High-fidelity Human Body Modelling from User-generated Data

Zongyi Xu

A thesis submitted to the University of London for the degree of
Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary University of London

August 2018

Abstract

Building high-fidelity human body models for real people benefits a variety of applications, like fashion, health, entertainment, education and ergonomics applications. The goal of this thesis is to build visually plausible human body models from two kinds of user-generated data: low-quality point clouds and low-resolution 2D images.

Due to the advances in 3D scanning technology and the growing availability of cost-effective 3D scanners to general users, a full human body scan can be easily acquired within two minutes. However, due to the imperfections of scanning devices, occlusion, self-occlusion and untrained scanning operation, the acquired scans tend to be full of noise, holes (missing data), outliers and distorted parts. In this thesis, the establishment of shape correspondences for human body meshes is firstly investigated. A robust and shape-aware approach is proposed to detect accurate shape correspondences for closed human body meshes. By investigating the vertex movements of 200 human body meshes, a robust non-rigid mesh registration method is proposed which combines the human body shape model with the traditional nonrigid ICP. To facilitate the development and benchmarking of registration methods on Kinect Fusion data, a dataset of user-generated scans is built, named Kinect-based 3D Human Body (K3D-hub) Dataset, with one Microsoft Kinect for XBOX 360.

Besides building 3D human body models from point clouds, the problem is also tackled which estimates accurate 3D human body models from single 2D images. A state-of-the-art parametric 3D human body model SMPL is fitted to 2D joints as well as the boundary of the human body. Fast Region based CNN and deep CNN based methods are adopted to detect the 2D joints and boundary for each human body image automatically. Considering the commonly encountered scenario where people are in stable poses at most of the time, a stable pose prior is introduced from CMU motion capture (mocap) dataset for further improving the accuracy of pose estimation.

Acknowledgements

I would like to first express my most sincere gratitude to my supervisor, Dr. Qianni Zhang. Thanks for her continuous supports during my PhD. It is an excellent experience to dive into the research ocean under her extraordinary guidance. I would also deeply appreciate Prof. Ebroul Izquierdo for his sharp insight into research and patience to every tiny research problem I raise. I would also thank Dr. Pengwei Hao for his many valuable suggestions on both research and teaching. It was my fortune and honour to have the guidance from all of you, which shapes my way of research, changes my way of thinking and profoundly influences my vision of the future.

I am deeply grateful to the colleagues and friends who I worked with - Shiyang Cheng, Patrycia Klavdianos, Tuanfeng Y. Wang and Heng Yang. My research can not be completed without the insightful communication with all of them. Meanwhile, I would like to appreciate Prof. Dongdong Zhang, Dr. Faranak Sobhani, Dr. Petar Palaek and Dr. Juan Miranda, who show me the so many possibilities in and after PhD. Additionally, my PhD journey would not have been completed without the help of all MMVers (too many names!). I would also like to thank all my current and previous roommates - Qian Yu, Jingya Wang, Jing Tian, Zhaoyang Xu and Wanlin Lin et al for their kindness, supports and sharing in the life. Special thanks go to Xuefeng Chen for his accompany and sharing my happiness and depression in every step of my PhD.

Most of all, I would deeply appreciate my family, especially my parents and grandparents, for their love and unlimited supports.

Contents

List of Figures

List of Tables

List of abbreviations

3D-GAN	3D Generative-adversarial network
ANICP	Active Nonrigid Iterative Closest Point
ARAP	as-rigid-as-possible
C2FCM	coarse-to-fine combinatorial matching
CT	Computed tomography
FAUST	Fine Alignment Using Scan Texture
FOV	Field of view
GMM	Gaussian Mixture Model
GPS	Global Point Signature
HKS	Heat Kernel Signature
ICP	Iterative Closest Point
IR	Infrared radiation
K3d-hub	Kinect based human body dataset
LNBPs	Local Normal Binary Patterns
LD-SIFT	Local depth SIFT
MABR	Multilevel Active Body Registration
MDS	Multi-Dimensional Scaling
mocap	Motion capture
Mosh	Motion and Shape Capture from Sparse Markers
MRI	Magnetic resonance imaging
NICP	Nonrigid Iterative Closest Point
PCA	Principal component analysis
RMS	Root mean square
SCAPE	Shape Completion and Animation of People
SDF	signed distance field
SI-hks	scale-invariant HKS
SMPL	A Skinned Multi-person Linear Model
SP	Stitched puppet
SVD	Singular Value Decomposition
TSDF	Truncated signed distance function

Statement of Originality

I, Zongyi Xu, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature:

Date: 1th of August, 2018

Details of collaboration and publications:

Portions of this research detailed in this thesis have been presented in international scholarly publications, as follows :

- Chapter 3: was published as a conference publication in the **22nd International Conference of Multimedia Modeling**.

  Zongyi Xu, and Qianni Zhang. "Symmetry-aware human shape correspondence using skeleton." In International Conference on Multimedia Modeling, pp. 632-641. Springer, Cham, 2016.

- Chapter 4: The work in Section 4.2 was published in the **Journal of Multimedia Systems**.

  Zongyi, Xu, Qianni Zhang, and Shiyang Cheng. "Multilevel active registration for kinect human body scans: from low quality to high quality." Multimedia Systems, 24(3), pp.257-270.

  The work in Section 4.3 was published by the **International Conference on Multimedia and Expo (ICME)**.

  Zongyi Xu, Qianni Zhang. "Region based User-generated Human Body Scan Registration." In IEEE International Conference on Multimedia and Expo (ICME), 2018.

- Chapter 5: The work was published in the **Eurographics** as poster.

  Zongyi Xu, Qianni Zhang. "Boundary-Aided Human Body Shape and Pose Estimation from a Single Image for Garment Design and Manufacture." Eurographics (Poster), 2018.

Chapter 1

Introduction

The past few decades have witnessed the dramatic changes in the ways of acquisition and generation of 3D objects and scenes. Traditionally, professional artists manually build models with 3D modelling tools, which takes a few days to several months for a faithful outcome. The later appearance of 3D scanning equipment facilitates quick and easy acquirement of 3D objects [141, 178, 120]. Today, we are surrounded by numerous low-cost 3D scanning equipment and 3D-related consumptions such as virtual reality and augmented reality, stimulating the interests of the general public to create 3D models for the objects around us. The human body is one of the most interesting subjects to model in 3D.

Being the carrier of our idea, our spirit and our stories, a human body is essential to the physical interaction we have with the world and between each other. Modelling 3D human body models from the data, like point clouds from affordable scanners or single selfies from smart phones, that is acquired by the general users, plays important roles in many applications that are closely related to mass life, such as fashion industry, online social networking [114, 82], fashion industry [129, 62], education [1] and even military [137]. By building high-fidelity human body models, we can create realistic and collaborative

[1] http://www.reveriefp7.eu/

online environments which bring together realistic inter-personal communication and interactions. In the fashion industry, through building realistic human body avatars for customers, the virtual dressing system can provide more accurate size and style recommendation. Realistic human body models for real people are also a revolution in filmmaking. Instead of using real actors, their virtual counterparts can be used as a substitute in dangerous scenes. Indeed, simulating high-fidelity 3D virtual human body avatar is tightly coupled with the next-generation applications.

Nowadays, the acquirement of high-quality human body meshes is still expensive. Besides building meshes with professional tools, which requires expertise and takes time and efforts, realistic human body scans can be captured with large and high-resolution scanning systems [15, 19, 197]. The 3D stereo capture system in Max Planck Institute [2] has 22 stereo capture modules tied on an aluminium frame, occupying a whole room. Some online 3D model providers use a large number of cameras to build a high-quality model. For example, TEN 24 [3] uses 170 cameras to capture a full body and integrates a head capture system with 40 cameras to build high-resolution head models. In the aforementioned scanning systems, besides a large number of expensive cameras and the complex setup, a specific space for scanning is also needed. Although high-resolution and delicate human 3D models can be acquired, these requirements are hard to be satisfied by massive consumer groups in common markets.

In most applications that are involved in our daily life, it is highly desirable to acquire 3D human avatars straightforwardly, accurately, and economically, with simple and affordable data sources that are easy to access by general users. Here, two commodity 3D scanners are taken as an example - Microsoft Kinect [4] and Structure Sensor [5]. Although these sensors are inexpensive and family-affordable, the captured data is usually low-quality. As we can see in Figure 1.1, the depth images in the bottom are very noisy with a number of no-data values delivered while performing depth measurement represented

[2] https://ps.is.tuebingen.mpg.de/pages/3dcapture
[3] http://ten24.info/3d-scanning/
[4] http://www.xbox.com/kinect
[5] https://structure.io/

by dark blue in the left bottom for Kinect and black in the right bottom for Structure sensor. In order to acquire high quality scanning results, some works start from refining

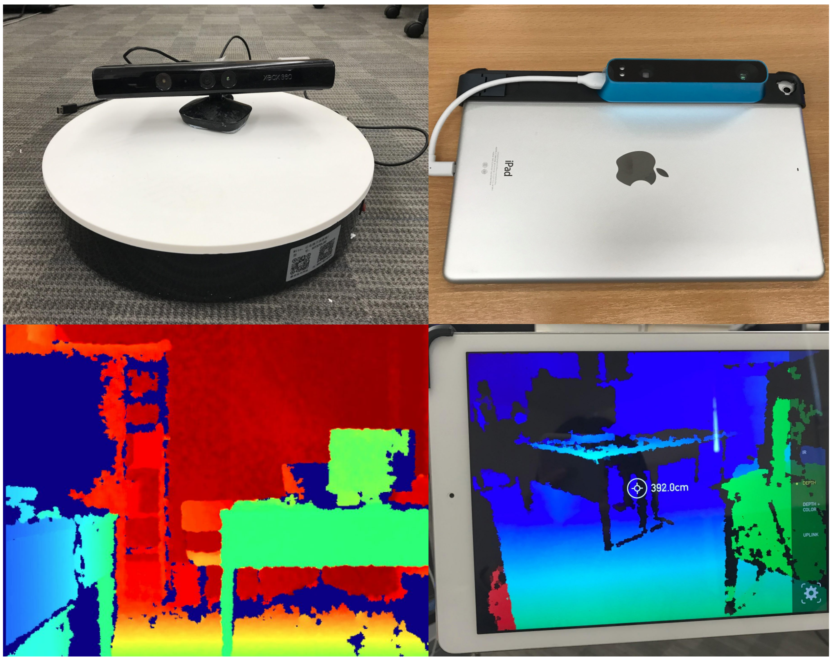


Figure 1.1: The affordable devices we used for capturing data. The left is the Microsoft Kinect and the right is structure sensor.

the depth image. In [153], a depth layer is used to produce well-defined depth edges. Moving objects are first detected to improve edge stability before filling holes [187]. A temporal consistency constraint is combined with bilateral filtering in [26] to predict the depth value in the undetected region. Although these denoising approaches largely reduce the noise near the edge and holes, the depth error increases with the distance between the scanner. When the untrained/unprofessional users scan bodies, the instability and changes in the distances make the acquired points distort heavily. Currently, some template based surface modelling methods [206, 192, 93] are proposed to acquire the high-quality meshes from noisy scans. In these methods, either the deformation is confined to naked human body models or each template has to be constructed for each target. Other template-free reconstruction methods [109, 89, 110] reconstruct the human body frame be frame in real time or offline. These systems are wither easily to be disturbed by the capturing environment so that the acquired meshes are not accurate enough or only constructed from 3D data without considering another common data source of 2D images.

In this thesis, the goal is to start with user-generated data that is accessible to general users and end with believable human body models. Two kinds of data sources are considered: as shown in Figure 1.2, noisy human body point clouds and single 2D images. The observation of point cloud is captured with the setup of a running turntable and one Microsoft Kinect xbox360. The detailed capturing process is described in Section 4.1. The observation of single images is obtained with the smart phones from the front view. The quality of the image, positions of the cameras and the background are not required when taking pictures. The experiments show that using single images, it is possible to estimate accurate enough human body shapes.

