Our E_c energy term prevents loss of tracking. From left to right: $\mathcal{S}$, $\mathcal{T}$, template deformation without E_c, and with E_c.

To compute $\mathcal{T}$ over a given sequence we do not count as components shells with an area smaller than $0.1m^2$. To compute dense correspondence we use the code of Brunton *et al*. Brunton et al. [2014] with $\epsilon = 0.25m$. To increase robustness and computational efficiency, we set $\tau_s = 0.1m$ and $\tau_d = 0.25m$.

The weights (w_c, w_d, w_s, w_t) in the template fitting step are fixed by solving the optimization problem in several stages with different energy weights at each stage as proposed by Allen *et al*. Allen et al. [2003]. So that when the template is far away from the target frame, correspondence and smoothness terms can lead the deformation and we set the weights to $(1, 400, 0, 0.1)$. Now the template is already closely enough to the target, we can use nearest neighbor information. We decrease the weight of spatial smoothness and increase the data term and set the weights to $(1, 200, 10, 0.1)$. Next, we turn off the weights of the dense correspondence term and optimize the energy based on the other three weights. This is to address inaccurate dense correspondence information. We set the weights to $(0, 100, 10, 0.1)$ and finally to $(0, 50, 10, 0.1)$. The nearest neighbors are

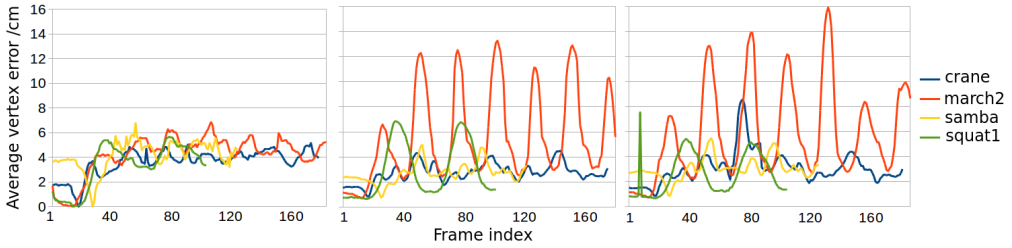


Figure 4.7 – Per-frame result on 4 sequences from Vlasic et al. [2008]. From left to right: our method, Allain et al. Allain et al. [2014] and Allain et al. Allain et al. [2015].

recomputed for each vertex every 20 iterations of optimization. The energy $E_{template}$ is optimized using a quasi-Newton method Liu and Nocedal [1989].

4.4 Evaluation

For better visualization, please refer to the supplemental video *.

Registered dataset We have tested our method against the methods of Allain et al. [2014] and Allain et al. [2015] on four sequences (`crane`, `march2`, `samba`, `squat1`) from Vlasic *et al*. Vlasic et al. [2008]. Since this dataset is temporally registered we can use it as ground truth for our evaluation. Our method takes as input the original preregistered dataset and we compute as error the Euclidean distance between our estimated vertex position and the ground truth. Both methods by Allain *et al*. require preprocessing of the data, which involves manual selection of the template and downsampling of the input mesh. We first map the processed template to the preregistered template by a nearest neighbor

*. https://hal.inria.fr/hal-01367791

approach and then proceed with the evaluation as described above. The preprocessing is done in favor of the reference algorithms.

All three methods share similar average quantitative results. The average vertex error over all sequences is 3.6cm, 3.9cm and 3.9cm for Allain et al. [2014], Allain et al. [2015] and our method respectively. The percentage of the vertices within 10cm error is 96.22, 94.68, and 96.83 respectively. Figure 4.7 shows the per-frame error on each sequence and for each method. Our method is generally more robust in the sense the maximum error for a sequence is lower. As shown in Figure 4.8, our method also better follows the deformation of wide clothes and makes good use of the prior knowledge of the underlying human whereas the two other methods loose tracking of the knee.

Dataset with sparse markers We have also tested our method on eight sequences from the dataset we have acquired in Section 3.2. These sequences include two subjects, one male and one female, wearing either layered or wide clothes, either performing a walk or rotating the body. Reconstruction artifacts like holes are present. Along with the mesh sequence, marker trajectories are also provided. We map each marker position to the nearest template vertex and use this mapping and correspondence information to compute the distance error of the alignment. We compute error as the Euclidean distance between our estimated location and the ground truth location of the marker.

Figure 4.9 shows a quantitative evaluation of our method on these sparse markers. The cumulative error curve, from which the template frame is excluded, shows that our method performs as well as the methods

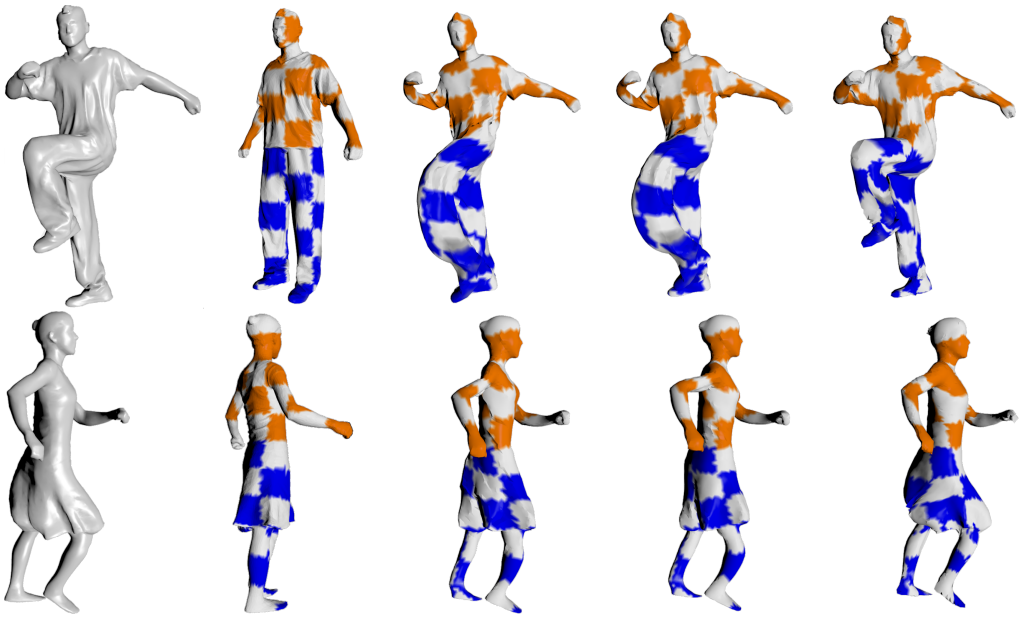


Figure 4.8 – Results on sequences `march2` (top row, frame #55) and `samba` (bottom row, frame #90) from Vlasic et al. [2008]. From left to right: $\mathcal{S}$, $\mathcal{T}$, result with Allain et al. [2014], result with Allain et al. [2015], and our result. See also the supplementary video.

by Allain *et al*. The bottom curves show that our method does not drift in time, because our dense correspondence is calculated from the template to each frame independently and is based on a partial near-isometric deformation model.

Figure 4.10 shows alignment results on representative sequences of our dataset. As for the registered dataset, these results also demonstrate that our method is more robust than Allain et al. [2014] and Allain et al. [2015], where limbs may be switched or the tracking of clothing may get lost. The color-coding shows that our method does not suffer from significant drift. Artifacts produced in the visual hull reconstruction even have a chance to be repaired such as the tunnel appearing on the left leg for the

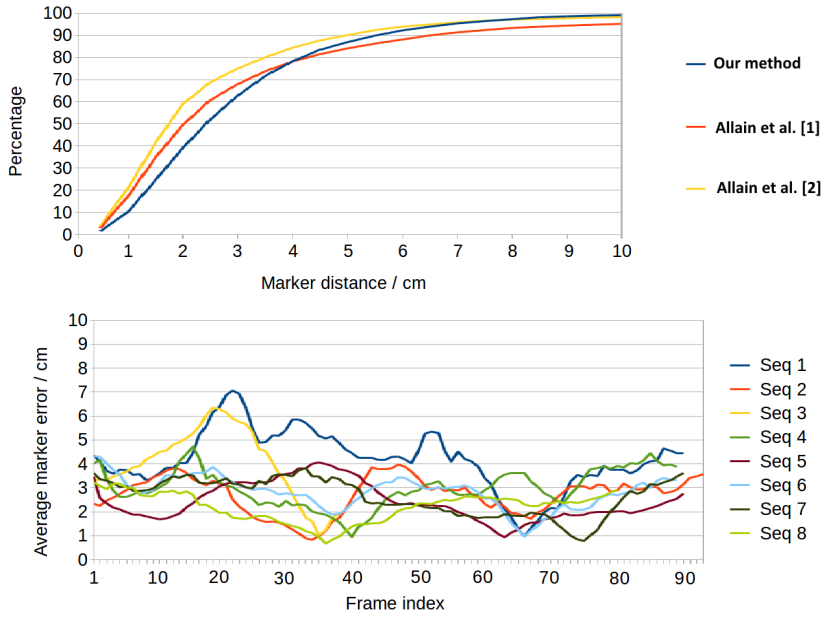


Figure 4.9 – Quantitative results of sparse markers on our dataset. Top: cumulative error curves for Allain et al. [2014], Allain et al. [2015] and our method. Bottom: average marker error per frame of our method for all eight sequences.

last example.