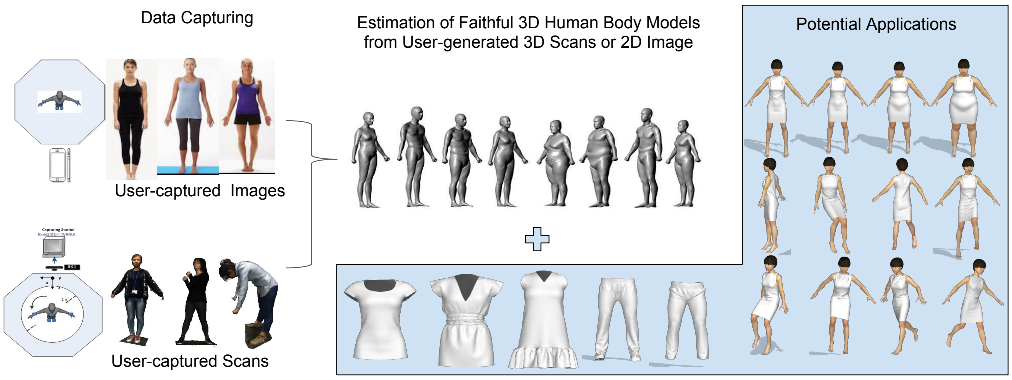

Data Capturing
User-captured Images
User-captured Scans
Estimation of Faithful 3D Human Body Models from User-generated 3D Scans or 2D Image
Potential Applications

Figure 1.2: Building human body models from two types of user-generated data sources and the potential applications. In this thesis, human body models are constructed from point clouds as well as single 2D images. An example of the potential application is given afterwards. Clothing model is retargeted to the acquired models for personalized virtual try-on applications.

The goal of this thesis is to improve the state of the art in the human body modelling from user-generated data which are point clouds and 2D images, by improving the robustness and accuracy of algorithms. Please note that the 2D image-based human body modelling is based on single images rather than reconstructing from multi-view images.

1.1 Problem Statement

The primary problem is the modelling of human bodies including shapes and poses from observations. The observations being discussed in this thesis are two types: point clouds and single 2D images, both of which are captured by general users. As shown in Figure 1.1, the devices that are used for capturing data are low-cost, family-affordable scanners, like Microsoft Kinect XBOX or smart phones. A single hand-held scanner is used to capture the human body point clouds. While the recent advances in 3D scanning techniques contribute to 3D mesh acquisition, the quality of the scans tends to be greatly affected by the light condition, the surrounding space and the scanning manners. Besides, self-occlusion and turbulences during scanning are very common in the user-generated scan data. These factors will cause noise, holes and distorted parts. Some examples of flaws in the user-generated scans are shown in Figure 1.3. It can be seen that the captured hand in Figure 1.3 (a) is so distorted that it does not even keep the shape of a hand. In Figure 1.3 (b), big holes exist on the top of the head. In addition, the example shown in Figure 1.3 (c) presents that part of the hand shape is missing and holes are prevalent on the surface, especially near armpit and crotch.

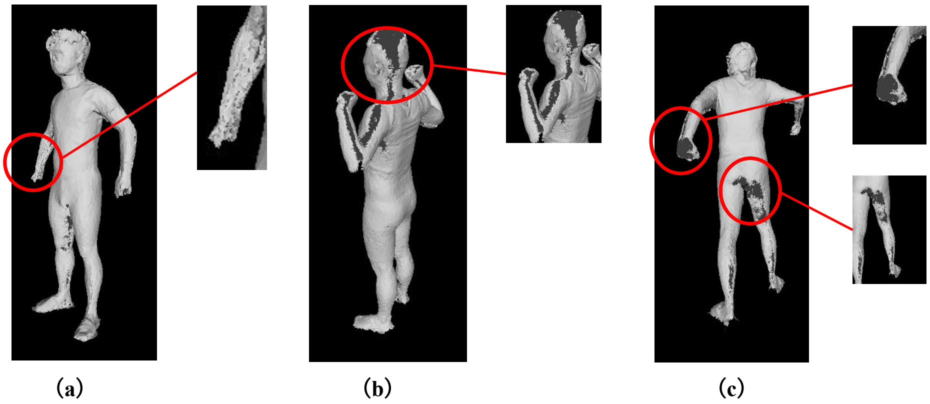


Figure 1.3: Some examples of the flaws existing in the user-generated scans.

More importantly, in everyday scanning for ordinary people, the requirement of naked scanning is hard to meet. Even though we ask subjects to wear tight clothes, some muscle and soft tissue details are squeezed to change or shielded by the clothes. Therefore, the modelling method must be robust to noise, outliers, occlusions as well as be able to describe the deformation caused by the worn clothes on targets. In this research, the modelling of the human body from user-generated scanning data is implemented by fitting a high-quality template mesh to the target scans. The key challenges of human body registration from low-quality scans are tackled. In this scenario, we can hardly acquire naked body scans but assume that scans with tight clothes can be obtained. Dealing with scanned targets wearing loose clothes is beyond the scope of this thesis because body features are fully covered.

Another common data source is the 2D images we take everyday with smart phones. The reconstruction from images taken with multiple cameras from different views is not considered because in practice, it can hardly be satisfied to have several synchronized cameras be around. It is difficult to establish the required capturing environment. Therefore, to guarantee the usability of the proposed methodologies in day-to-day scenarios, the constraint of only relying on single 2D images is imposed in this research. The problem of human body modelling from user-generated image data is defined as building a believable model from a single image in the front view. The built model can satisfy the potential applications, such as personalized virtual try-on, anthropometric measurement [9] and image retouching techniques [204] et al.

1.2 Motivation

The motivation of this research can be distinguished in two aspects: the scientific and practical motivation.

The scientific motivation of this research is to robustly build high-fidelity human body models from point clouds with various imperfections and single 2D images. In the case of low-quality point clouds, human body shape models are defined in both holistic and regional level. These shape models describe the principle of the movements of each vertex for the human body models. The combination of the shape models with the classic non-rigid registration methods improves the registration robustness significantly. In the case of low-resolution single 2D images, deep learning based methods are used to detect the joints and boundary for images. These automatically-detected clues are used to deform the parametric human body model, SMPL [95], to infer the human body shapes and poses. This addresses the depth-lacking problem for the human body modelling using single 2D images.

Then practical motivations are to facilitate general users to simulate, manipulate and use 3D human body models straightforwardly with the most common data source and simplest acquirement manner. Unlike scanning human body with complex and highly accurate scanners and scanning setups [79, 198, 15, 103, 85], high-fidelity human body models are acquired here from noisy, obscure human body scans captured from ordinary people or single 2D images captured with smart phones. This relieves 3D artists from tedious 3D modelling work and enables ordinary people to access their own realistic 3D human body models whenever and wherever possible. It is helpful to discover and design potential applications of high-fidelity human body models in education, entertainment, life and health domains. In the following subsections, the potential practical applications of high-fidelity human body modelling are presented.

1.2.1 Augmented/Virtual Reality Industry

The rapid development of social networks and communication technologies has promoted people's social life to a whole new level. People communicate with their family and friends instantly using social networks such as Facebook, WhatsApp, Twitter, exchanging messages and audio-visual content even if they live far away from each other, but the communication is limited to these modalities: text, voice and videos. Augmented/Virtual reality gains more attention these years due to the desire for immersion in the interaction environment, and virtual human avatar is the core element to enable immersive experience [138, 159, 200, 64]. Instead of interacting with a pre-designed human avatar, a believable virtual avatar of real people contributes to improving users' immersive experience in social networking applications and others. For example, Holoportation [114] merges the real-time and realistic human body reconstruction into the interaction in virtual reality, providing an immersive experience of communication between two remote users, almost as if they are in the same physical space.

1.2.2 Clothing Industry

Online clothes shopping reduces the cost for the retailers, providing the convenience for customers to search in a wider range of products and help overcome geographical limitations. According to the latest PwC report [6], in the year of 2017, over 50% of global shoppers prefer to purchase clothing and footwear online. However, this newly emerged industry continuously suffers from increased dissatisfaction from consumers due to the difficulty in clothes fitting which is the main reason for the high return rate over 25%. To improve the fitting experience in fashion ecommerce, the technologies that enable the realistic individualised virtual fitting at a low cost is of crucial importance. Moreover, Hong et al [68] design a virtual reality system based on the 3D body scans to design garments for the physically disabled people. Therefore the precise estimation of human body shape and reconstruction of a 3D virtual human based on user-generated data from low-cost 3D scanners or even normal smart phone cameras is the key to solve the fitting problem.

1.2.3 Health

Currently, there are many medical imaging technologies, like X-ray, CT, MRI, used in detecting diseases. However, these kinds of data need interpretation by professionals and are not intuitive. Building 3D models from this kind of data is beneficial to the patient in understanding the diseases and the doctors to conduct operations more intuitively

[6] https://www.pwc.com.au/publications/retail-consumer-market-insights/quarter-3-2017.html

[35, 100]. The most recent work of [65] reconstructs the complete 3D volumes from a single 2D x-ray imagery. As for the human body, some medical research suggests that many diseases are relevant to people's weight and body shape changes. For instance, obesity could be an indication of a number of serious and potentially life-threatening conditions such as heart disease or cancer. Obesity can also affect the quality of life and lead to psychological problems, such as depression and low self-esteem. Estimation of human body shape helps to perform accurate anthropometric measurements, like weight, height, waist circumference which are useful indicators of health and provide references for personal exercise and suggestions from doctors.

1.2.4 Ergonomics

Digital human models are becoming more commonly used by ergonomics and other engineers to design both equipment and work environments to meet the needs of human operators [122, 39, 112]. They allow the designers to explore the potential advantages and disadvantages of different design configurations without requiring the construction of expensive physical mockups as for the past. For ergonomics applications, high-fidelity personalised human body models with accurate body size and shape can be employed. As shown in Figure 1.4, human avatars can substantially benefit human-related design applications, like vehicle design [13], chair and seat design [123] et al. Traditional ergonomic analysis is performed based on the real experiment users' operation on a certain equipment. If a high-fidelity human body model can be estimated for real users with concerned body sized and shapes, the cost for ergonomic engineering simulation can be greatly reduced.

1.3 Challenges and Solutions

The challenges of acquiring high-fidelity human body model based on user-generated data stem from both the data types as well as human body states. The main challenges

Figure 1.4: Human avatars used for task analysis [133].

are summarized as follows, along with the proposed solutions in this research.

Point Clouds The majority of traditional registration methods are tailored for high-quality scans from expensive scanners [95, 198, 15]. Following the introduction of the low-quality scans from cost-effective devices, the 3D data capturing of the human body becomes more convenient and easier with many reconstruction strategies appearing for static or dynamic scenes in offline or realtime settings [109, 110, 193]. However, due to the inevitable holes, noise and outliers in the low-quality scan, the registration of the human body becomes even more challenging [175, 131]. How do we represent the registration problem? The rigid ICP [12] is extended in the non-rigid setting. The traditional rigid ICP iteratively revises the transformation needed to minimize the distance between the source and the target point clouds. It optimizes a global optimal transformation matrix (combination of rotation and translation) for all the points of the source. The shape of the source point clouds will not be changed. However, in the case of the human body, the different body shapes and poses require non-rigid deformation of the source point clouds. The transformation for each vertex in the source point clouds should be computed. The whole registration problem is formulated as energy minimization below.

$$E(C, W) = E_{fit}(C, W) + \lambda E_{reg}(C, W), \tag{1.1}$$

where $E_{fit}(C, W)$ is the distance between pairs of correspondences, $E_{reg}(C, W)$ is the prior term which is learned from prior knowledge and enables higher robustness, C is the correspondences that are iteratively updated and W is the transformation matrix. How do we robustly register the noisy, incomplete and sometimes even wrong point clouds? When we scan around the human body, there exist some parts that the scanners cannot see, such as the armpit, the part between two legs and the head top. These occluded parts that the scanners cannot see will cause incomplete parts on the surface. Also, noise and distortion commonly exist on the scanned surface. Features are covered or lost in these cases [4]. Therefore, in this research, the template-based registration method is investigated. Principal component analysis (PCA) [121] is exploited to train prior knowledge and combined with non-rigid ICP to regularize the deformation of the template. Specifically, in Chapter 4, two levels of statistical shape models are presented: a holistic body shape model to describe the basic figure of human and a set of shape models for every body part to capture more body details. The fitting procedure follows a coarse-to-fine approach that is robust and efficient. How do we register target point clouds of different poses? Human body models are in great variance in poses. When the initial poses of template model differ a lot from the target mesh, it means the ICP cannot have a good initialization to start. Due to the local iterative procedure the ICP adopts, the registration may easily get trapped into local minima [188]. Therefore, to address this problem, both prior knowledge and a set of sparse landmarks are used. The landmarks will drive the deformation of the template to make it pose in a similar way as the targets, providing a better initialization for registration. The prior knowledge will improve the robustness of the registration.