4.5 Conclusion

In this chapter we have presented a method for non-rigid registration of motion sequences of humans in clothing. In order to be robust to geometric and topological artifacts, we rely on body landmarks which are mapped to the clothed surface and serve as starting points to grow partial near-isometric patches. These patches allow to compute dense pairwise correspondence information between an automatically selected template surface, based on topological and geometric criteria, and every surface in other frames of the sequence. Correspondences are then used to guide

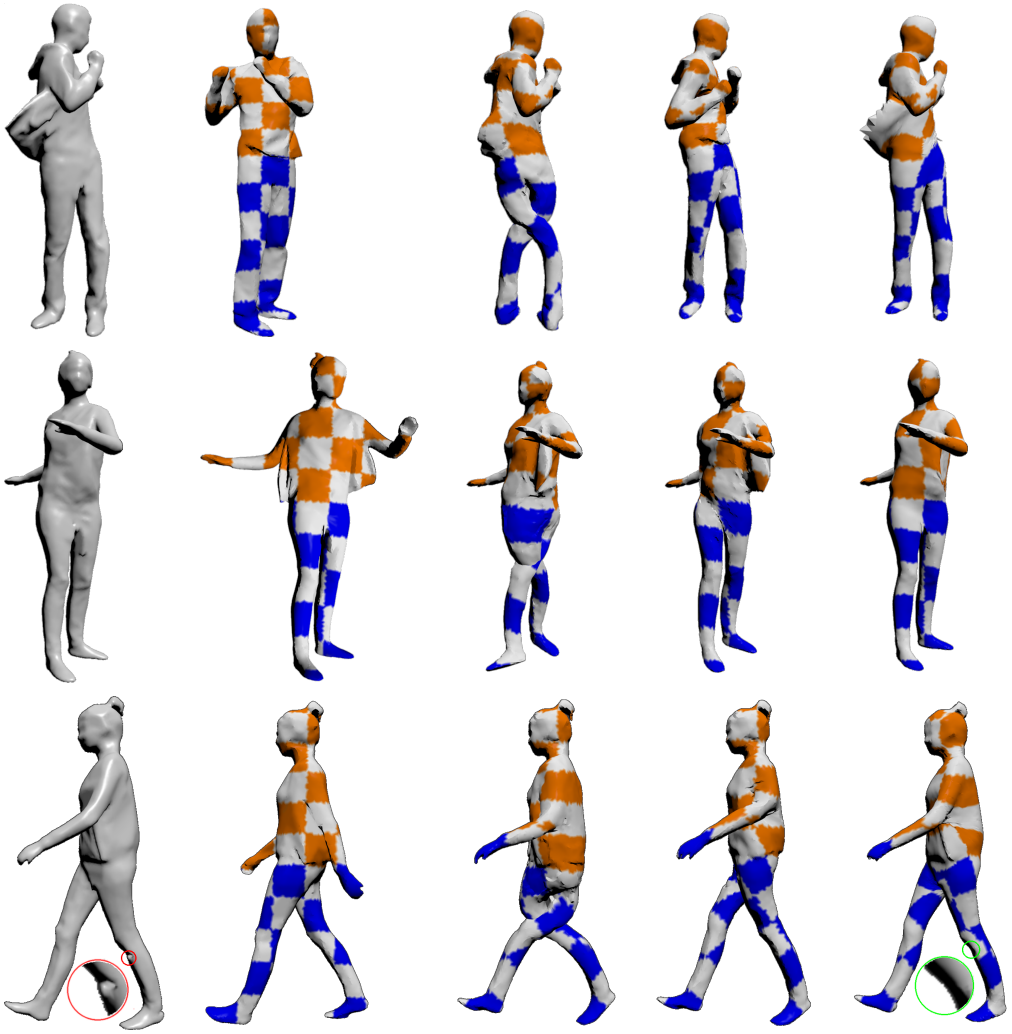


Figure 4.10 – Alignment result on three representative sequences of our dataset. From left to right: $\mathcal{S}$, $\mathcal{T}$, result with Allain et al. [2015], result with Allain et al. [2015] and our result.

the deformation from the template to the other frames.

This method relies on a few assumptions, which are valid in most practical cases. In particular, cloth is supposed to be inelastic and not to slip much along the body. When this assumption is no longer valid, for example, imaging dancing with a skirt, the projected anatomical points are not meaningful anymore. As a remedy to the proposed method, one needs to find reliable landmarks on the clothing surface. Another assumption is that the topology of the scene is supposed to stay constant, and at least one frame in the sequence needs to capture the correct topology. This is usually easy to achieve in controlled acquisition, including all the acquisition we have done concerning this thesis, since a separate static "A" pose acquisition and reconstruction can be arranged. However, if this assumption is violated in other application, for instance, rotating with an open jacket, the selected "template" does not possess the right topology, therefore, it can not be used to register the entire sequence just by performing near isometric deformation, but needs to be cut and separated, which could be a very difficult problem. A bypass could be to construct a template from the sequence, as described in Tevs et al. [2012]. Apart from the above two assumptions, the proposed method also suffers from occasionally non-isometric deformation. It happens especially when there is a merge in the observation. In this case the locally partial isometry assumption is violated, so that no dense correspondence in the area can be computed. Without these dense correspondences, the deformation tends to be non-isometric, since there is no regularization term in the energy function that controls the non-isometric behavior. For further improvement, it will be reasonable to also include regularization

terms such as as-rigid-as-possible or as-isometric-as-possible terms into the energy function.

After registering the clothed sequence, we can now represent it as a deformable surface. Now we will study the statistics of the deformation and recover the effect of various controlling variables to the shape in next chapter.

Chapter 5

Clothing deformation modeling

Contents

5.1 Introduction

In the previous two chapters, we have proposed approaches to process raw 3D acquisitions in order to extract body and clothing shape information. In this chapter we will propose a data-driven learning approach to generated the shape of moving clothed human based on simulated or captured training data. We propose to leverage existing 3D motion sequences by performing a statistical analysis of the dynamically deforming clothing layer in order to allow for efficient subsequent synthesis of 3D motion. Performing statistical analysis on the dynamically deforming clothing layer is challenging for two main reasons. First, the motion of the clothing is influenced by numerous factors including the body shape, posture, and direction and speed of motion of the underlying human as well as the material composition of the cloth. To allow for controlled synthesis of these effects, they must be explicitly modeled. Second, the geometry of the person with clothing may be significantly different from that of the underlying human body, e.g. in case of a dress, and the clothing may slide w.r.t. the human skin. Hence, computing assignments between the clothing layer and the body is complicated, especially across different subjects who wear the same type of clothing (e.g. shorts and T-shirt).

Two existing lines of learning approaches analyze the clothing layer to allow for synthesis (details in Section 2.4). The simulation-based methods are limited by the quality of the simulated synthetic training data, which can hardly handle complicated clothing. The capture-based methods usually can only be trained on one specific actor wearing a fixed outfit, and can hence not be used to synthesis changes in the clothing itself (e.g.

changes of fit or materials).

The method proposed in this chapter combines the advantages of both methods by enabling to train from various motion data that can either be simulated or captured. To this end, we perform a statistical analysis to model the geometric variability of the clothing layer. Our main contribution is that the proposed analysis is versatile in the sense that it can be used to train and regress on semantic parameters, thereby allowing to control e.g. clothing fit or material parameters of the synthesized sequences. Our statistical analysis models the deformation of the clothing layer w.r.t. the deformation behavior of the underlying human actor represented using a statistical model of 3D body shape. For the analysis, we consider two fairly straightforward models. First, we model the layer variations with a linear subspace. Second, we model the variation of the clothing layer using a statistical regression model, in order to capture some of the underlying causal dynamics in relatively simple form. Our experiments show the validity of the representation with qualitative and quantitative captured sequence reconstruction experiments based on these parameterizations. We further qualitatively demonstrate the value of our approach for three applications. First, following Neophytou and Hilton [2013], we train from multiple sequences of the same actor in the same clothing acquired using dense 3D motion capture and use this to exchange body shape and motion. Second, following Guan et al. [2012], we train from multiple simulated sequences showing the same actor in clothing of different materials and use this to change the material of the clothing. Third, to demonstrate the novelty of our approach, we train from multiple sequences of different actors in the same type of clothing acquired using

dense 3D motion capture and use this to change the clothing.

5.2 Methodology

In this chapter, we propose a general framework to study the deformation of the clothing layer. First, we estimate the human body shape using the method introduced in chapter 3 and extract the offset clothing layer in a way that is robust to situations where the geometry of the dressed person differs significantly from the geometry of the human body, as in the case of a dress. A fuzzy vertex association from the clothing surface to the body surface is established, so that we can represent the clothing deformation as an offset mesh based on the body surface. Second, we use statistical analysis to analyze the geometric variability of the clothing layer to greatly reduce self-redundancies. Third, we show how to capture some of the underlying causal dynamics in relatively simple form, by modeling the variation of this clothing layer as a function of body motion as well as semantic variables using a statistical regression model.

The input to our method is a set of 3D sequences showing the dense geometry of a human in clothing performing a motion. These sequences may have been generated using physical simulation or motion capture setups. Before further processing, we require for each sequence an estimate of the underlying body shape and motion and an alignment of the clothing layer. Note that the clothing layer may optionally include the geometry of the body itself, i.e. show the body with clothing. If the sequences were generated using physical simulation, this information is typically readily available. For captured data, any of the previously reviewed

methods (in Chapter 2, Subsection 2.2 and Section 2.3) may be used to compute this information. We estimate the underlying body shape using the method proposed in Chapter 3 and compute alignments of the complete deforming surface (i.e. the human in clothing) using an embedded deformation model based on Li *et al*. Li et al. [2009] without refining the deformation graph.

In the following, we denote the registered sequences of the clothing layer by $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n$ and the corresponding sequences of underlying body shape estimates by $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_n$. Furthermore, let $\boldsymbol{C}_{i,k}$ and $\boldsymbol{B}_{i,k}$ denote the k-th frames of $\boldsymbol{C}_i$ and $\boldsymbol{B}_i$, respectively. Thanks to the alignment, $\boldsymbol{C}_{i,k}$ has the same number of corresponding vertices as $\boldsymbol{C}_{j,l}$. Similarly, $\boldsymbol{B}_{i,k}$ has the same number of corresponding vertices as $\boldsymbol{B}_{j,l}$. While sequences $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$ (and similarly $\boldsymbol{B}_i$ and $\boldsymbol{B}_j$) may consist of different numbers of frames, $\boldsymbol{C}_i$ and $\boldsymbol{B}_i$ contain corresponding clothing layer and body estimate and therefore consist of the same number of frames.

The body estimates in sequence $\boldsymbol{B}_i$ can be expressed using a generative statistical body model that decouples the influence of identity and posture variation Neophytou and Hilton [2013], Loper et al. [2015], Pishchulin et al. [2017]. This allows to represent $\boldsymbol{B}_i$ using one vector $\boldsymbol{\beta}_i$ for identity information and a vector $\boldsymbol{\theta}_{i,k}$ per frame for pose information. These generative models allow for two important modifications. First, the body shape of the actor can be changed while keeping the same motion by modifying $\boldsymbol{\beta}_i$. Second, the body motion can be changed by modifying $\boldsymbol{\theta}_{i,k}$ for each frame.

In this work, we use S-SCAPE as generative model Pishchulin et al.