Images To estimate 3D human body mesh from 2D images, rather than reconstructing human body models from dozens of images taken from multiple views [104, 105] or videos [205, 192], only one image in the front view is used to estimate the body shape. Single image based reconstruction is an under-constrained problem as single images do not provide enough information for inferring the depth information for each pixel. The human bodies with different poses could have the same projection on the 2D plane. How do we infer the human body shape from a single image? We assume that 2D joints and boundary contain a great amount of 3D shape information. Fitting only 2D joints and boundary can produce plausible estimates of 3D body shape. The template-based method is also a part of the solution [14, 107, 81]. An objective function is defined and, pose and shape are optimized directly so that the projected joints of the template model are close to the 2D joints of images. How do we deform the general template model to targets with different poses? There are a large number of varieties in human poses. The human pose is represented as a set of joint positions at a certain state. The movement is regarded as the transformation from one pose to another pose. The transformation is performed by changing the positions of each joint through the kinematic chain [54]. Thus, the template is reposed with the joint information to make it pose the same way as the target does. Specifically, in Chapter 5, joint and boundary of 2D image provide clues for better estimation of human body shapes. New pose prior are trained for better estimates of human poses.

1.4 Contributions

The contributions of this research are about acquiring high-fidelity models from two user-generated observations, point clouds and 2D images. To achieve this goal, for user-scanned point clouds, the non-rigid registration approach with a trained statistical shape model is applied to align a high-quality template to targets. For single user-captured images, 2D joints and boundaries are exploited to estimate the 3D body models.

- In an ideal situation where closed meshes are acquired, the shape correspondence establishment is discussed. Traditional approaches are usually based on the intrinsic distortion, like the discrepancies in geometric distance to locate the closest shape features but the symmetric flips cannot be distinguished by intrinsic shape characteristics. Therefore, in this thesis, an approach of locating accurate shape correspondences is proposed. Given that the acquirement of skeletons of a human body is much easier due to the ubiquity of the use of 3D depth sensors, the skeleton is exploited to distinguish the left and right to remove the symmetric flips.

- In most cases of user-generated scans, the captured meshes are usually not watertight ones where the acquired meshes are ideally closed and there are no holes on the surface. The watertight meshes are usually synthesized by artists using professional tools. In practice, even big holes are prevalent on the surface, which leads to the intrinsic feature based approaches are not feasible to detect reliable shape correspondences. Thus, for this kind of data, finding accurate shape correspondences based on features is impracticable. In this thesis, non-rigid registration approaches are investigated to locate the accurate closest points on the target iteratively. During this process, non-rigid deformation of the template mesh is performed to achieve the closest results. For target human body scans with a standard A pose as shown in Figure 1.5, a fully automatic registration method which combines statistical PCA model extracted from aligned high-quality human body meshes with non-rigid ICP. For human body scans with various poses, landmark based human body registration is proposed. Sparse landmarks are used to guide to pose template.

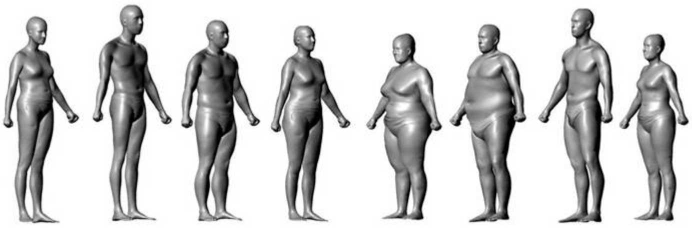


Figure 1.5: The examples of the standard A pose [189].

- Besides 3D scanning data, another important and more commonly available type of user-generated data is 2D images. The estimation of the 3D human body from a single 2D image is an ill-posed problem as single 2D images lack sufficient information to infer the depth. Numerous 3D subjects can have the same projection on the 2D plane. To tackle this problem, the statistical human body model which is trained from thousands of 3D human body models with various poses is used as the template model and 2D joints and boundary information are utilised to accurately estimate the realistic and personalised human avatars.

- In order to evaluate the proposed method for registering the user-generated 3D scans, we collect a 3D human body dataset which can not be found in the literature. It is scanned with commodity Microsoft Kinect for XBOX 360, called K3D-hub Dataset. The dataset has 255 real human body scans and can be used to evaluate the robustness of registration algorithms in case of low quality scans.

1.5 Thesis Structure

The remaining chapters of this thesis are structured as follows.

Chapter 2 Background presents a background on high-fidelity human body modelling technologies, with particular emphasis on state-of-the-art registration and modelling methods using point clouds and typical 3D reconstruction methods from images.

Chapter 3 Symmetry-aware human shape correspondence under closed mesh illustrates the proposed method to address the problem of symmetrically flipped correspondences in human body mesh correspondence establishment via a novel refinement algorithm based on skeleton and base vertex set.

Chapter 4 Building high-fidelity human body from user-scanned point clouds describes the proposed robust registration methods that integrate the prior knowledge of the human body deformation into classic non-rigid ICP and can be applied to both high-quality and low-quality human body scans in various poses.

Chapter 5 Estimation of 3D human body models from user-captured 2D images presents the proposed single image-based human body shape and pose estimation method. This method improves the accuracy of shape estimation by using both joints and the boundary information. The pose estimation is also improved by training a stable pose prior for the commonly encountered scenes in our daily life. In the end, the potential application of virtual dressing is presented.

Chapter 6 Conclusions and future work gives the general conclusion and comments on the limitation of this research. Some future works for improving and expanding the proposed approaches are also discussed.

Chapter 2

Background

In this chapter, technologies related to human body modelling are reviewed, with special focus on those relying on two-types of data: low-quality 3D scan data and images. Due to the appearance of low-cost RGBD cameras, 3D data can be captured much more easily than before, while many imperfections on the captured data bring more challenges to the process afterwards. To build a complete human body model from 3D scanning data, a number of human body reconstruction systems register point clouds/ meshes captured from different camera views. Thus, to be compatible with this process, in this chapter, a typical type of low-cost 3D scanner - Kinect is firstly introduced, and then the background of surface registration is presented which includes the rigid and the non-rigid approaches, followed by an introduction of current human body parametric models. Based on these foundations, a general review of the state-of-the-art approaches to human body modelling with low-cost 3D scanners is presented.

Another important kind of source data - 2D images, plays an indispensable role in 3D reconstruction. Therefore, the image based reconstruction methods are also reviewed in this chapter.

2.1 Introduction to Low-cost 3D Scanners

Emerging at the end of 2000s, light coding technology paved the way for new technology allowing low-cost sensory inputs for consumer and commercial markets. With years going by, this tool has gradually evolved, allowing general users to easily find affordable and friendly 3D scanners to use. One of the typical examples is Microsoft Kinect.

Microsoft released Kinect for Xbox 360 device in November of 2010, as the first commercial release of a depth camera device. Then in February of 2012, Microsoft released the version of Kinect for windows. Kinect sensors consist of a red, green, blue (RGB) camera, an infrared (IR) emitter and an IR camera. As shown in Figure 2.1, the RGB camera is in the middle. The IR emitter is on the left and the IR camera is on the right. A multi-array microphone is used for audio localization and recognition. The motorized tilt can be used to adjust the angle of Kinect.

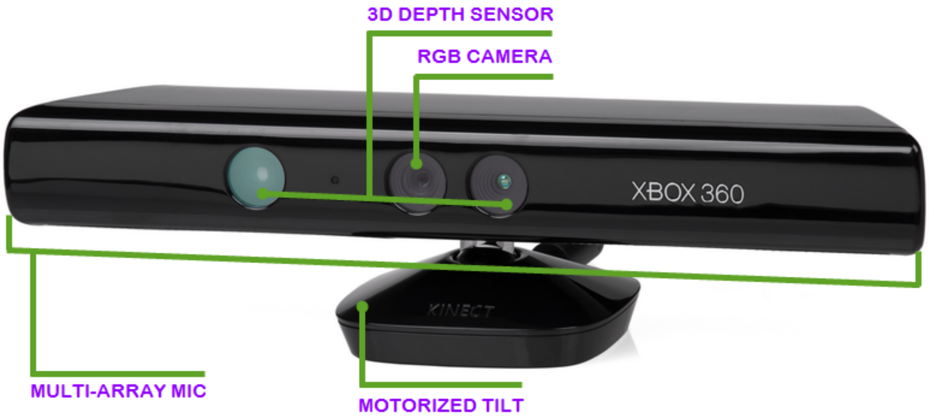


Figure 2.1: The hardware structure of Kinect.

An overview of the Kinect's work flow is shown in Figure 2.2. When a Kinect starts to detect the object with its field of view (FOV), the IR emitter will send a known pattern of IR dots and they will be recorded by IR cameras. When an object appears in FOV, the new light pattern will be recorded by the IR camera. The appearance of the new object causes distortion of the light pattern and it will be compared with the pre-recorded IR pattern. The measured distortion of the recorded IR pattern will be used to calculate the depth distance of the point from the camera to the objects.

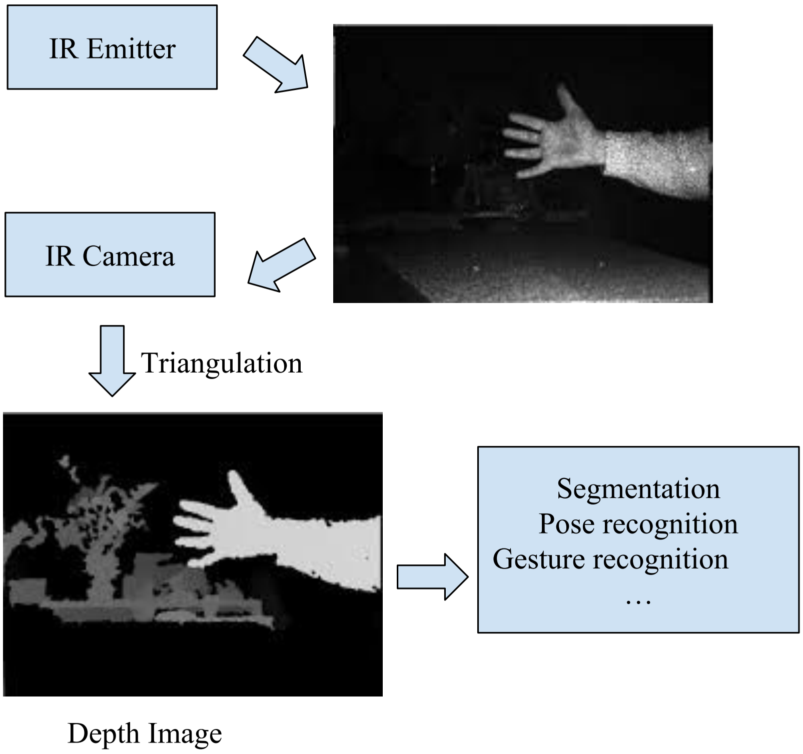


Figure 2.2: The overview of how Kinect works.

Although structured light technique provides fast and cheap measurements of the depth distance, there are some situations in which depth measurements cannot be acquired. First, the sun light which includes infrared will wash out some speckle patterns so that the IR sensor cannot capture all the light patterns. This causes loss of some depth data, thus the Kinect is usually limited to indoors. Second, when several Kinects are setup to simultaneously capture an objects, the Kinects will confuse each other. Third, when an object is occluded by another, the IR emitter cannot send speckle patterns to the occluded parts and therefore the IR camera cannot receive light patterns of the occluded parts. Fourth, different objects' materials will also affect the capturing of light patterns. If the material is light-absorbing, this means the emitted light can be absorbed by the material. As a result, the IR camera cannot receive the projected light and calculate the depth information. If the material is a smooth plane, the IR projected speckle beam may produce specular reflection on the surface of the object, which also causes the failure in capturing the light pattern.

Similar to Kinect, several other structured light sensors have appeared, such as Structure sensor, XtionPro [1], Intel RealSense [2] et al., but they also suffer from the same problems of low-quality scanning results mentioned above. Stimulated by the emergence of commodity 3D scanners, technologies focusing on using their data for 3D human reconstruction [180, 168, 119], modelling [63] and human detection and tracking [185, 186, 76] have been developed in the research community. In this research, the proposed methods mainly address the problems of high-fidelity human body modelling from low-quality data captured by Microsoft Kinect for XBOX 360.

2.2 Current Human Body Dataset

With the development of the 3D scanning technology, a number of 3D human body datasets have been captured. Here, several commonly-used datasets are described below.

SCAPE human dataset [6] was built by Dragomir et al. in 2005. It is composed of a pose dataset and a shape dataset. The pose dataset contains scans of 70 different poses of a particular person. The shape models consist of 45 different people in a similar but un-identical pose. In the dataset, one mesh is chosen as the template mesh and registered to other instance meshes. Each mesh has 25000 triangle faces and 12500 vertices. Although the original work makes use of both shape and pose data, only the pose data is distributed together with its skeleton information. Meshes in SCAPE are hole-filled using the algorithm by Davis et al. [38]. SCAPE model also constructs a skeleton for the template mesh based on the fact that vertices on the same skeleton joint are spatially contiguous and exhibit similar motion across different scans. Thus, after

[1] https://www.asus.com/3D-Sensor/Xtion_PRO/

[2] https://realsense.intel.com/

scanning the pose instance for a particular person, the mesh is decomposed into several approximately rigid parts and obtains the location of the parts in different pose instances as well as the articulated object skeleton linking the parts. Based on the pose dataset, a tree-structured articulated skeleton is automatically constructed with 16 parts.

TOSCA [22] consists of 80 objects, including 12 females in 7 poses and 2 males in 20 poses, as well as 11 cats, 9 dogs, 3 wolves, 8 horses, 6 centaurs, 4 gorillas. This dataset is synthesized by artists. Meshes within the same category are in the same correspondence. All meshes are watertight and some parts do not present realistic deformation of muscles and soft issues.

FAUST (Fine Alignment Using Scan Texture) [18] contains 300 real human scans of 10 different subjects in 30 different poses, acquired with a high accuracy 3D multi-stereo systems. Each scan is a high-resolution, triangulated, non-watertight mesh. Compared to synthetic meshes, FAUST presents more challenging features: missing data, different topologies, realistic deformation and self contacts. All meshes are brought into alignment with a common template mesh.

MPII Human Shape [127] is a collection of 3D human body shape meshes. It is developed by aligning meshes in the CAESAR dataset [3] and learns the human shape space from thousands of alignments. It is the largest available dataset of aligned human meshes to date but it only covers the shape space in the canonical pose.

2.3 Surface Registration

Registration is a long-term problem in the research community of computer graphics, computer vision and even medical research [196, 47, 140]. Both the large variance of human body poses and the deformation of human shapes pose great challenges to the registration of human body. The problem of human body surface registration is defined as aligning two or more human body meshes as closely as possible. Due to the limitations

[3] https://store.sae.org/caesar/

of scanning technology, most scanning systems provide partial surface data that must be aligned and merged to obtain a complete digital representation of the scanned object [199, 89]. Besides 3D shape acquisition, registering a template to a set of target scans makes it feasible to transfer texture and skeleton, and perform statistical shape analysis and shape interpolation, which play important roles in computer graphics and computer vision [93, 167, 71]. Surface registration can be classified into dense registration and sparse registration. The dense registration is to find a mapping from each point in the template to the target, while the sparse registration is to find correspondences only for selected feature points. According to the absence or presence of shape changing, the surface registration also can be divided into rigid registration and non-rigid registration.

2.3.1 Rigid Registration

In rigid registration, only rotation and translation are performed for the mesh. The resultant mesh shares a common coordinate system and overlaps the most of parts with the target. Its own shape, however, is invariant. Rigid transformation is usually a preprocessing step before nonrigid registration. Sparse correspondences should be given before rigid registration. However, the limitations of the 3D scan data pose a lot of challenges for the accurate correspondence establishment. On the one hand, the noise, outliers and missing data could cover or lose the features of the key points; on the other hand, the intrinsic symmetry of the human body makes it challenging to distinguish the right and the left so that the current algorithms usually suffer from the symmetry flip problem [191, 83]. The flipped correspondences will cause erroneous rigid registration results.

The main goal of rigid registration is to compute a rigid transformation between two meshes, making them share a common coordinate system. At least four pairs of correspondences are sufficient to compute rigid registration. For human body modelling, it mainly undergoes articulated changes where bones undergo large rigid transformation and local nonrigid surface deformation (bending or stretching) near joints [163]. If rigid transformation is to be performed, the joint positions are good correspondences. Many works use skeleton to perform rigid transformation. The method proposed in [2] predefines a skeleton for the mesh. Since at least four joints are sufficient for rigid registration, more recent works do not use explicit skeletons. In [40], five extrema that represent head and limbs are located to perform rigid transformation. The approach described in [202] constructs a consensus skeleton from successive frames of point clouds. In [28], it is able to build a graph-based representation for the scans and estimate the joints automatically. With given correspondences, the rotation and translation for two meshes A and B can be computed with singular value decomposition. Without explicit correspondences, Iterative Closest Point (ICP) is an option. Here, these two kinds of typical methods for rigid registration are introduced.

Iterative Closest Point

Iterative Closest point (ICP) is an algorithm employed to minimize the distance between two meshes. ICP is often used to reconstruct 2D or 3D surfaces by aligning consecutive pieces of scans captured from different views [48, 53]. The algorithm iteratively updates the transformation (a combination of rotation and translation) aiming at minimizing the distance from the source to the reference point cloud. If two corresponding point clouds are given:

$$\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}, \tag{2.1}$$

$$\mathbf{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_n\}. \tag{2.2}$$

Usually, the corresponding center of mass μ is subtracted from every point in each point sets before calculating the transformation, which can be regard as moving the two point sets to the origin of the coordinate. The resulting point sets are:

$$\mathbf{X}' = \{\mathbf{x}_i - \mu_{\mathbf{x}}\} = \{\mathbf{x}'_i\}. \tag{2.3}$$

$$\mathbf{P}' = \{\mathbf{p}_i - \mu_{\mathbf{p}}\} = \{\mathbf{p}'_i\}. \tag{2.4}$$

In the ICP algorithm, the goal is to find the translation $\mathbf{t}$ and rotation $\mathbf{R}$, such that the sum of the squared error is minimized [7]:

$$E\left(\mathbf{R},\mathbf{t}\right) = \frac{1}{N}\sum_{i=1}^{N} \parallel \mathbf{x}_i' - \left(\mathbf{R}\mathbf{p}_i' + \mathbf{t}\right) \parallel^2, \tag{2.5}$$

where N is number of the points, $\mathbf{x}_i$ and $\mathbf{p}_i$ are corresponding points of the two point clouds. At each step, the parameter set of $\mathbf{R}$ and $\mathbf{t}$ is updated. In most cases, the point correspondence pairs are unknown. Therefore, we usually calculate rotation and translation in an iterative way as shown in Figure 2.3.1.

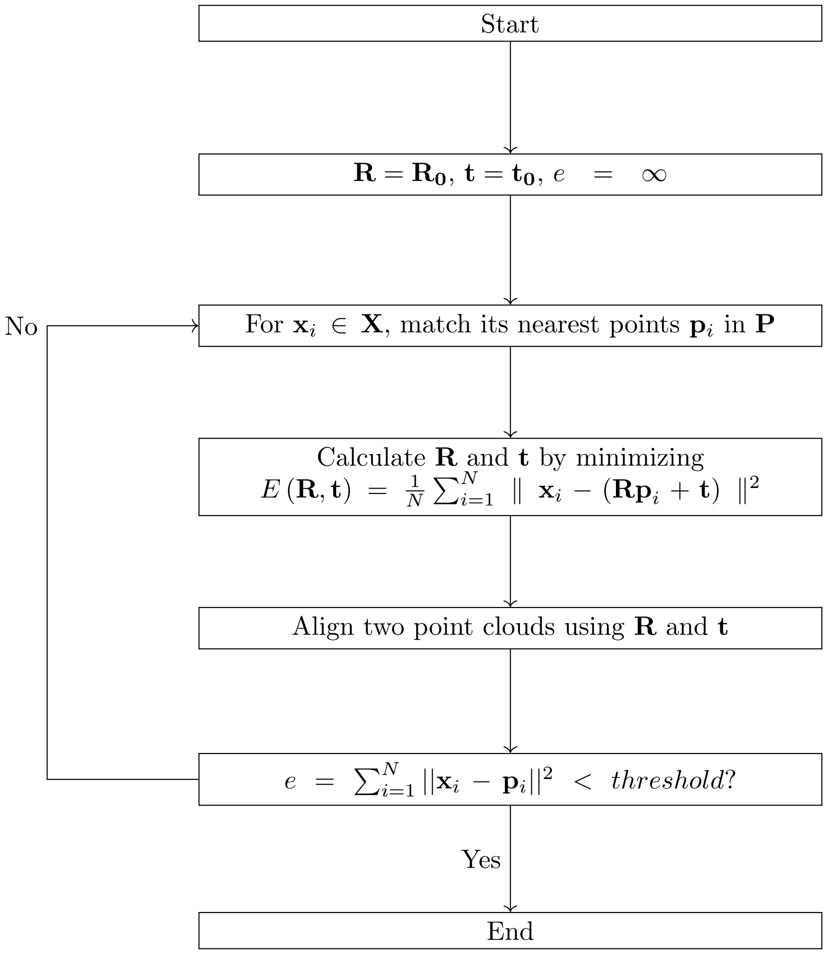

Start
$\mathbf{R} = \mathbf{R_0}$, $\mathbf{t} = \mathbf{t_0}$, $e = \infty$
No
For $\mathbf{x}_i \in \mathbf{X}$, match its nearest points $\mathbf{p}_i$ in $\mathbf{P}$
Calculate $\mathbf{R}$ and $\mathbf{t}$ by minimizing
$E\left(\mathbf{R},\mathbf{t}\right) = \frac{1}{N}\sum_{i=1}^{N} \parallel \mathbf{x}_i - \left(\mathbf{R}\mathbf{p}_i + \mathbf{t}\right) \parallel^2$
Align two point clouds using $\mathbf{R}$ and $\mathbf{t}$
$e = \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{p}_i||^2 < threshold?$
Yes
End

Figure 2.3: The flowchart of ICP algorithm.

The ICP algorithm is widely used for geometric alignment of three-dimensional models and its many variants are proposed. However, it lacks robustness when ICP algorithm is used in minimizing a non-convex cost function bacause of the local minima. Thus, many variants are proposed to improve the robustness of ICP. In the stage of selecting initial points, the original version of ICP uses all available points [12]. The proposed method in [170] uniformly subsamples available points. Random sampling is also used to select points. The method in [101] selects different sample of points at each iteration. In some cases, meaningful points with special properties will speed up the converge of ICP algorithm. In [177], points are selected with high intensity gradients to aid the alignment. The normal or face information are also considered to select effective points. The strategy to select points is finding correct and meaningful correspondences. A typical solution for finding correspondence manually is to use markers placed on subjects when capturing. An alternative solution for accelerating the marker and correspondence selection procedure is to automate the detection of landmarks on the mesh.