[2017], which uses the A-pose as standard pose $\boldsymbol{\theta}_0$. S-SCAPE combines a linear space learned using principal component analysis (PCA) to represent variations due to identity with a linear blend skinning (LBS) model to represent variations in pose. Consider the j-th vertex $\boldsymbol{v}^{\boldsymbol{B}}_{i,k,j}$ of frame $\boldsymbol{B}_{i,k}$. This vertex is generated by transforming the j-th vertex $\boldsymbol{\mu}^{\boldsymbol{B}}_j$ of the mean body shape in standard pose $\boldsymbol{\theta}_0$ as $\boldsymbol{v}^{\boldsymbol{B}}_{i,k,j} = \boldsymbol{T}_j(\boldsymbol{\theta}_{i,k})\boldsymbol{T}_j(\boldsymbol{\beta}_i)\boldsymbol{\mu}^{\boldsymbol{B}}_j$, where $\boldsymbol{T}_j(\boldsymbol{\theta}_{i,k})$ and $\boldsymbol{T}_j(\boldsymbol{\beta}_i)$ are (homogeneous) transformation matrices applying the transformations modeled by LBS and learned by PCA. We can hence use S-SCAPE to define an operation called *unposing* in the following. This operation changes the pose of $\boldsymbol{B}_{i,k}$ to the standard pose $\boldsymbol{\theta}_0$ while maintaining body shape by replacing vertex $\boldsymbol{v}^{\boldsymbol{B}}_{i,k,j}$ for all j by

$$\tilde{\boldsymbol{v}}^{\boldsymbol{B}}_{i,k,j} = \left(\boldsymbol{T}_j(\boldsymbol{\theta}_{i,k})\right)^{-1} \boldsymbol{v}^{\boldsymbol{B}}_{i,k,j}. \tag{5.1}$$

Offset clothing layer extraction

We model the clothing layer as an offset from the body. To this end, we need to find corresponding vertices on the body mesh for each clothing vertex. Because $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n$ and $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_n$ are temporally coherent, respectively, we can establish this correspondence on a single pair of frames $(\boldsymbol{C}_{i,k}, \boldsymbol{B}_{i,k})$ and propagate this information to all sequences. In practice, a pair of frames with few concavities is preferred because it enhances the robustness of the sparse association when created using a ray shooting method (see next paragraph). However to prove the generality of our approach, in our experiments, the association is simply estimated on the first frame of the first sequence. Since the following description is limited to a single pair of frames $(\boldsymbol{C}_{1,1}, \boldsymbol{B}_{1,1})$, for simplicity, we will drop frame and sequence index in this subsection.

$\boldsymbol{C}$ and $\boldsymbol{B}$ usually consist of a different number of vertices and have possibly significantly different geometry. Hence, a bijective association is in general not achievable. As our final goal is to model the deformation of the clothing layer using the body layer, our main interest is to find one or more corresponding vertices on $\boldsymbol{B}$ for each vertex on $\boldsymbol{C}$. We achieve this by computing a sparse correspondence that is subsequently propagated to each vertex on $\boldsymbol{C}$ using a probabilistic geodesic diffusion method. Note that unlike Pons-Moll *et al*. Pons-Moll et al. [2017], our method works for difficult geometries such as skirts without manual intervention.

Sparse association For each vertex $\boldsymbol{v}_j^{\boldsymbol{B}}$ on $\boldsymbol{B}$ we shoot a ray along the surface normal outwards the body. If there is an intersection $\boldsymbol{p}_j^{\boldsymbol{C}}$ with $\boldsymbol{C}$ and the distance between $\boldsymbol{v}_j^{\boldsymbol{B}}$ and $\boldsymbol{p}_j^{\boldsymbol{C}}$ is within a threshold of $15cm$, we search for the vertex $\boldsymbol{v}_i^{\boldsymbol{C}}$ on $\boldsymbol{C}$ closest to $\boldsymbol{p}_j^{\boldsymbol{C}}$. Such pairs $\left(\boldsymbol{v}_i^{\boldsymbol{C}}, \boldsymbol{v}_j^{\boldsymbol{B}}\right)$ are considered to be associated. If multiple body vertices are associated with the same clothing vertex, we only keep one pair per clothing vertex to put the same weight to each sparsely associated $\boldsymbol{v}_i^{\boldsymbol{C}}$. The pairs $\left(\boldsymbol{v}_i^{\boldsymbol{C}}, \boldsymbol{v}_j^{\boldsymbol{B}}\right)$ are defined as *sparse association*.

Fuzzy dense association We now propagate the sparse association to every clothing vertex. Intuitively, if a clothing vertex $\boldsymbol{v}_i^{\boldsymbol{C}}$ is associated to a body vertex $\boldsymbol{v}_j^{\boldsymbol{B}}$ then there is a high probability that the neighboring vertices of $\boldsymbol{v}_i^{\boldsymbol{C}}$ should be associated to the neighboring vertices of $\boldsymbol{v}_j^{\boldsymbol{B}}$. Based on this idea, for any pair $\left(\boldsymbol{v}_k^{\boldsymbol{C}}, \boldsymbol{v}_l^{\boldsymbol{B}}\right) \in \boldsymbol{C} \times \boldsymbol{B}$ we initialize the association probability $P\left(\boldsymbol{v}_k^{\boldsymbol{C}}, \boldsymbol{v}_l^{\boldsymbol{B}}\right)$ to be 0. Then we loop on all the sparse association pairs $\left(\boldsymbol{v}_i^{\boldsymbol{C}}, \boldsymbol{v}_j^{\boldsymbol{B}}\right)$ and update the association probability of any

vertex pair $\left(\boldsymbol{v}_k^{\boldsymbol{C}}, \boldsymbol{v}_l^{\boldsymbol{B}}\right)$ according to:

$$P\left(\boldsymbol{v}_k^{\boldsymbol{C}}, \boldsymbol{v}_l^{\boldsymbol{B}}\right) = P\left(\boldsymbol{v}_k^{\boldsymbol{C}}, \boldsymbol{v}_l^{\boldsymbol{B}}\right) + exp\left(-\left(r\left(\boldsymbol{v}_k^{\boldsymbol{C}}, \boldsymbol{v}_i^{\boldsymbol{C}}\right) + r\left(\boldsymbol{v}_l^{\boldsymbol{B}}, \boldsymbol{v}_j^{\boldsymbol{B}}\right)\right)/\sigma^2\right), \tag{5.2}$$

where $r(,)$ computes the squared geodesic distance between two vertices. In our implementation we set σ to $1cm$. To simplify the computation we only consider vertices $\boldsymbol{v}_k^{\boldsymbol{C}}$ and $\boldsymbol{v}_l^{\boldsymbol{B}}$ that lie within $3cm$ geodesic distance from $\boldsymbol{v}_i^{\boldsymbol{C}}$ and $\boldsymbol{v}_j^{\boldsymbol{B}}$. For the dense association, for each vertex on $\boldsymbol{C}$ we choose a constant number n_f of vertices on S_B that have the highest association probability values as associated vertices. We normalize the association probability to form fuzzy association weights, and store the indices of the n_f associations in a list I. This step does not only compute body vertex matches for previously unassociated clothing vertices but can also correct wrong matches from the sparse association and make the association more meaningful in situations where $\boldsymbol{C}$ and $\boldsymbol{B}$ differ significantly. This is illustrated in the case of a skirt on the right of Figure 5.1 (see also Section 5.3 for a discussion).

Offset representation of clothing layer Since we have established correspondence between $\boldsymbol{C}$ and $\boldsymbol{B}$, we can now get the offset clothing layer by subtracting $\boldsymbol{B}$ from $\boldsymbol{C}$. However, this Euclidean offset depends on the human pose and the global rotation. To account for this, we first unpose both $\boldsymbol{B}$ and $\boldsymbol{C}$. The body estimate $\boldsymbol{B}$ is unposed using Equation 5.1, and the clothing layer $\boldsymbol{C}$ is unposed with the help of the fuzzy dense association by replacing vertex $\boldsymbol{v}_j^{\boldsymbol{C}}$ for all j by

$$\tilde{\boldsymbol{v}}_j^{\boldsymbol{C}} = \left(\sum_{i=1}^{n_f} \omega_i \boldsymbol{T}_{I_j[i]}(\boldsymbol{\theta})\right)^{-1} \boldsymbol{v}_j^{\boldsymbol{C}}, \tag{5.3}$$

where ω_i are the fuzzy association weights and $I_j[i]$ denotes the i-th entry of the index list I associated with vertex $\boldsymbol{v}_j^{\boldsymbol{C}}$. The offset of each clothing vertex is then obtained as:

$$\boldsymbol{d}_{i,j} = \tilde{\boldsymbol{v}}_i^{\boldsymbol{C}} - \tilde{\boldsymbol{v}}_j^{\boldsymbol{B}}, \quad \boldsymbol{d}_{i,j} \in \mathbb{R}^3, \tag{5.4}$$

where $\left(\tilde{\boldsymbol{v}}_i^{\boldsymbol{C}}, \tilde{\boldsymbol{v}}_j^{\boldsymbol{B}}\right)$ form a fuzzily associated pair. We stack all the $\boldsymbol{d}_{i,j}$ from one frame pair $(\boldsymbol{C}, \boldsymbol{B})$ to form a single vector denoted by $\boldsymbol{d} \in \mathbb{R}^{3\times n_f \times n_v}$, where n_v is the number of vertices in $\boldsymbol{C}$.

Clothing layer deformation space reduction

The deformation of the offset clothing layer is now encoded in $\boldsymbol{d}$. To reduce the self-redundancies in $\boldsymbol{d}$, we perform PCA on $\boldsymbol{d}$ from all frame pairs $\left(\boldsymbol{C}_{i,k}, \boldsymbol{B}_{i,k}\right)$. This allows for the clothing deformation to be represented by PCA coefficients $\boldsymbol{\alpha}_k$. Note we do not assume $\boldsymbol{d}_k$ to form a Gaussian distribution. The purpose of PCA is only to reduce the dimensionality of the space, not to sample from it.

We would like to learn a mapping from semantic parameters of interest, denoted by $\boldsymbol{\gamma}$,to the clothing layer deformation. After obtaining a low dimensional representation, this is equivalent to finding a mapping from $\boldsymbol{\gamma}$ to $\boldsymbol{\alpha}$. The PCA representation of the offsets successfully gets rid of self-redundancies in clothing layer. Furthermore, in PCA space, we can choose the number of principal components to use in order to balance the speed, storage, and quality.

Neural network for regression

To allow control of the offset clothing layer deformation, we study the relationship between its variation and semantic parameters γ, where γ can be body motion, clothing style, clothing material and so on. We treat this as a regression problem that learns the mapping from γ to α. Due to the nonlinearity of the problem itself and the potentially large sample size, we choose a fully connected two-hidden-layer neural network to train the regression, with the size of input layer equal to the dimensionality of the semantic parameters and the size of output layer equal to the number of principal component used. The sizes of the first and second hidden layers are 60 and 80, respectively. In our implementation, the neural network is implemented with OpenNN lib. For each experiment, we set 20% of the frames from training data as validation frames. We choose mean square error as loss function, quasi-Newton method as optimization strategy, and stop the training once validation error starts to increase.