Singular Value Decomposition In [7], given correspondences, Singular Value Decomposition (SVD) [151] is used to compute the rotation and translation matrix. As shown in Figure 2.4, given the known corresponding vertex sets $\mathbf{P_A} \in \Re^{3\times N}$ and $\mathbf{P_B} \in \Re^{3\times N}$ for point clouds A and B with N points, the alignment from A to B is illustrated as follows. Firstly, we compute the centroids of two meshes as below.

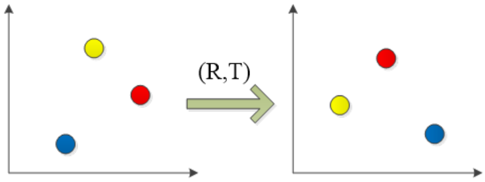


Figure 2.4: The rigid transformation between two datasets. The corresponding points have the same color. **R** is the rotation and **T** is the translation.

$$\bar{\mathbf{A}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{P_A}^i, \tag{2.6}$$

$$\bar{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{P_B}^i, \tag{2.7}$$

where $\bar{\mathbf{A}} \in \mathfrak{R}^{3\times 1}$ and $\bar{\mathbf{B}} \in \mathfrak{R}^{3\times 1}$ are the centers of two point clouds. With the centroids, we then compute the covariance matrix :

$$\mathbf{Cov} = \sum_{i=1}^{N}[(\mathbf{P_A}^i - \bar{\mathbf{A}})(\mathbf{P_B}^i - \bar{\mathbf{B}})^T], \tag{2.8}$$

Then, the directions U_A and U_B of the main axes can be computed by singular value decomposition:

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = SVD(\mathbf{Cov}). \tag{2.9}$$

As illustrated in [7], we have

$$\mathbf{R} = \mathbf{V}\mathbf{U}^T. \tag{2.10}$$

If the determinant of $\mathbf{R}$, $det(\mathbf{R}) = +1$, then $\mathbf{R}$ is the desired rotation solution; if $det(\mathbf{R}) = -1$, then $\mathbf{R}$ is the relection where SVD based algorithm fails. This occurs if the two point clouds are coplanar or colinear. Thus, [171] proposes Eq. 2.11 that always gives the correct transformation parameters even when the data is corrupted.

$$\mathbf{R} = \mathbf{V}\mathbf{S}\mathbf{U}^T, \tag{2.11}$$

where $\mathbf{S}$ is chosen as

$$\mathbf{S} = \begin{cases} \mathbf{I} & if \quad det(\mathbf{V}\mathbf{U}^T) = 1 \\ diag(1, 1, ..., 1, -1) & if \quad det(\mathbf{V}\mathbf{U}^T) = -1 \end{cases} \tag{2.12}$$

As for the translation $\mathbf{t} \in \mathfrak{R}^{3\times 1}$, it is formulated as:

$$\mathbf{t} = -\mathbf{R}\bar{\mathbf{A}} + \bar{\mathbf{B}}. \tag{2.13}$$

According to [98], compared with other solutions proposed in [173, 70, 69] , SVD methods provide the best overall accuracy and stability taken the degenerated data into account.

2.3.2 Non-rigid Registration

Non-rigid registration is a more challenging task, as it needs to consider shape deformation. In the case of human body, each person have unlimited poses and the shape of each person is different from the other. Nonrigid registration is necessary if we want to fully align two human body meshes. The concept of “full alignment” is that we deform one template mesh to the target mesh so that the appearance (shape and pose) of the deformed template is totally the same with the appearance of the target mesh. In order to perform a good nonrigid deformation, dense reliable correspondences are needed. From the analysis of shape correspondence before, establishing dense correspondences for 3D shape is a challenging problem. Missing data poses much more challenges for non-rigid registration than rigid registration. Here, we mainly focus on the state-of-the-art methods nonrigid registration in the following.

In order to align two meshes, namely the template and the target, as closely as possible, the template mesh undergoes a series of transformations while the target remains unchanged. The registration is called non-rigid registration if the transformations include changes in shape such as morphing, or articulation [150]. There are many challenges in performing non-rigid registration. Noise, outliers, limited overlap and missing data make the nonrigid alignment error-prone. Missing data results in holes in the scans. In these hole parts, the nearest points of template will be found on the boundary points of the hole. This will lead the nearby faces to converge, decreasing the quality of the deformed mesh. If the hole is too large, the deformed mesh will lose the original shape. Besides the challenges coming from the imperfections in data, deformation itself faces a lot of problems. Unlike rigid registration in which a few correspondences are sufficient to define a rigid transformation, nonrigid registration needs reliable dense correspondences to compute. However, establishing accurate dense correspondences is nontrivial. Although many works [144, 143, 75, 146] are able to find the dense correspondences accurately with spectral methods, they are not feasible in real scans with noise, outliers and missing data.

Compared to the methods like multidimensional scaling (MDS) in [45], Generalized Multidimensional Scaling (GMDS) in [21], or Laplace-Beltrami operator used in [115] where the mesh is embedded into a different domain where Euclidean distances can approximate geodesic distances so that the isometry errors can efficiently be measured and optimized in the embedding space, another type of vertex based methods are proposed to approach the problem of human mesh registration [67, 5, 32, 16]. In these works, the non-rigid registration is regarded as an optimization problem with the objective function in Eq. 2.14.

$$E = E_{distance} + E_{regularization}, \tag{2.14}$$

where the distance term $E_{distance}$ aims to minimize the distance between the template and target and the regularization term $E_{regularization}$ is a smoothness term which controls the deformation of neighboring vertices to make the deformation smooth. In [1], Laplacian mesh regularizer [158] is adopted to enforce the smoothness of the deformation. In [67], besides the spatial smoothness of the deformation, an additional constraints term that controls the magnitude of the effect of the pose-dependent deformation model is added. Some other regularization constraints are also used to improve the registration performance. Texture information in [15] is formulated as an additional term to keep more details in deformation. Markers are another commonly used constrains such as used in SCAPE [6] and Mosh data [97]. However, adding markers requires professional knowledge and is hard to be satisfied in routine use.

Following the framework defined in Eq. 2.14, many efforts are made to achieve the non-rigid iterative closest point [5, 32] which computes the affine transformation (translation, rotation, scale and shear) at each vertex of template to allow non-rigid registration of template and scans.

2.4 3D Human Body Models

After the above introduction of the surface registration, it is known that a template is adopted to achieve alignment. Here, we give a survey of human body models that can be used as templates and introduce their use in the human body related research like pose estimation, segmentation, registration and animation. Human body modelling has a long history. The early models were stick-like figure and were used for coarse pose estimation from images. Later models using geometric primitives help coarse human body shape segmentation. Current human body models are built from real people scans with more realism and detailed shape variation, and are able to infer more details. Statistical models of 3D human body shape and pose learned from scan datasets have been developed into valuable tools to solve a variety of computer vision and graphics problems, such as the generation of realistic synthetic body models or the reconstruction and tracking of detailed 3D body models from input images or point clouds.

2.4.1 Models using Geometric Primitives

In early days, some human body models are composed of geometric primitives. An example is the puppet model produced by Hinton [66] which uses 15 rectangles to represent the human body parts and these 15 rectangle representations have the following relationships: the length of a part, measured along the proximal-distal axis must be greater than its width; the truck must be wider than any of the upper limb-parts, and each of these, in turn, must be wider than its connected lower limb-part. Ju et al. [80] define a "cardboard person model" in which a person's limbs are represented by a set of connected planar patches, as shown in Figure 2.5. The model is used to estimate the motion of a person and applied in tracking and activity recognition.

Besides 2D rectangles, in Figure 2.6, Hanavan et al. [59] propose a mathematical human body models from a set of simple polygonal shapes. In [154], authors propose a loose-limbed model where parts are loosely connected and each body part is modeled by

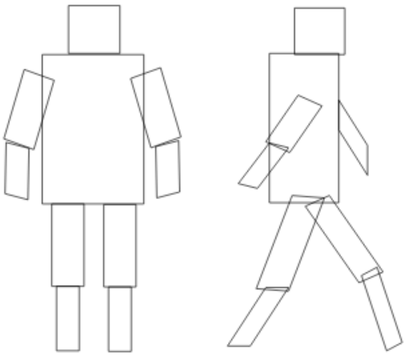


Figure 2.5: The cardboard human model. The limbs of a person are represented by planar patches [80].

a tapered cylinder with an elliptical cross-section. Each part model is associated with 6 fixed and 6 estimation parameters.

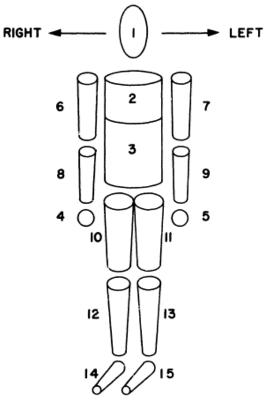


Figure 2.6: The mathematical model [66].

As shown in Figure 2.7, the fixed parameters correspond to the part length, width at the proximal and distal ends, the offset of the proximal and distal joints along the axis of the limb. This loose-limbed model can be used for tracking and the size of body parts is manually defined. The connections between two parts are modeled with Gaussian potentials, then exploited for a non-parametric form of belief propagation.

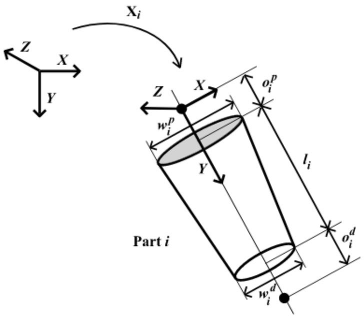


Figure 2.7: Parametrization of a 3D body part [154].

Sminchisescu et al. [157] present a human body model consists of a "skeleton" and "flesh" built from super quadric ellipsoids. This human body model is targeted to recover three-dimensional human body motion from video sequences as shown in Figure 2.8. Stoll et al. [160] model the human body by sums of spatial Gaussians to allow for performing motion capture from multi-view video sequences. The model comprises a kinematic skeleton and a statistical model that represents the shape and appearance of the human. The body is represented in 3D with a set of spheres of various sizes as shown in Figure 2.9.

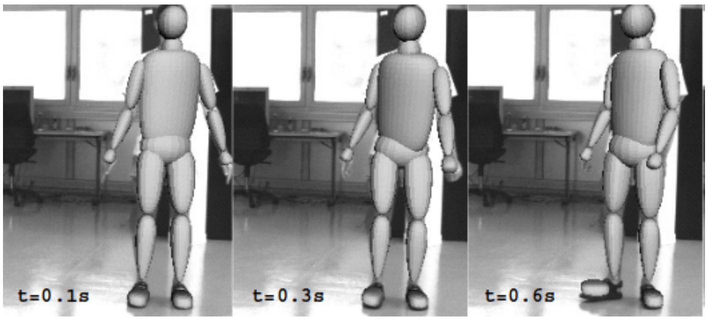


Figure 2.8: Human tracking with human body model [157].

As the geometric primitive representations of human body is simple, they usually have the advantage of efficient computation. They tend to be used to assist human related processing in the 2D image domain, like motion tracking, recognition, re-identification et al. Geometry primitive based representation only can describe the human body shape

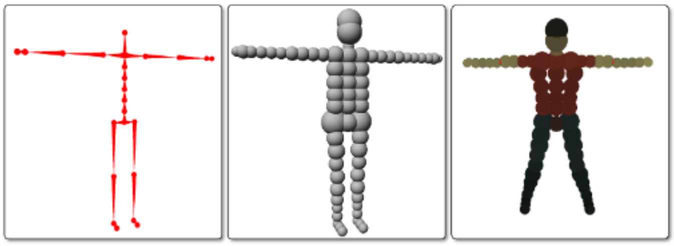


Figure 2.9: SoG-based body model. From left to right: skeleton, default SoG model, actor-specific model [160].

and pose roughly. However, in the computer graphic community, visual quality is mandatory. Creating visually plausible soft deformation by merely using geometric primitive representations can hardly be achieved.

2.4.2 Artist-driven Skinned Models

In animation industry like Disney, Lucasfilm or Dreamworks, many artists create animation characters with powerful professional tools. Such animation requires a lot of human inputs and is not very straightforward. In the process of creating faithful representations of the human body, a vexing problem is modelling body joints: while a kinematic chain can be used to drive surfaces, surfaces near two or more bones can be difficult to model. Building realistic models requires a lot of manual effort from animators for rigging and skinning characters.

2.4.3 Simulation-based Models

In simulation based models, the skin deformation is caused by the underlying motion of the skeleton and the information of the muscles [181, 8]. These methods provide very high-level realism but they are difficult to build and animations with these methods usually require expensive simulations.

2.4.4 3D Statistical Models

Although there are various scanning systems for capturing bodies in 3D, the acquired data can hardly be readily used as a human body model. The captured human body scans usually contain holes and they are in different topologies. Removing the noise and outliers, filling holes and alignment have to be performed before being used in practical applications. If animation is expected to perform, rigging and skinning for each scan are also indispensable. Rigging is to create a skeleton and a set of controls to animate the characters and skinning is the process which assigns bone influence to the mesh in order to generate visually plausible deformations with pose changes. The storage of these scans comes in the form of sets of vertices and faces, which is also inefficient.

However, a statistical human model is able to represent variations in human physiques and poses using low-dimensional parameter spaces. Methods for learning shape models depend on identified corresponding points across many 3D scans, namely, registered data. In a word, building parametric human body model is to learn a parameter space from a series of registered 3d meshes. This model can represent human bodies with different identities and poses by adjusting the shape and pose parameters. The pipeline of building the statistical model is shown in the Figure 2.10. A general template model is registered to different target scans to acquire corresponding alignments. With the collection of alignments, the statical model M is trained with which new shapes and poses can be generated by adjusting the shape and pose parameters.

In the early stage, human body modelling is performed by simulating tissue deformation on top of modelled skeletal bones [42, 148]. This kind of approaches has been researched extensively but involves a lot of manual modelling, since the surface as well as the bones, muscles, and other tissues have to be designed. Additionally, these methods tend to be computationally expensive since they involve physically based tissue simulation [61].

With easier acquisition of 3D scans becoming possible , many systems are proposed

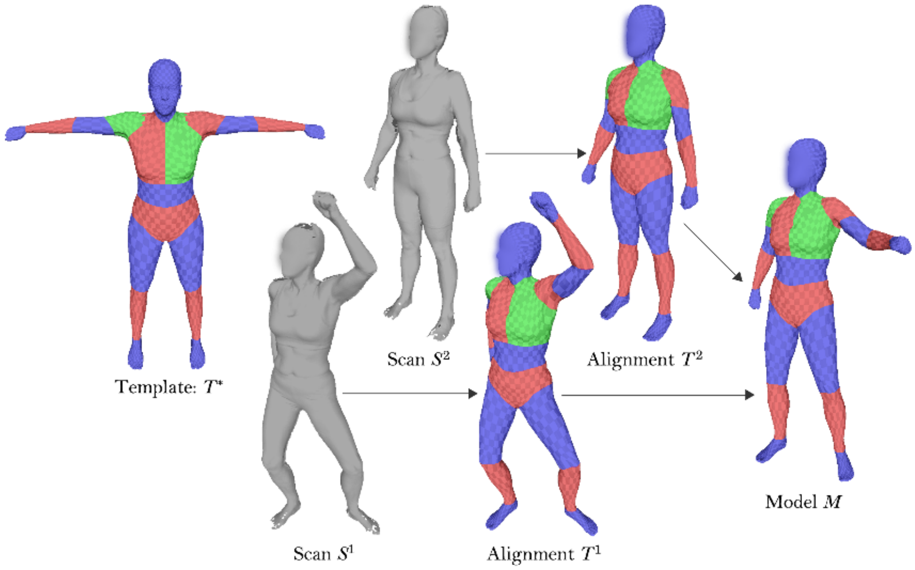


Figure 2.10: The pipeline of building the statistical model [67].

to build human body model from 3D scans [6, 67, 61, 2, 125, 15, 31]. The earliest 3D statistical shape model [2] learns a space of the human body from the registrations of 250 laser scans in a fixed A pose. The statistical shape model is trained by performing Principal Component Analysis (PCA) over shape deformation with respect to a template. New human body shapes can be synthesized by adjusting the coefficients of learned shape components. Allen et al. [3] later improve this work by considering the pose deformation.

SCAPE model is one of the most popular parametric human body models. It is trained from 70 registered human body poses and 50 registered human body shapes.

It learns two separate models of human pose and human shape, and combines them to produce 3D surface models for different people in different poses. It learns a pose deformation model from a subject with multiple poses and learns a shape model from different identities in a standard "A" pose. The deformation of the SCAPE model is as follows. Firstly, the template is decomposed into discrete triangles. The deformation is performed on triangle edges. Secondly, the single triangle is individually deformed according to a sequence of pose- and shape- dependent linear transformations. There are

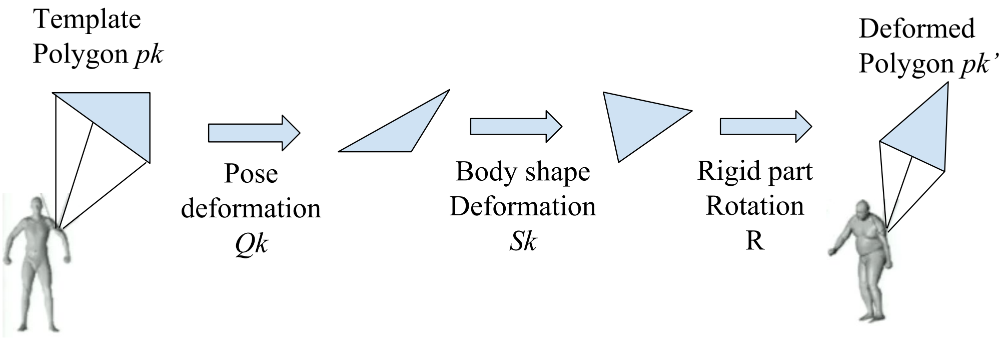


Figure 2.11: SCAPE models the deformation of of each triangle face.

three sequential deformation matrices that are applied to each triangle T_f. $R_f(\theta)$ is the rotation that controls the pose-dependent deformation; D_f controls the shape deformation between different persons; and $Q_f(\theta)$ accounts for pose-induced shape changes like muscle bulging and skin wrinkling. Therefore, the deformation of each triangle can be written as:

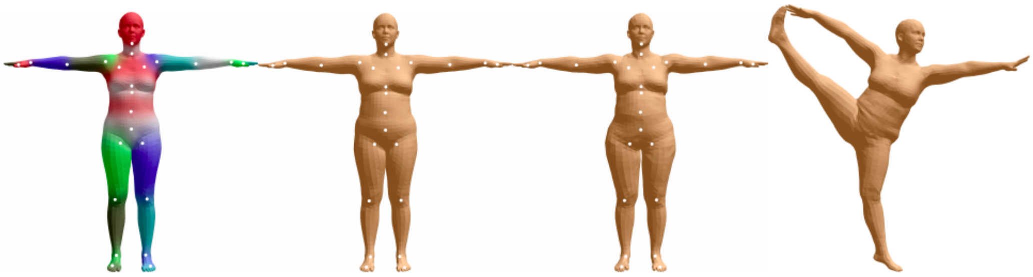
$$f^* = R_f(\theta)D_fQ_f(\theta)f, \tag{2.15}$$

where f is the edge of each triangle. After each triangle is re-targeted to new positions, these disconnected triangles are stitched together to build a new mesh. SCAPE model transfers the nonrigid deformation of the human body into rigid deformations of each triangle. Although this transformation simplifies the problems, the final stitched body may fold or collapse near joints as the triangles are rotated independently.

In order to address this problem, BlendSCAPE proposed in [67] rotates each triangle with a linear blend of the rotation of each part in the skeleton tree. The rotation of each triangle is defined as $B_f(\theta) = \sum_i w_{fi}R^i$. R^i is the rotation of each part. With $B_f(\theta)$, BlendSCAPE is able to smooth the deformation of the boundaries. Therefore, BlendSCAPE approximates a scan in the dataset with a model M that poses the model using $B_f(\theta)$, deforms it to the shape of the person using a person-specific parameter D_f and describes the non-rigid shape changes using $Q_f(\theta)$. BlendSCAPE aligns the triangulated template mesh to a corpus of 3D scans and simultaneously trains the parameters of the model. The problem is formulated as the optimisation of a non-linear objective function.

One of the main disadvantages of the SCAPE model is that it trains the model of body shape and pose independently. It learns a pose deformation model from a subject with multiple poses and learns a shape model from many subjects with a standard pose, which neglects the correlations between the body shape and pose. Chen et al. [31] demonstrate that the decoupling of shape and pose deformations in the SCAPE model has a major limitation - 3D meshes of different individuals change in a similar manner for the same pose change. This is unreasonable because the body surface deformation generated by the motion performed by the athlete is largely different from the same motion carried out by a person with less skeletal muscle. Therefore, methods proposed in [31, 61] use shared encoding for human body modelling. BlendSCAPE utilizes a co-registration method which trains the model in the process of aligning scans with different shapes and poses. The most recent work on human body modelling is the Stitched puppet (SP) which builds a part-based human model [207]. In SP, each body part is independently translated and rotated rigidly and non-rigidly to fit into the target scans. Then, these parts are stitched together via potential functions. Besides the above scan-targeted fitting, some works are based on point clouds. Dey et al. [40] utilize the invertible finite volume method to control the template tetrahedral mesh to the target point clouds. The process, however, is tedious as the template mesh has to be converted into a volume mesh before fitting and the invertible finite volume method is needed to invert the possibly twisted tetrahedra to preserve the volume of the mesh.

SMPL (Skinned Multi-Person Linear model) [95] model is a learned linear human body model which can describe human body shape and poses in a realistic way. SMPL is a skinned vertex-based model that accurately represents a wide variety of body shapes in natural human poses. The parameters of the model are learned including the rest pose template, blend weights, pose-dependent blend shapes, identity-dependent blend shapes, and a regressor from vertices to joint locations. Unlike previous models, the pose-dependent blend shapes are a linear function of the elements of the pose rotation matrices. Compared with the SCAPE model, we can see in Figure 2.12 that SMPL model represents the identity-related blendshape and pose-related blendshape in a linear way which greatly improves the efficiency and accuracy.



(a) $\bar{\mathbf{T}}, \mathcal{W}$ **(b)** $\bar{\mathbf{T}} + B_S(\vec{\beta})$, $J(\vec{\beta})$ **(c)** $T_P(\vec{\beta},\vec{\theta}) = \bar{\mathbf{T}}+B_S(\vec{\beta})+B_P(\vec{\theta})$ **(d)** $W(T_P(\vec{\beta},\vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W})$

Figure 2.12: SMPL model [95]. (a) Template mesh in T pose with blend weights and joints; (b) Template with identity-related blend shape and joints; (c) Template with both identity-related and pose-related blend shapes; (d) Deformed template by adjusting shape and pose parameters.

Table 2.1: The comparison of the statistical models.