5.3 Method validation

To validate each step of our method, we train on small training sets consisting of a single sequence each ($n = 1$) using ten existing sequences of the Adobe Vlasic et al. [2008] and Inria Yang et al. [2016] datasets showing fast, large-scale motion in ample clothing as this is especially challenging to model. In all following experiments, body motion is parameterized by global speed, joint angles and joint angular speed. For offset clothing layer extraction, we show that we can extract the entire clothing layer regardless of the clothing geometry. Then we validate our PCA step to show that

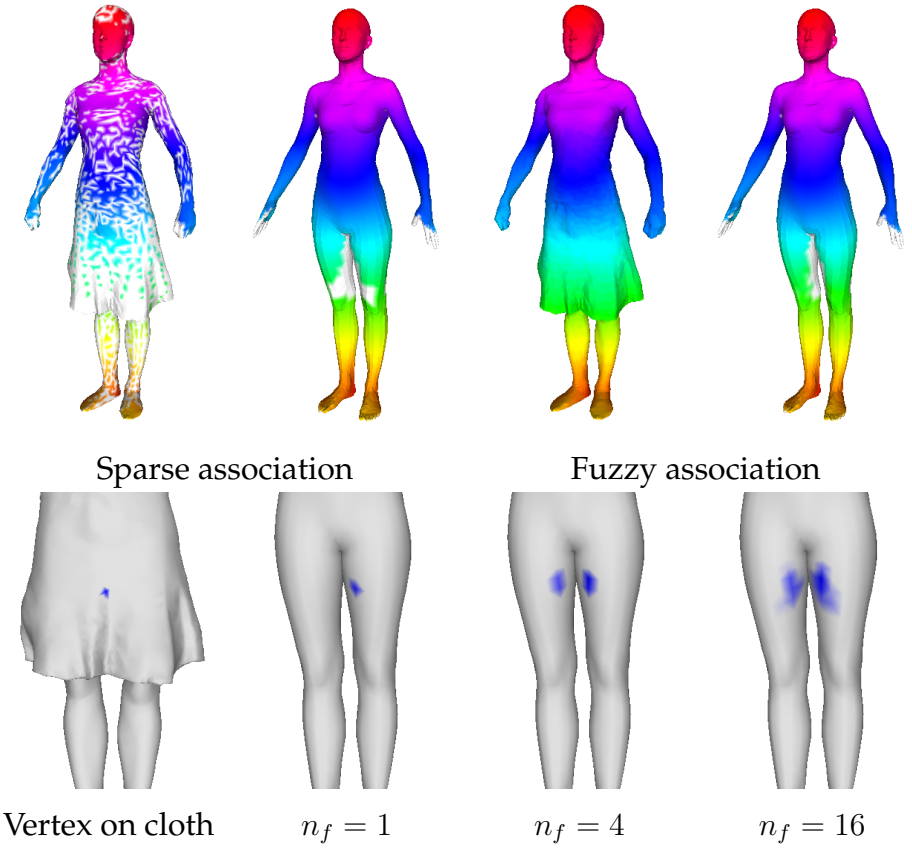


Figure 5.1 – Top: associations of clothing and body layer ($n_f = 1$), where color indicates the association. Bottom: a blue vertex on the skirt is associated to body vertices with different n_f. The intensity of the blue color is proportional to the association weight.

it greatly reduces the deformation space with acceptable reconstruction error. Finally, we validate the neural network regression by showing that both training error and testing error are satisfying.

Offset clothing layer extraction

We model the offset by first constructing a sparse correspondence between clothing and body, and then propagating the correspondence to each clothing vertex. Figure 5.1 (left) shows an example of the sparse and

fuzzy association. Note that if we only use sparse association to store the information about clothing deformation, the information of the lower part of the dress is not recorded sufficiently.

The geometry of the clothing layer and the underlying human body differs significantly in the case of a skirt. Hence, $n_f = 1$ may not be meaningful and robust enough as having a single associated vertex is prone to form a seam in the middle of the front and the back faces of the skirt as some vertices around those areas are associated to the left thigh while neighboring ones are associated to the right thigh. Using higher n_f, such a skirt vertex is associated to both legs, therefore preventing seams. This is illustrated in Figure 5.1 (right).

We use our fuzzy association to directly transfer the offset clothing layer on data from Pons-Moll *et al*. Pons-Moll et al. [2017]. Compared with their work, our method achieves similar results, shown in Figure 5.2, without the need for manual intervention.

PCA deformation space reduction

In our experiments, the dimension of the offset clothing layer vector generally varies from 20,000 to 80,000. To reduce this dimensionality, we perform PCA on the extracted offset of the clothing layer. To analyze how many PCs to keep, we reconstruct the sequence with different numbers of PCs. We compare the reconstruction against the original sequence by computing the average vertex position error. Table 5.1 gives errors per sequence for different numbers of principal components. Figure 5.3 visualizes the effect of increasing the number of PCs for one example. Such

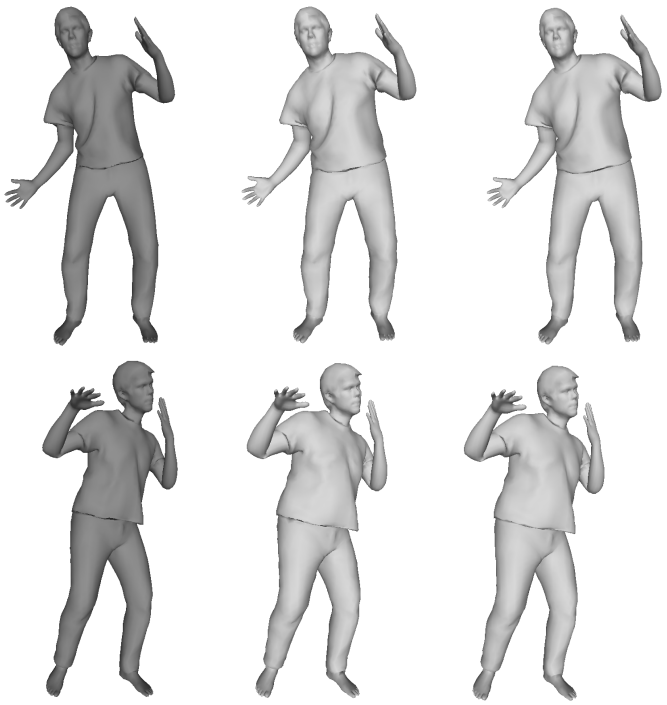


Figure 5.2 – Two example frames of comparison to Pons-Moll et al. [2017]. From left to right: original acquisition, transferred clothing layer with our method, and with Pons-Moll et al. [2017]. Both methods produce very similar results.

an analysis allows to choose the number of PCs to satisfy requirements on accuracy, speed or memory usage. In all following experiments, when training on a single subject with fixed clothing, we use 40 PCs, and when training on multiple subjects or multiple clothings, we use 100PCs as we found these datasets to contain more variation

Neural network regression

We validate our neural network by regressing 40 PCA coefficients to human body motion. Each sequence consists of 95-275 frames. We choose 20% of the frames from each sequence as testing data and the remaining

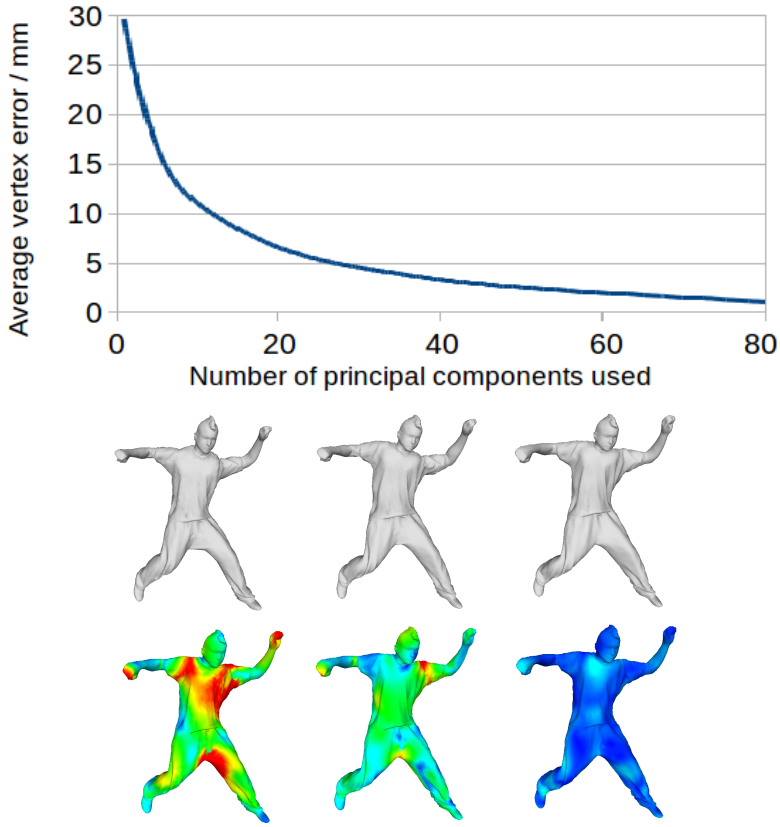


Figure 5.3 – Left: curve shows that the average reconstruction error drops when more principal components are used. Right: one example frame from *bouncing* sequence with the first row showing PCA reconstruction and the second row showing the error in color. From left to right 1 PC, 5 PCs and 40 PCs. Blue $= 0mm$, red $>= 50mm$.

80% to be the training data. After training, we feed the motion parameters for all frames to the network and get 40 PCA coefficients for each frame to reconstruct the sequence. This reconstruction is then compared against the ground truth. Table 5.2 shows the quantitative error of the regression. The training error and the prediction error are generally low and close to the reconstruction error when using 40 principal components, which means our neural network regression is accurate and does not overfit the training data. Figure 5.4 shows the visual result of some examples of the

Sequences	1 PC	5 PCs	10 PCs	20 PCs	30 PCs	40 PCs	50 PCs	All PCs
Bouncing	32.48	22.52	16.77	11.30	8.35	6.49	5.14	0.10
Crane	26.53	14.87	10.47	6.27	4.18	3.08	2.35	0.11
March 1	22.95	13.94	8.88	5.18	3.64	2.68	2.08	0.11
March 2	23.03	14.39	9.79	6.33	4.54	3.44	2.70	0.12
Squat 1	46.33	18.94	10.39	5.74	3.70	2.67	1.99	0.09
Squat 2	24.73	12.35	8.00	4.55	3.19	2.33	1.73	0.09
Samba	24.05	14.75	8.95	5.15	3.45	2.40	1.75	0.10
Swing	30.42	19.68	13.78	7.97	5.24	3.62	2.66	0.09
s1 mul. w.	27.24	13.73	8.42	4.50	3.01	2.17	1.58	0.13
s1 wide w.	23.92	13.28	9.08	5.05	3.28	2.30	1.64	0.12
s6 mul. w.	28.20	12.94	5.89	2.87	1.91	1.38	1.02	0.09
s6 wide w.	26.38	12.25	7.09	3.94	2.71	1.94	1.39	0.13

Table 5.1 – The table shows the reconstruction error for each sequence using different number of principal components.

regression error. Both training and prediction error are almost always low.

5.4 Applications

This section shows the virtue of the proposed method by applying it to three scenarios. The first trains from multiple sequences of the same actor in the same clothing and uses this to synthesize similar clothing on new body shapes and under new motions. The second trains from multiple simulated sequences showing the same actor in clothing of different materials and uses this to change the material of the clothing. The third application trains from multiple sequences of different actors in the same type of clothing and uses this to change the fit of the clothing. This entirely new way of synthesizing clothing is possible thanks to our regression to

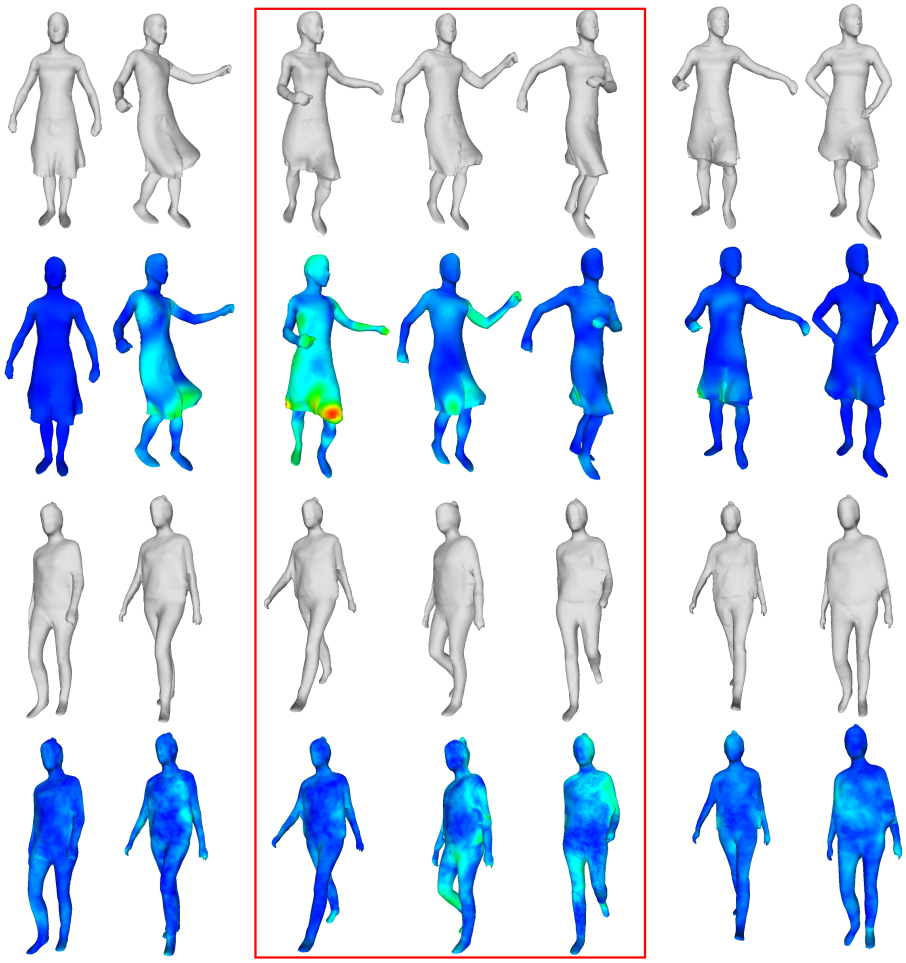


Figure 5.4 – Example frames from two sequences showing the reconstruction based on trained neural network. The first and the third row show the reconstruction and the second and the forth show the corresponding vertex error on ground truth meshes. The three columns in the red box are predictions from testing frames; others are from training frames. Dark blue $= 0$ millimeters, bright red $>= 5$ millimeters.

Sequences	Training error	Prediction error
Bouncing	9.95	10.27
Crane	4.71	4.28
March 1	3.54	3.93
March 2	4.39	5.67
Squat 1	6.57	4.31
Squat 2	3.85	4.09
Samba	4.27	7.44
Swing	7.72	5.95
s1 multi walk	10.87	8.57
s1 wide walk	9.09	9.21
s6 multi walk	7.71	8.19
s6 wide walk	5.97	4.82

Table 5.2 – The table shows the reconstruction error based on regression for each sequence. Error is calculated as average vertex euclidean distance error over all the training or testing frames.

semantic parameters. For better visualizations of the results, refer to the supplemental material.

Clothing dynamics modeling

Change body shape After extracting the offset of the clothing layer, we can add this offset to any body shape under normalized pose and update the pose of the body with clothing using the relations of Equations 5.3 and 5.4. Figure 5.5 shows two examples of changing the body shape of a given motion sequence.

Change clothing dynamics In this part, we trained our regression model from multiple sequences of the same actor in the same clothing

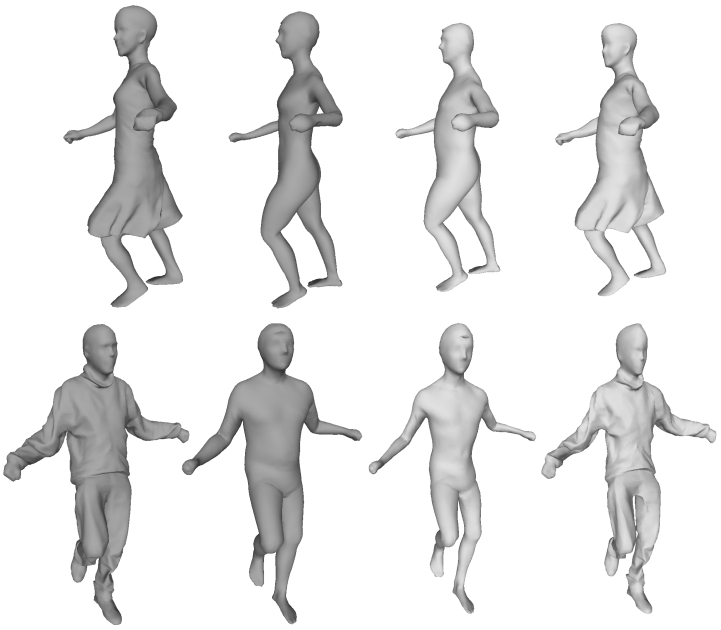


Figure 5.5 – Change body shape. From left to right: original clothing mesh, estimated body, changed body, new clothing mesh.

acquired using dense 3D motion capture. We use the regression model to learn the mapping from the body motion parameters to the PCA coefficients of the offset vectors. To synthesize new sequences, we feed new motion parameters to the model. Figure 5.6 shows examples of the resulting changes in the clothing dynamics. Note that realistic wrinkling effects are synthesized.

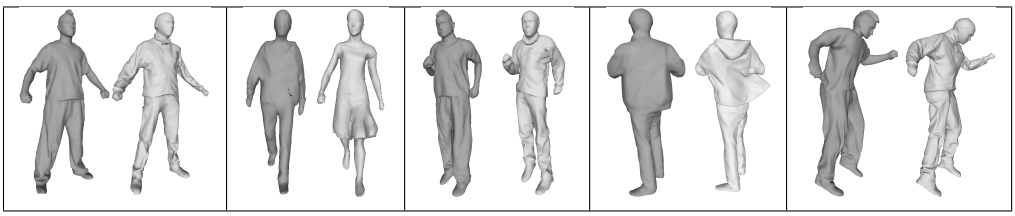


Figure 5.6 – Examples of changing clothing dynamics. The brighter gray meshes are not in the training data but generated by feeding the motion parameters of the darker gray meshes to the neural network trained on sequences containing the brighter gray clothes.

Clothing material modeling

This section shows how to model material parameters using our method. As material parameters are not readily available for captured data, we train from synthetic data generated using a state-of-art physical simulator Li et al. [2018]. For training, we simulate 8 sequences of the same garment pattern, worn by the same actor in a fixed motion, with varying materials. We choose a detailed garment pattern with garment-to-garment interaction during motion as this generates rich wrinkles that are challenging to model. The materials were generated using 39 parameters Wang et al. [2011], and to allow for easier control of the parameters, we reduced their dimensionality to 4 using PCA before regressing from material parameter space to offset space. We used 7 materials to train our regression model, and left 1 material for testing. To avoid over-fitting to these 7 material points, we added a Gaussian random noise to the material parameters for all frames when training the regression. After training, we predicted the clothing layer from the motion parameters and new material parameters. Since for simulated data, segmented and aligned clothing and body meshes are available for each frame, our method uses this information. That is, we use the clothing layer directly and fit the S-SCAPE model to the mesh of the undressed body model used for simulation.

Figure 5.7 shows the comparison between our prediction and the ground truth for the test sequence. Note that a globally correct deformation is predicted even though the cloth deformation is far from the body shape. In spite of the globally correct deformation, our prediction lacks some detailed wrinkles. We suspect this detailed loss is due to dimension

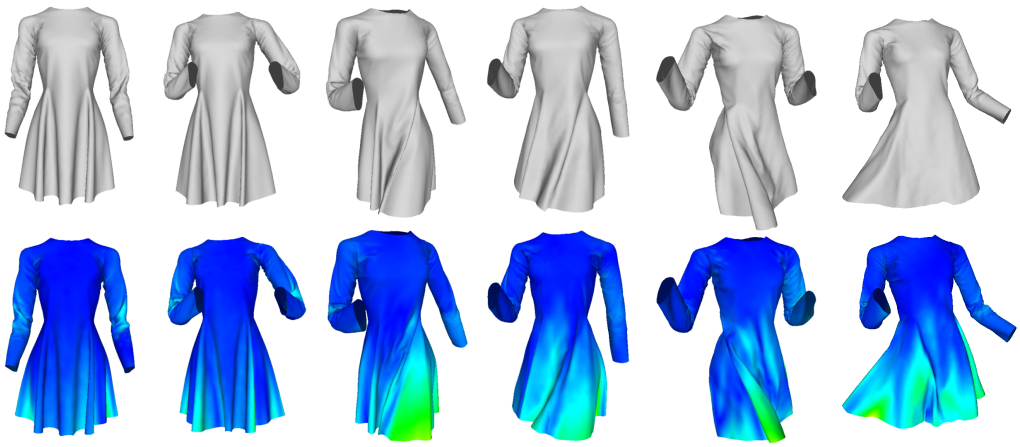


Figure 5.7 – Comparison to ground truth. First row: our predicted clothing deformation. Second row: ground truth colored with per-vertex error. Blue $= 0cm$, red $= 10cm$.

reduction on both material space and deformation space, as well as the limitation of the training data size. For qualitative validation, we randomly sampled material parameters in the PCA subspace and used them to synthesize new sequences. Figure 5.8 shows some examples. Note that visually plausible sequences are synthesized.