Models	Modelling for	Linear?	Modelling on	PDD?	V #	F #
[2]	Shape	✓	vertex	×	6890	13776
[61]	Shape&Pose	×	vertex	×	6890	13776
SCAPE [6]	Shape&Pose	×	face	✓	12500	25000
BlendSCAPE[67]	Shape&Pose	×	face	✓	12500	25000
SP[207]	Shape&Pose	×	vertex	×	12500	25000
Tenbo[31]	Shape&Pose	×	face	✓	12500	25000
SMPL [95]	Shape&Pose	✓	vertex	✓	6890	13776

Here, we summarise the differences between the popular statistical models in the Table 2.1. As it can be seen that only [2] and SMPL model are linear models but [2] only builds the shape models. Pose dependent deformation (PDD) describes the deformation induced by different poses. Triangular-deformation based modelling methods, like SCAPE, BlendSCAPE and Tenbo, allow the composition of different transformation, like body shape variation, rigid part rotation and pose-dependent deformation. To increase the expressive power, the vertex-based SMPL model includes a pose blend shape. According to [95], the vertex-based skinned SMPL model is more accurate than a triangular deformation based model like BlendSCAPE trained on the same data. Comaring to other models, SMPL is the most accurate model with lowest computation complexity.

The statistical 3D shape model has also been explored in the human face. There is a lot of research on the modelling of human faces and much of them are valuable for human body modelling. A high level of detail is one of the key requirements for human face modelling as emotion is a very important property for the face. Therefore, the modelling approaches for face usually keep more details than human body modelling methods.

Part-based modelling techniques statistical models usually capture the shape deformation using the principal component analysis technique. One of the disadvantages of PCA is that such deformation components are global. When a user intends to morph the shape locally by specifying the locally changed vertex positions, the deformed shape tends to have unrelated areas deformed as well, due to the global nature of the basis [164]. Part-based modelling techniques are commonly used in faces and human body modelling since they allow for richer representations and enable the fitting of different parts to be tailored specifically. Many linear techniques are locally used to face and body modelling. This is because the linear techniques outperform non-linear methods in terms of computational complexity and simplicity. The description ability of linear approaches is weaker than non-linear techniques. For detail modelling, many approaches decompose a mesh into parts to improve expressiveness. Tena et al. [166] develop a region-based linear approach for face modelling. Each subpart is controlled by a PCA model which is independently trained. For the sake of boundary continuity, the method adds a constraint on the boundary parts. In this way, the method allows for flexibility while maintaining coherence. To define the part-based model, a segmentation of the training shapes into meaningful parts is required. The segmentation of the face model in [166] is based on spectral clustering. The active nonrigid ICP [32] uses a segmented face model-annotated model of the face to perform the region-based face modelling. Besides these segmentation methods which are directly performed on the shape space, Brunton et al. [24] develop a wavelet decomposition of 3D face geometry. The method requires no explicit segmentation of the face into parts; the wavelet transform decomposes the surface hierarchically into overlapping patches and the inverse transform recombines them.

2.5 Surface Modelling

After discussing currently popular human body statistical models in the above, the human body surface construction approaches are investigated in the following. We are focusing on the surface modelling methods that build full body meshes using commodify sensors. The methods are divided into two kinds: template-free surface modelling and template-based surface modelling.

2.5.1 Template-free Surface Modelling

In the initial stage of reconstruction with commodity scanners, many offline reconstruction systems are proposed for scenes, human bodies and human faces [20, 90]. KinectAvtar [36] firstly applies the super-resolution algorithm to acquire new super-resolved depth maps with much higher resolutions and less noise. Then global rigid and non-rigid alignment steps combine the super-resolved scans into a final model. For each scan, it takes around 14 minutes. The work of 3D Self-portraits implements the scanning of users with a single 3D Kinect by rotating the same pose for a few different views [89]. Then it non-rigidly registers the scans captured in each view into a watertight surface. The work in [43] reconstructs high-quality objects based on a single stream from Kinect but it requires several hours to complete.

KinectFusion is the first work that implements the reconstruction scenes in real time with commodity scanners [109]. Four components make up the complete real-time scanning systems. (a) A pre-processing step is performed to generate point clouds and calculate the normal for each vertex in the point clouds; (b) the camera pose is estimated in a frame-to-model way which aligns the point clouds of current frame with the globally fused model; (c) given the camera pose determined by tracking the depth data from a new sensor frame, surface is reconstructed by updating the scene model maintained with a volumetric, truncated signed distance function (TSDF) representation. The TSDF represents the surface voxels as zeros, the voxels in the free space as positive values that increases with the distance from the surface, and the voxels in the inner space as negative values; (d) the point clouds are acquired from the depth map of current frame by ray casting the TSDF and normals are computed for the alignment with next frame. As KinectFusion is proposed under the assumption of static scenes, DynamicFusion [110] implements the real-time reconstruction of dynamic scenes under non-rigid deformation. The key insight of DynamicFusion is building a canonical model and estimating the volumetric model-to-frame warp field parameters. With the estimated warp field, the current frame depth map is fused into the canonical space. To capture the newly emerging surface geometry, the warp filed is extended to ensure deformations are represented. DynmicFusion inspires a lot of follow-up works, like VolumeDeform [73], which uses sparse RGB feature matching to improve tracking robustness and handle scenes with little geometric variation, allowing for reconstruction of newly emerging parts in real time.

Although DynamicFusion proposes the approach to reconstruct meshes in real time with both scenes and camera moving, some major changes in shape and topology are still hard to accommodate. For example, it is difficult to reconstruct the fast interaction between two people or with objects. This is caused by the volumetric representation which is incrementally updated with new depth input. This reference model confines so that it is hard to reconstruct the cases where quick and dramatic changes that happen in topologies and shapes. Therefore, Fusion4D [44] proposes to address the problem of the reconstruction with dramatic changes in shapes and topology. Fusion4D takes multiple RGB frames as input and first estimates a segmentation mask per camera. Different from DynamicFusion where the first frame is fixed as the reference frame, Fusion4D periodically updates the reference to a fused data volume, called key volumes.

The fixed reference in DynamicFusion constraints the reconstruction of the case where new topology emerges or disappears, while the updating reference allows Fusion4D to be more responsive to new data. The impressive reconstruction performance makes it be the supporting technical behind the popular application of Holoportation [114]. In Body-Fusion, Tao et al. [193] narrow the research domain into the human body and provide a more robust real-time fusion method for dynamic scenes. It uses an articulated skeleton to define the warp field instead of a coarse general deformation graph. As the skeleton parameterization is low-dimensional, the tracking problem is drastically simplified and the approach produces more stable reconstructions for the human case. Unlike DynamicFusion or VolumeDeform which require highly-controlled motion resulting from the underlying mesh-based correspondence estimation between frames, KillingFusion [155] implements real-time non-rigid 3D reconstruction from a single RGBD stream without any priors, such as skeleton, shape prior. It can handle fast motion and topology changes. KillingFusion uses a signed distance field (SDF) to represent the shape and incrementally evolve the projective SDF of the current frame towards the target SDF. During fusing the current frame to the target SDF, the main energy component is designed to align the current frame to the cumulative model by minimizing their voxel-wise difference of signed distances - thus without explicit correspondence search and the approach is suitable for parallelization.

Beyond only reconstructing the geometry, the recent work in [57] does some optimizations for the reflectance and illuminations. This has multiple advantages. For example, it makes the tracking robust against the changes of scene illuminations. In addition, it enables some useful applications like relighting in random environment.

2.5.2 Template based Surface Modelling

The real-life scenarios where objects move and interact non-rigidly pose many challenges to the research community of 3D reconstruction. The difficulty stems from a large number of unknown parameters and the inherent ambiguity of the problem, as various deformations can generate the same shape. This situation can be alleviated by template-based approaches.

Tong et al. [168] use three Kinects to capture the upper, middle and lower parts of a human body. Using the measurements of the first frame, Tong et al. construct a quite rough template using the statistical body shape model which was proposed in

[60]. Then pairwise non-rigid deformation is performed between successive frames and then deformed back to the first frame. The frame-to-frame registration could suffer from error accumulation where the first and last frames do not match well. Thus, the second step of global registration is applied to distribute errors in the deformation space. As this system fuses frames from three Kinects, it requires calibration before capturing and several minutes to construct a complete model. Liu et al. [92] automatically build human body models with a template without any human-assigned markers. It firstly captures a raw scan of a human body with calibrated multiple Kinect sensors. A simplified template is deformed to the scan to build the correspondences between the template and the scans. With the established correspondences, a higher-resolution version of the template is deformed to the target scan to acquire the complete and believable human body models. This work also needs the calibration process in the first place. It requires around 10 seconds to reconstruct a high-resolution human body model.

Besides the above works that use multiple sensors, there are many systems using one sensor to reconstruct bodies. Weiss et al. [179] take four views of the body with a single Kinect and estimate 3D point clouds by deforming SCAPE parametric human body model with the constraints of the silhouette of RGB image and range data. In [16], the 3D geometry and appearance of the human body are estimated from a monocular RGB-D sequence of a user moving freely in front of the sensor. It brings the range data into alignment with a proposed parametric 3D body model, called Delta. Delta is an extension of BlendScape by proposing a variable-detailed shape model for both bodies and heads. Chen et al. [29] reconstruct 3D human models with a Kinect. Firstly, the depth images captured from multi-views with a single Kinect are registered. Then a statistical human model is utilized to iteratively augment and complete the human body information by fitting the statistical model to the registered depth image.

The above works build complete and high-quality human body meshes by fitting a template to target depth image or point clouds. However, they rely on priors of an offline learned model. In the following, we introduce some works based on a template but can process in real time. Zollhöfer et al. [206] deform a template to incoming depth in real time. Rather than using a statistic model of a specific domain, it firstly builds a template by scanning the subject as they move rigidly. This geometry prior avoids strong scene assumption, but at the same time, it requires the subject to stay absolutely still during the template generation. Next, a GPU pipeline performs non-rigid registration of live RGB-D data to the smooth template based on an as-rigid-as-possible (ARAP) framework.

2.6 3D Reconstruction from Images

For a human, it is usually an easy task to perceive the 3D structures shown in an image. However, it is difficult to estimate the true 3D geometry of the objects because of the loss of depth information in the projection process. Here, the 3D reconstruction methods using the multi-view images and the deep learning based methods are discussed.

2.6.1 Multi-View 3D Reconstruction

Infinite different 3D surfaces may produce the same set of images. If we take images from multiple views, we can infer the geometry of the object more accurately. In the situation of 3D reconstruction of objects or human bodies from a collection of images taken from multiple views, the camera positions and the internal parameters are assumed to be known or they can be estimated by the images. Usually, there are three steps to reconstruct 3D models from multiple views. First of all, the correspondences between multiple images are established using image features. Second, with the correspondences, the position and orientation of cameras are estimated using triangulation. Third, with the estimated camera parameters, depth information of the key points in the images can be calculated. Many researchers seek to estimate a complete depth map. Narayanan et al. [108] use a traditional multi-baseline stereo matcher for a complete depth map. Goesele et al. [52] improve the calculation of depth map by only reconstructing the portion of the scene that can be matched with high confidence in each input view. Then the following merging step fills the holes that are close to silhouettes or caused by oblique surfaces, occlusions, highlights, low-textured regions et al. These above methods require many views (for example up to 317 views required). Otherwise, the holes will appear in the reconstructed results. In [152, 87], the texture information is utilised to help carve away the inconsistent points but in practice texture information is not always available. In the above point based methods, establishing correspondences is not a simple and clean method, as point correspondences can only be reliable for the subset of the scene in some cases. For regions without salient features, it is not easy to establish reliable correspondences. Additionally, even in the case where hundreds of images are given, the reconstruction from 2D is still an ill-posed problem as there are several different 3D models consistent with an image sequence. Therefore, some approaches use prior knowledge to regularize the problem. Jin et al. [77] estimate the 3D shape and appearance of a scene from a calibrated set of views by relying on an affine subspace constraint that must be satisfied when the scene exhibits "diffuse + specular" reflectance characteristics. This constraint is used to define a cost function for the discrepancy between the measured images and those generated by the estimate of the scene, rather than attempting to image-to-image match directly. This method is able to reconstruct the scenes with strong specularities that are a challenge to methods relying on image-to-image matching.