Clothing fit modeling

The proposed analysis is versatile in terms of the parameters of interest we wish to regress to. This allows for entirely new applications if sufficient training data is available. We demonstrate this by explicitly modeling the clothing size variation from acquisition data, which has not been done to be best of our knowledge. For training, we use 8 sequences of an extended version of the Inria dataset Yang et al. [2016] of different subjects (4 male and 4 female) wearing different shorts and T-shirts while walking. These sequences are tracked with a common mesh topology. For each

Figure 5.8 – Two synthesized sequences with new material parameters.

sequence, we manually assign a three dimensional vector to describe the size of the clothing, containing the width and length of the shorts and the size of the T-shirt. To model relative fit rather than absolute size, the sizes are expressed as ratio to corresponding measurements on the body. During training, to avoid over-fitting to these 8 sizing points, we add a Gaussian random noise to each size measurement. The regression learns a mapping from the body motion and size parameters to the PCA of the offsets. After training, new size parameters along with a motion allow to synthesize new sequences. Figure 5.9 shows modifications of the clothing fit on one frame of a sequence. Note that although our method learned certain clothing size variations, the three dimensions of our measurements are not completely separated, as e.g. the "large T-shirt" also introduces wider shorts. We believe this is caused by the limited size of training

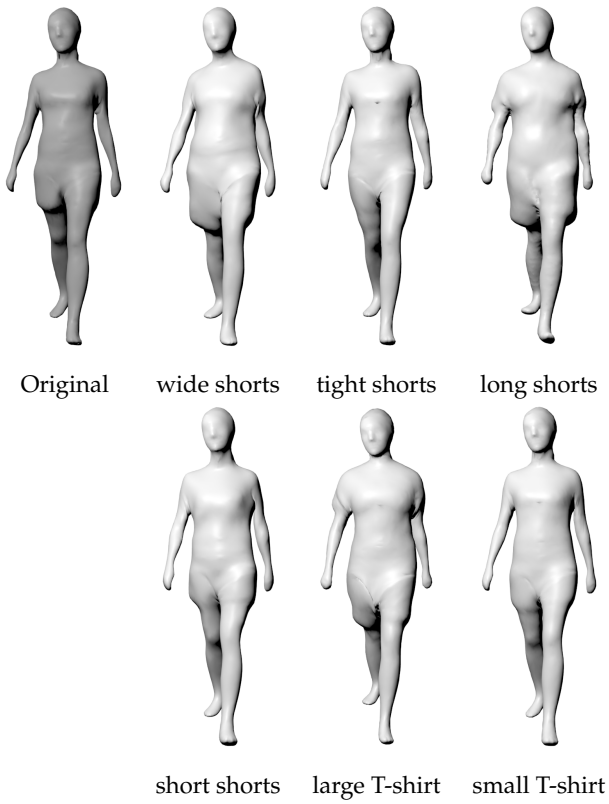


Figure 5.9 – Change the clothing fit shown on one frame of a sequence.

data. Since the regression also models body motion, our method not only captures size variation, but also dynamic deformation caused by motion. Figure 5.10 shows examples of this.

5.5 Conclusion

In this chapter we have presented a statistical analysis and modeling of the clothing layer from sets of dense 3D sequences of human motion. Our analysis shows PCA to be a suitable tool to compress the geometric variability information contained in the clothing layer. The regression component of our model is shown to properly capture the relation be-

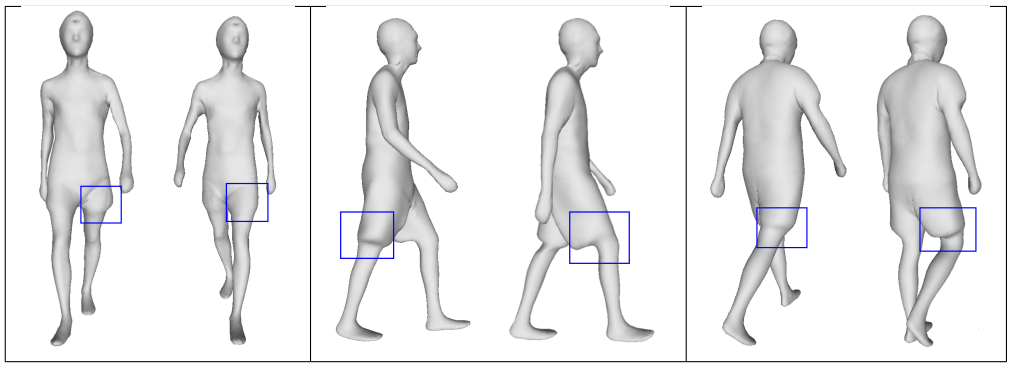


Figure 5.10 – Our approach captures dynamics caused by both clothing fit and body motion.

tween layer variations and semantic parameters as well as the underlying motion of the captured body. This allows predictions of the clothing layer under previously unobserved motions, with previously unobserved clothing materials or clothing fits.

Our model opens a large number of future possibilities. First, it can be extended to include more variability under different clothing worn by a large number of subjects. Second, more elaborate clothing layer motion subspace models could be devised. Third, several semantic regression groups could be simultaneously considered.

Despite of the promising potentials, the proposed methods also suffer from certain limitations. The first limitation comes from the clothing offset layer extraction. It relies heavily on the result of non-rigid registration of the clothing surface and the human body estimation. The registration quality is essential for later analysis and regression. In practice, human intervene to monitor and improve the non-rigid registration is inevitable. It would be a great interest to explore the possibility of extracting the off-set clothing layer without registering the clothing surface. An offset

from body along body surface normal is a possible alternative, as used in Pons-Moll et al. [2017]. Another limitation is the regression model. Currently a two-hidden-layer neural network is used. As the dataset becomes larger and larger and more and more semantic variables need to be considered, this simple regression model may not suffice. A more elaborate regression model that better captures the functionality should be explored.

Chapter 6

Conclusion and future work

Contents

6.1 Conclusion

In this thesis, we have proposed a data-driven learning approach to generate the dynamic geometry of clothing that corresponds to the input inquiry states, including human motion, human body shape, clothing material and so on.

In the research field of dynamic clothing geometry synthesizing, our proposed approach possesses a number of important advantages. Compared with other solutions based on physics simulation, our proposed method does not require intensive computation in the synthesizing phase, and is able to generate shapes that are observed from the real world. Com-

pared with some data-driven approaches, for example, Xu et al. [2014a] and Neophytou and Hilton [2014], our proposed method does not only blend examples from a certain dataset, but learns the underlying shape distribution and the causality between the shape and the controlling variables. Compared with some other data-driven learning approaches, for example, Guan et al. [2012], the proposed method is not limited to learn from synthetic data, which is clean of noise yet hardly able to model complex clothes, but is also able to learn from real 3D captures, which allows much richer data source.

To make use of real 3D captures, we proposed two steps to extract information that is essential to the task of clothing shape modeling in Chapter 3 and Chapter 4. The first information we extract is the human body under the clothing. With the help of an automatic method (Stitched Puppet Zuffi and Black [2015]) for pose initialization and based on a simplified statistical human body shape model (S-SCAPE Jain et al. [2010]) we proposed a first automatic method that extract both human pose and body shape from a raw captured 3D sequence of a moving human with clothing. We simply assume that the human body should always stay within the observed surface and lies as close as possible to it. Estimating human body from a 3D sequence rather than a single 2D images, the proposed method is much more robust to the ambiguity in 2D images and is able to exploit the richer information contained in 3D sequence. Compared with some state-of-the-art methods, we achieve similar quality without any manual intervention. The second information we extract is the motion of the clothing surface. The raw captured 3D sequence does not contain temporal information. Using the estimated human body

and isometric patches, we start the non-rigid registration from sparse corresponding projected anatomic points on the clothing surface and then propagate the correspondence along the surface according to isometric deformation to get a dense surface registration, which finally leads to a non-rigid registration for the whole sequence. This proposed method makes use of the fact the most clothes are not significantly stretchable, and facilitates the isometric matching by starting from previously obtained human body. It is a straightforward and effective solution compared to other methods, as shown in Section 4.4.

With the ability to extract important information from raw captures, our solution can make use of both synthetic and captured data. Along with the versatile learning pipeline, the proposed solution opens various possibilities. As shown in section 5.4, we can use physical simulators to systematically generate clothing deformation with respect to different materials, and learn the functional relationship between geometry to material parameters. From captured data, we can easily obtain example sequences with different clothing fit, and then learn the relationship between clothing deformation and clothing fit parameters. With a much larger dataset, we believe that more unprecedented applications can be explored.

6.2 Discussion and Future work

Modeling the shape space of a moving human in clothes is a very difficult problem. Many works can be done to improve current solutions and many other potential solutions can be explored. The related discussion on detailed aspects has been addressed in the conclusion sections of Chapter

3, 4, and 5. We will only talk about high level topics here.

Single surface model vs body surface plus clothing surface model. In the field of capture-based clothing deformation modeling, there are mainly two basic definition of clothing layer. One considers the entire surface of human as clothing layer while the other one separates the actual clothing from the observed human surface and models dressed human as several meshes: a body mesh plus clothing meshes. The former can be found in the work of Neuphytou and in this thesis; the latter can be represented by works of Pons-Moll et al. [2017] and Zorah et al. [2018]. Separating body and clothing is physically correct and should be the ideal direction to explore in the future, yet currently it still suffers from certain disadvantages. The very first obstacle is the inadequacy of current dataset. It comes from two aspects. On one hand, there is not enough publicly available dataset that contains registered separate clothing. On the other hand, most current dataset of 3D dressed human is not designed to perform clothing segmentation and registration methods that are proposed in Pons-Moll et al. [2017] and Zorah et al. [2018], which require near mono-color single-layer clothing or clothing with marked boundaries. Besides the inadequacy of dataset, the separated layer model also needs to solve interpenetration among body surface and clothing surfaces explicitly, while the single surface model does not. Considering the potential application of this research, we often need to generate shapes that visually plausible. Yet with texture, human visual perception seems to be much more sensitive to interpenetration rather than slightly mis-deformed clothing shapes. An extra interpenetration solver makes the separated layer model more complicated than single layer model. In summary, the single surface

model is simple and suitable for most current dataset and the separated layer model is physically more accurate and potentially capable of modeling challenging scenarios. More research effort should be devoted into separated layer model in the future.

Visual quality vs physical accuracy. Both physics-base methods and learning approaches in clothing deformation modeling wish to be able to generate good synthesis. But the standard of good can be different for different people. Typically, good usually means high visual fidelity in computer graphics, while in computer vision, good usually means close to physical reality. Particularly in terms of clothing deformation modeling, a physically accurate result generally guarantees high visual quality. But it is not true the other way around. Looking through current works on clothing deformation modeling using capture-based learning approaches, it is not difficult to find out that almost all validate the result according to visual quality rather than error against the ground truth (captures). There are several reasons for this. First, the majority of the applications of this research, such as virtual dressing room and special effect in the movie, requires high visual fidelity rather than physical accuracy. Another reason may be that clothing deformation seems to be chaotic. If the initial states of the clothing or the body motion are a little bit different from each other, it seems that they tend to develop into much larger different shapes in the future. Due to such empirical observation, it is difficult to propose a fair comparison between synthesis and captured ground truth. Apart from the above two reasons, it is also obvious that good visual quality is a much easier object to achieve compared with physical accuracy. Therefore, high visual quality is considered the first target and major focus

of the community is put on visual quality, instead of physical accuracy. However, we should keep in mind that the ultimate goal of our research should be modeling the mechanism of our physical world, particularly the mechanism that governs clothing deformation. It can be either explicit modeling, such as physics-based simulation, or implicit modeling, such as learning approaches. To evaluate the modeling of the reality, we need to compare with real world ground truth. That is where physical accuracy matters. If we have a further vision about clothing deformation modeling, it is not difficult to figure out the above mentioned three reasons are not sufficient. Specifically, apart from applications of generating vivid shapes, the research could also contribute in other applications where physical accuracy matters. For example, given an observed sequence of deformed clothing, it can predict physically plausible deformation in the next frames, so that temporal registration of dressed human could benefit. Regarding the suspect of chaotic nature, more research effort is required. Questions that need to be answer include: if both low frequency shape and high frequency shape chaotic; instead of giving a single solution to the deformation in the next frame, can we compute a probability distribution of all the shapes in the next frame; if clothing deformation is chaotic, like the weather system, can we approximate the initial state according to observed history and then be able to predict the future deformation just as people have been doing in the weather forecast? To summarize, visual quality is modestly a proper goal for current research on clothing deformation, but physical accuracy is our next target and should start being considered in our current research.

Dataset For any data-driven approach, the primary problem is to obtain the data. It involves data generation/collection, data cleaning and preprocessing. Yet the publicly available 3D sequences of dressed human are limited in both quantity and quality. In terms of quantity, besides limited number of sequences and frames, existing dataset also lacks structure, in the sense that subject, clothing and motion are not well seperated and combined. For example, it is difficult to find sequences of different subjects wearing the same clothing and performing the same motion. But such sequences will be of great help to study the influence of body shape to clothing deformation. In terms of quality, apart from the requirement on spacial and temporal resolution, we also prefer the data to be as close as possible to the physical ground truth, instead of visually plausible. For synthetic data, this means the simulators should represent the real world with high fidelity. And for captured data, this means low noise in raw data and a robust non-rigid surface registration that follows physical ground truth. In the future, building a rich dataset of 3D dressed human should be considered a crucial and urgent task.

Deep learning We have witnessed the great success of deep learning methods in recent years. Started as a solution to classification problem, the deep learning method has evolved into a versatile tool, which is able to compress high dimensional space with Autoencoder, generate data with Generative Adversarial Neural network and predict sequence with Long Short Term Memory network. Those elaborate deep learning methods cover some key components of clothing deformation modeling. Now with more captured and synthesized data, we can start to explore the possibility of modeling the problem using deep learning. Zorah

et al. [2018] predicts coarse PCA shape coefficient using Long Short Term Memory network and generates detailed wrinkles using 2D conditional Generative Adversarial Network, where normal map is used. A further research in this direction is of great interest, including possible topics such as learning the low dimensional clothing deformation representation using Autoencoder, learning the clothing shape variation across different clothing type, and learning clothing dynamics on unregistered sequences using voxel or point cloud based deap learning methods.

Other solutions Apart from data-driven learning approach, there are other data-driven methods and physics based approaches that have been explored. Many of them achieve high visual fidelity or real-time performance. People should continue working on those solutions and trying to combining different solution to compensate each other. For example, state-of-the-art high resolution physics-based clothing simulation can generate detailed clothing geometry, but usually suffer from heavy computation, thus difficult for real-time application. Blending example shapes from dataset to generate new shapes as described in Xu et al. [2014a] is computationally cheap in the synthesizing phase. We can combine this data-driven method and the state-of-the-art physics-based simulation to potentially have high-quality real-time performance. The major focus for this combination should be placed on the aspect of efficiently generate and store the dataset, as well as the aspect of the blending techniques. Another possible combination is that we can combine low resolution physics-based simulation for low frequency geometry and Zorah et al. [2018] for high frequency geometry synthesis, using neural network to augment the simulation result.

Bibliography

OpenNN library. http://www.opennn.net/. Cited on page 96.

B. Allain, J.-S. Franco, E. Boyer, and T. Tung. On mean pose and variability of 3d deformable models. In *European Conference on Computer Vision*, pages 284–297. Springer, 2014. Cited on pages x, x, x, 65, 66, 80, 81, 82, and 83.

B. Allain, J.-S. Franco, and E. Boyer. An efficient volumetric framework for shape tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 268–276, 2015. Cited on pages x, x, x, x, x, 6, 34, 36, 65, 66, 80, 81, 82, 83, and 84.

B. Allen, B. Curless, and Z. Popović. The space of human body shapes: reconstruction and parameterization from range scans. In *ACM transactions on graphics (TOG)*, volume 22, pages 587–594. ACM, 2003. Cited on pages 69, 75, 76, and 79.

D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *ACM transactions on graphics (TOG)*, volume 24, pages 408–416. ACM, 2005a. Cited on pages 3, 19, and 20.

D. Anguelov, P. Srinivasan, H.-C. Pang, D. Koller, S. Thrun, and J. Davis. The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces. In *Advances in neural information processing systems*, pages 33–40, 2005b. Cited on page 20.

M. Aubry, U. Schlickewei, and D. Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1626–1633. IEEE, 2011. Cited on page 31.

A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, pages 15–29. Springer, 2008. Cited on pages 27, 42, 47, and 51.

D. Baraff and A. Witkin. Large steps in cloth simulation. In *Conference on Computer Graphics and Interactive Techniques*, 1998. Cited on pages 2 and 35.

H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. Cited on pages 15 and 31.

S. Biasotti, A. Cerri, A. Bronstein, and M. Bronstein. Recent trends, applications, and perspectives in 3d shape similarity assessment. In *Computer Graphics Forum*, volume 35, pages 87–119. Wiley Online Library, 2016. Cited on page 32.

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 *European Conference on Computer Vision*, pages 561–578. Springer, 2016. Cited on page 28.

R. Bridson, S. Marino, and R. Fedkiw. Simulation of clothing with folds and wrinkles. In *Symposium on Computer Animation*, 2003. Cited on page 35.

A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Efficient computation of isometry-invariant distances between surfaces. *SIAM Journal on Scientific Computing*, 28(5):1812–1836, 2006. Cited on page 33.

A. Brunton, M. Wand, S. Wuhrer, H.-P. Seidel, and T. Weinkauf. A low-dimensional representation for robust partial isometric correspondences computation. *Graphical Models*, 76(2):70–85, 2014. Cited on pages 69, 74, and 79.

C. Budd, P. Huang, M. Klaudiny, and A. Hilton. Global non-rigid alignment of surface sequences. *International Journal of Computer Vision*, 102(1-3):256–270, 2013. Cited on pages 34 and 77.

C. Cagniart, E. Boyer, and S. Ilic. Free-form mesh tracking: a patch-based approach. In *CVPR 2010-IEEE Conference on Computer Vision and Pattern Recognition*, pages 1339–1346. IEEE, 2010. Cited on pages 6 and 33.

Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016. Cited on page 61.

Y. Chen, Z. Liu, and Z. Zhang. Tensor-based human body modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 105–112, 2013. Cited on page 25.

Z. Chen and H.-J. Lee. Knowledge-guided visual perception of 3-d human gait from a single image sequence. *IEEE transactions on Systems, Man, and Cybernetics*, 22(2):336–342, 1992. Cited on pages vii and 20.

Z.-Q. Cheng, Y. Chen, R. R. Martin, T. Wu, and Z. Song. Parametric modeling of 3d human body shapea survey. *Computers & Graphics*, 71: 88–100, 2018. Cited on page 20.

G. K. Cheung, S. Baker, and T. Kanade. Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–375. IEEE, 2003. Cited on page 14.

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(4), 2015. Cited on page 36.

F. Cordier and N. Magnenat-Thalmann. A data-driven approach for real-time clothes simulation. In *Computer Graphics and Applications, 2004. PG 2004. Proceedings. 12th Pacific Conference on*, pages 257–266. IEEE, 2004. Cited on pages 1 and 2.

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 (TOG)*, 27(3):98, 2008. Cited on pages 59 and 65.

E. de Aguiar, L. Sigal, A. Treuille, and J. K. Hodgins. Stable spaces for real-time clothing. *ACM Transactions on Graphics*, 29(4), 2010. Cited on page 35.

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. *ACM Transactions on Graphics (TOG)*, 35(4):114, 2016. Cited on pages 5, 14, 17, 31, and 65.

M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi. Motion2fusion: real-time volumetric performance capture. *ACM Transactions on Graphics (TOG)*, 36(6):246, 2017. Cited on page 17.

M. Douze, J.-S. Franco, and B. Raffin. QuickCSG: Arbitrary and Faster Boolean Combinations of N Solids. Research Report RR-8687, Inria - Research Centre Grenoble – Rhône-Alpes ; INRIA, Mar. 2015. URL https://hal.inria.fr/hal-01121419. Cited on page 17.

O. Freifeld and M. J. Black. Lie bodies: A manifold representation of 3d human shape. In *European Conference on Computer Vision*, pages 1–14. Springer, 2012. Cited on page 25.

Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010. Cited on pages 14 and 15.

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 *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1746–1753. IEEE, 2009. Cited on page 32.

A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 963–968. Ieee, 2011. Cited on pages 14 and 15.

M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. 2007. Cited on pages 14 and 15.

R. Goldenthal, D. Harmon, R. Fattal, M. Bercovier, and E. Grinspun. Efficient simulation of inextensible cloth. In *ACM Transactions on Graphics (TOG)*, volume 26, page 49. ACM, 2007. Cited on page 2.

K. Grauman, G. Shakhnarovich, and T. Darrell. A bayesian approach to image-based visual hull reconstruction. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2003. Cited on page 14.

P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black. Drape: Dressing any person. *ACM Transactions on Graphics*, 31(4), 2012. Cited on pages 35, 89, and 112.

K. Guo, F. Xu, Y. Wang, Y. Liu, and Q. Dai. Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3083–3091, 2015. Cited on page 31.

K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (TOG)*, 36(3):32, 2017. Cited on page 32.

N. Hasler, C. Stoll, B. Rosenhahn, T. Thormählen, and H.-P. Seidel. Estimating body shape of dressed humans. *Computers & Graphics*, 33(3): 211–216, 2009a. Cited on page 27.

N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *Computer Graphics Forum*, volume 28, pages 337–346. Wiley Online Library, 2009b. Cited on page 25.

T. Helten, A. Baak, G. Bharaj, M. Muller, H.-P. Seidel, and C. Theobalt. Personalization and evaluation of a real-time depth-based full body tracker. In *2013 International Conference on 3D Vision (3DV)*, pages 279–286. IEEE, 2013. Cited on pages 28, 40, and 41.

R. Hess. *Animating with Blender: how to create short animations from start to finish*. Focal Press, 2012. Cited on page 1.

M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision*, pages 362–379. Springer, 2016. Cited on page 14.

A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. Moviereshape: Tracking and reshaping of humans in videos. In *ACM Transactions on Graphics (TOG)*, volume 29, page 148. ACM, 2010. Cited on pages 3, 22, 68, and 112.

H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. Cited on pages 26 and 61.

J. M. Kaldor, D. L. James, and S. Marschner. Efficient yarn-based cloth with adaptive contact linearization. *ACM Transactions on Graphics*, 29 (4), 2010. Cited on page 35.

A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on page 27.

L. Kavan, S. Collins, J. Žára, and C. O'Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 39–46. ACM, 2007. Cited on page 1.

M. Klaudiny and A. Hilton. Cooperative patch-based 3d surface tracking. In *2011 Conference for Visual Media Production*, pages 67–76. IEEE, 2011. Cited on page 33.

S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen. Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 10. ACM, 2017. Cited on page 26.

A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162, 1994. Cited on pages 14 and 16.

A. Letouzey and E. Boyer. Progressive shape models. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 190–197. IEEE, 2012. Cited on pages 68 and 72.

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*, pages 165–172. ACM Press/Addison-Wesley Publishing Co., 2000. Cited on page 2.

H. Li, R. W. Sumner, and M. Pauly. Global correspondence optimization for non-rigid registration of depth scans. In *Computer graphics forum*, volume 27, pages 1421–1430. Wiley Online Library, 2008. Cited on page 31.

H. Li, B. Adams, L. J. Guibas, and M. Pauly. Robust single-view geometry and motion reconstruction. In *ACM Transactions on Graphics (TOG)*, volume 28, page 175. ACM, 2009. Cited on pages 31 and 91.

J. Li, G. Daviet, R. Narain, F. Bertails-Descoubes, M. Overby, G. Brown, and L. Boissieux. An implicit frictional contact solver for adaptive cloth simulation. *ACM Transactions on Graphics*, 37(4):15, 2018. Cited on page 105.

L. Li. Time-of-flight camera–an introduction. *Texas Instruments-Technical white paper*, 2014. Cited on page 17.

D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989. Cited on page 80.

J. Long, K. Burns, and J. J. Yang. Cloth modeling and simulation: a literature survey. In *International Conference on Digital Human Modeling*, pages 312–320. Springer, 2011. Cited on page 2.

M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34 (6), 2015. Cited on pages 3, 19, 23, 24, and 91.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. Cited on pages 15 and 31.

T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer vision and image understanding*, 81 (3):231–268, 2001. Cited on page 19.

A. Mustafa, H. Kim, and A. Hilton. 4d match trees for non-rigid surface alignment. In *European Conference on Computer Vision*, pages 213–229. Springer, 2016. Cited on page 34.

A. Neophytou and A. Hilton. Shape and pose space deformation for subject specific animation. In *3D Vision-3DV 2013, 2013 International Conference on*, pages 334–341. IEEE, 2013. Cited on pages 19, 22, 89, and 91.

A. Neophytou and A. Hilton. A layered model of human body and garment deformation. In *3DV*, pages 171–178, 2014. Cited on pages vii, 3, 28, 29, 36, 37, 40, 41, 42, 51, 59, 60, 68, and 112.

R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition, pages 343–352, 2015. Cited on pages 14, 17, 31, 36, and 65.

M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (TOG)*, 31(4):30, 2012. Cited on page 31.

L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67:276–286, 2017. Cited on pages 23, 42, 54, and 91.

G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics (TOG)*, 34(4):120, 2015. Cited on page 24.

G. Pons-Moll, S. Pujades, S. Hu, and M. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics*, 36 (4), 2017. URL `http://dx.doi.org/10.1145/3072959.3073711`. Cited on pages vii, x, x, 36, 37, 93, 98, 99, 110, and 114.

T. Popa, Q. Zhou, D. Bradley, V. Kraevoy, H. Fu, A. Sheffer, and W. Heidrich. Wrinkling captured garments using space-time data-driven deformation. In *Computer Graphics Forum*, volume 28, pages 427–435. Wiley Online Library, 2009. Cited on page 74.

A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black. Generating 3d faces using convolutional mesh autoencoders. *arXiv preprint arXiv:1807.10267*, 2018. Cited on page 26.

K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, and S. Fleming. Civilian american and european surface anthropometry resource (cae-

sar), final report. volume 1. summary. Technical report, SYTRONICS INC DAYTON OH, 2002. Cited on page 23.

K. Rohr. Human movement analysis based on explicit motion models. In *Motion-based recognition*, pages 171–198. Springer, 1997. Cited on pages vii and 20.

J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):245, 2017. Cited on page 25.

B. Rosenhahn, U. Kersting, K. Powell, R. Klette, G. Klette, and H.-P. Seidel. A system for articulated tracking incorporating a clothing model. *Machine Vision and Applications*, 18(1):25–40, 2007. Cited on pages 6 and 32.

R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. Citeseer, 2009. Cited on page 31.

Y. Sahillioglu and Y. Yemez. 3d shape correspondence by isometry-driven greedy optimization. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 453–458. IEEE, 2010. Cited on page 33.

T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. *ACM SIGGRAPH computer graphics*, 20(4):151–160, 1986. Cited on page 34.

H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *European Conference on Computer Vision*, 2000. Cited on page 19.

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. *International Journal of Computer Vision*, 87(4), 2010. Cited on page 19.

L. Sigal, M. Mahler, S. Diaz, K. McIntosh, E. Carter, T. Richards, and J. Hodgins. A perceptual control space for garment simulation. *ACM Transactions on Graphics*, 34(4), 2015. Cited on page 35.

M. Slavcheva, M. Baust, D. Cremers, and S. Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, page 7, 2017. Cited on page 34.

C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *The International Journal of Robotics Research*, 22(6), 2003. Cited on page 19.

O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007. Cited on page 33.

C. Stoll, J. Gall, E. De Aguiar, S. Thrun, and C. Theobalt. Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics*, 29(6), 2010. Cited on pages 35 and 51.

R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. In *ACM Transactions on Graphics (TOG)*, volume 26, page 80. ACM, 2007. Cited on page 33.

G. K. Tam, Z.-Q. Cheng, Y.-K. Lai, F. C. Langbein, Y. Liu, D. Marshall, R. R. Martin, X.-F. Sun, and P. L. Rosin. Registration of 3d point clouds and meshes: a survey from rigid to nonrigid. *IEEE transactions on visualization and computer graphics*, 19(7):1199–1217, 2013. Cited on page 32.

A. Tevs, A. Berner, M. Wand, I. Ihrke, M. Bokeloh, J. Kerber, and H.-P. Seidel. Animation cartographyintrinsic reconstruction of shape and motion. *ACM Transactions on Graphics (TOG)*, 31(2):12, 2012. Cited on page 85.

F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *European conference on computer vision*, pages 356–369. Springer, 2010. Cited on page 31.

O. Van Kaick, H. Zhang, G. Hamarneh, and D. Cohen-Or. A survey on shape correspondence. In *Computer Graphics Forum*, volume 30, pages 1681–1707. Wiley Online Library, 2011. Cited on page 32.

G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. *arXiv preprint arXiv:1804.04875*, 2018. Cited on page 28.

D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics*

(TOG), volume 27, page 97. ACM, 2008. Cited on pages x, x, 5, 6, 30, 32, 59, 65, 80, 82, and 96.

P. Volino, N. Magnenat-Thalmann, and F. Faure. A simple approach to nonlinear tensile stiffness for accurate cloth simulation. *ACM Transactions on Graphics*, 28(4), 2009. Cited on page 35.

H. Wang, J. F. O'Brien, and R. Ramamoorthi. Data-driven elastic models for cloth: modeling and measurement. *ACM Transactions on Graphics*, 30(4), 2011. Cited on page 105.

A. Weiss, D. Hirshberg, and M. J. Black. Home 3d body scans from noisy image and range data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1951–1958. IEEE, 2011. Cited on pages 28, 40, and 41.

S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang. Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding*, 127:31–42, 2014. Cited on pages 28, 40, 41, 54, 59, 60, and 68.

W. Xu, N. Umentani, Q. Chao, J. Mao, X. Jin, and X. Tong. Sensitivity-optimized rigging for example-based real-time clothing synthesis. *ACM Transactions on Graphics*, 33(4), 2014a. Cited on pages 35, 112, and 118.

W. Xu, N. Umentani, Q. Chao, J. Mao, X. Jin, and X. Tong. Sensitivity-optimized rigging for example-based real-time clothing synthesis. *ACM Transactions on Graphics (TOG)*, 33(4):107, 2014b. Cited on page 3.

J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of human body shape in motion with wide clothing. In *European Conference*

on Computer Vision, pages 439–454. Springer, 2016. Cited on pages 96 and 106.

T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. *arXiv preprint arXiv:1804.06023*, 2018. Cited on pages 14, 31, and 65.

T. Yu12, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. 2017. Cited on page 14.

P. Zanuttigh, G. Marin, C. Dal Mutto, F. Dominio, L. Minto, and G. M. Cortelazzo. Operating principles of structured light depth cameras. In *Time-of-Flight and Structured Light Depth Cameras*, pages 43–79. Springer, 2016. Cited on page 16.

C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. Cited on pages 5, 28, 29, 36, 61, and 68.

X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. Cited on page 61.

M. Zollhöfer, E. Sert, G. Greiner, and J. Süßmuth. Gpu based arap deformation using volumetric lattices. In *Eurographics (Short Papers)*, pages 85–88, 2012. Cited on page 34.

M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (TOG)*, 33(4):156, 2014. Cited on pages 30, 31, and 33.

M. Zollhöfer, P. Stotko, A. Görlitz, C. Theobalt, M. Nießner, R. Klein, and A. Kolb. State of the Art on 3D Reconstruction with RGB-D Cameras. *Computer Graphics Forum (Eurographics State of the Art Reports 2018)*, 37(2), 2018. Cited on pages 30 and 33.

L. Zorah, D. Cremers, and T. Tony. Deepwrinkles: Accurate and realistic clothing modeling. In *European Conference on Computer Vision*. Springer, 2018. Cited on pages 114, 117, and 118.

S. Zuffi and M. J. Black. The stitched puppet: A graphical model of 3d human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3537–3546, 2015. Cited on pages viii, 9, 42, 45, 48, and 112.