2.6.2 Deep Learning based End-to-End 3D Reconstruction

Currently, with the available large dataset of 3D objects and the development of deep learning, a number of learning based methods are proposed to reconstruct 3D objects from 2D images. Lun et al. [99] propose to infer 3D shapes from 2D sketches by using a deep encoder-decoder network. The encoder converts the sketch into a compact representation which encodes shape information and the decoder converts this representation into depth and normal maps capturing the underlying surface from several output viewpoints. These multi-view maps are then fused into complete point clouds. Then polygon meshes are constructed for the point clouds. Choy et al. [34] propose a recurrent neural network, 3D-R2N2, to learn a mapping from images of objects to their 3D shapes from the large collection of 3D CAD models, ShapeNet which consists of 50,000 models and 13 major categories. The output of this work is voxel representation. The approach suffers from the limited resolution of the voxel representation.

Although currently emerging works on the reconstruction of the 3D objects using deep learning methods present inspiring results. Most of them are voxel- or point clouds-based. The voxel representation rasterizes a 3D shape as an indicator function or distance function sampled over dense voxels and applies a deep neural network over the entire 3D volumes. As the memory and computation costs grow cubically as the voxel resolution increases, the voxel resolution is usually limited in the current methods. Moreover, even though the ShapeNet provides the training set for learning, when it comes to the specific domain of human body shape and poses, the training dataset is still insufficient.

2.7 Summary

The objective of this chapter is to give the relative background of human body modelling. With the appearance of commodity sensors in recent years, the acquirement of 3D and 2D data facilitates the modelling of the human body with great ease. In this chapter, we firstly introduce several types of low-cost 3D scanners. Their advantages and imperfections are analyzed. The basic concepts of surface registration are illustrated in a later step. In the following, the 3D human body models are also introduced from the primary geometric models to the current realistic statistical body models. Finally, the current human body modelling approaches from 3D sensing data and 2D images are discussed in the end.

Chapter 3

Symmetry-aware Human Shape Correspondence under Closed Mesh

As discussed in the Introduction, the core insight of this thesis is to align the template to two types of target data: point clouds or images. In the case of point clouds, the most intuitive way is to establish the one-to-one correspondences between the template and the target. For this objective, the ideal scanning environment is usually considered where the captured scans are smooth and watertight. In this situation, the goal is to find the correspondences between the template and the target scans based on geometry features. Finding accurate shape correspondences can provide automatic rigid/non-rigid shape registration, which lays a solid foundation of shape completion, shape morphing and shape statistical modelling. In the research community of 3D shape analysis, finding sparse or dense correspondences is a fundamental but nontrivial problem. In the case of the human body, the various human body shapes/poses and incomplete parts (missing data) caused by occlusion or scanning artefacts make the local/global geometry features change. These factors pose more challenges on establishing reliable dense correspondences between user-generated scans.

Besides the above-mentioned problems, the symmetric flip problem exists in establishing correspondences for intrinsically symmetric models. It is challenging for the current correspondence establishment approaches to completely distinguish between the left and the right. To tackle the correspondence ambiguity problem that arises when matching two intrinsically symmetric shapes as well as improve the accuracy of locating the final corresponding pairs, in this chapter, the state-of-the-art approach is extended by using skeleton information to further remove symmetric flipped shape correspondences. To discriminate symmetric surface points, after the initial correspondences are located, the candidate sets for each point on the template are built, followed by making use of skeleton to remove the symmetrically flipped false candidates. In the remaining candidates, final correspondences are achieved by choosing those with the minimum geodesic error from a base vertex set, which is formed by sampling on the mesh. Experiments demonstrate that the proposed method can effectively remove all the symmetrically flipped candidates. Moreover, the final correspondence pair is more accurate than those of the state of the arts.

3.1 Overview

3D shape correspondence is a mapping from one point set on the source mesh to another on the target mesh. There exist three kinds of mapping: one-to-one, one-to-many and many-to-one. In this chapter, the goal is to address the problem of establishing the accurate one-to-one correspondence between intrinsic-symmetrically isometric human models.

The target of shape correspondence is to find the point pairs that are similar or semantically equivalent. Isometric shapes appear in various contexts such as different poses of an articulated human model or two shapes presenting different but semantically similar objects [145]. It is highly demanded to find isometric shape correspondence since most real world deformations are isometric. Moreover, finding shape correspondences between isometric shapes is of practical values. For instance, the deformation based on the isometric template will be much more efficient benefiting from their similar shapes. If two shapes are totally isometric, the geodesic distance between two points on one shape should be equal to the geodesic distance between their correspondences on the other shape [145].

Embedding-based methods are popular techniques for the 3D shape correspondences problem. In these methods, an original mesh is embedded into a new domain where matching can be easily performed. Euclidean embedding can be achieved by using various techniques such as least-square MDS [45], heat kernel embedding [116] and spectral embedding [75], with which the original mesh is mapped into the Euclidean domain. Besides embedding methods, other approaches [145, 144] minimize the isometric distortion directly in the 3D Euclidean space. However, most existing algorithms tend to be confused by the intrinsically symmetric features and suffer from symmetric flip problems. They can hardly discriminate symmetric points on the surface even if the mesh to be matched is not perfectly symmetric. Therefore, it is common that the correspondence of the point on the right hand of the source mesh is established on the left hand of the target mesh, as shown in Figure 3.1.

In this research, a robust method is proposed to find correspondences for human isometric shape model which is able to solve the symmetric flipping problem. The idea here is to combine skeleton information to distinguish intrinsic symmetry. Given two meshes with their skeletons, local features are firstly detected to find one-to-many correspondences between the two meshes. The candidate set for each feature point presents symmetric properties on the mesh. A skeleton segment associated with surface points is capable of discriminating symmetry. The final correspondence is located and refined by minimizing the isometric distortion with respect to based vertex set.

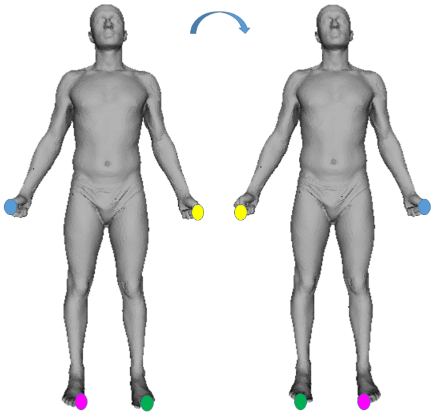


Figure 3.1: Flipping correspondences. Each correspondence pair is labelled with the same colour.

3.2 Related Work

Shape correspondence is a long- but non-trivial problem. In areas such as shape matching, 3D shape retrieval, mesh registration, and 3D reconstruction, many efforts are made on finding shape corresponding points on two meshes recently.

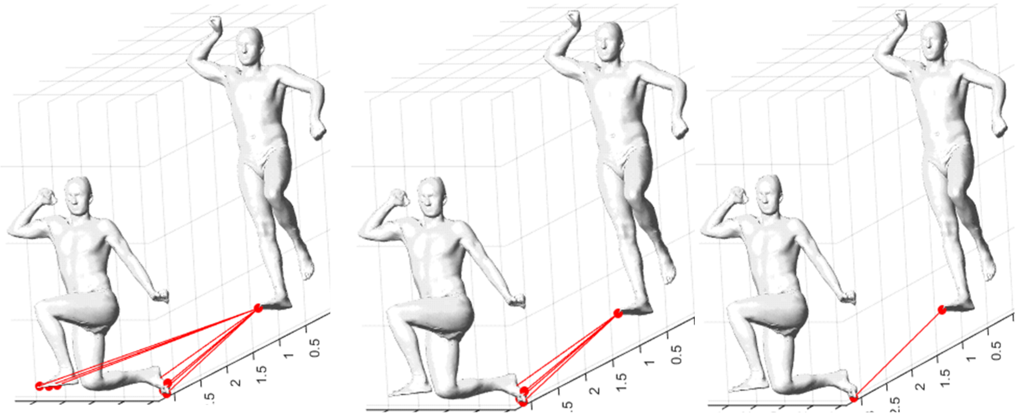
The SCAPE model uses markers to manually locate correspondences between two meshes [6]. Besides manual assignment, feature matching is applied in many works to automatically establish correspondences. Some local descriptors in 2D images are extended into 3D domains, such as MeshHOG [194] and 3D shape context [84] for 3D point sets, the Spin images [78] and Multi-scale features [91] for oriented points, and curvature maps [50] and salient geometric features [49] for surfaces. However, these shape features are hard to be preserved under non-rigid deformation which often occurs in the human body case.

The establishment of correspondence for shapes under non-rigid deformation can be accomplished by adopting the embedding-based method which is a more reliable approach when it comes to isometric deformation. Multidimensional Scaling (MDS) [45] approximates geodesic distance with Euclidean distance in embedding space and the correspondence problem is treated as a nearest neighbour search in the embedding space [143]. Dey et al. [40] use the Global Point Signature (GPS) [142] for spectral embedding of meshes and thereby find the shape correspondence at mesh extremities for initialization. The method in [144] also transfers vertices into the spectral domain and optimizes for the best match using the expectation-maximization algorithm. However, these methods sometimes provide false correspondence due to the presence of model symmetries.

Ovsjanikov et al. firstly represent to-be-matched shapes using the functional map proposed in [117] to identify the symmetric parts of an object, and then factor them out [118] but the symmetry map of one of the two shapes has to be known. Zhang et al. differentiate the intrinsic symmetric points by calculating a signed angle field from the gradient fields of the harmonic field which is derived from four points on the hands and feet [201]. Sahilliolu et al. propose a coarse-to-fine scheme to track symmetric flips [145]. Although the above methods improved the detection of accurate nonrigid shape correspondences based on the embedding approach, none of them can completely remove the symmetric flipped false correspondences in an automatic way.

3.3 Skeleton-based Symmetry-aware Shape Correspondence

The intrinsic symmetry leads to symmetric flipped correspondence between two meshes. Neither embedding-based methods like MDS, GMDS nor local descriptors can differentiate them effectively. Previous works which are solely relying on surface-related information, i.e geodesic distance or face normal, are unable to solve symmetric problems completely. However, with the help of a set of skeleton information where different skeleton segments have different labels and surface point and skeleton segment are associated, it is possible to address the symmetric flipping problem with skeleton. Moreover, the appearance of Kinect camera enables to obtain the skeleton of a mesh with ease. To perform the skeleton attachment process, the method in [10] can be used to attach skeletons for meshes, in which the input is the joint positions tracked by Kinect. The output is the skeleton attached to the human model. Therefore, it can be assumed that the meshes with the skeleton attachment are available in this chapter.



(a) Expanded candidates (b) Skeleton-filtered candidates (c) final correspondence pair

Figure 3.2: The workflow of proposed skeleton-based shape correspondence method: (a) the expanded candidate set for one point on the source; (b) the after-filtered candidate set using skeleton filtering; (c) the final one candidate point on the target based on base vertex set.

In the following of this chapter, the focus is on how to tackle the symmetric flip problem for locating the accurate one-to-one correspondences. The workflow is shown in Figure 3.2. Firstly, the candidate set for each source point is established. For each point on the source, 50 candidates on the target mesh are detected with HKS which are distributed on both sides of the body as shown in Figure 3.2 (a). Then, a skeleton is used to remove the symmetric-flipped false candidates which is shown in Figure 3.2 (b). Among the remaining candidates, the base vertex set that is described in Section 3.3.3 is used to locate the final one shape correspondence and improves the accuracy of the shape correspondences.

3.3.1 Correspondence Candidate Set

As mentioned before, in order to make sure that the candidate set includes the correct correspondence as much as possible, the one-to-many correspondences are computed using Heat Kernel Signature (HKS) [162], and the top N similar points are selected to construct the candidate set which is shown in Figure 3.2 (a). To compute the heat of point i at time t_i, we firstly perform the Laplace-Beltrami operator L on the mesh. The Laplace-Beltrami operator is an extension of the Laplacian to the manifold. Like the Laplacian, the Laplace-Beltrami operator is defined as the divergence of the gradient [149]. As for complex surfaces like the human body, it is difficult to acquire the explicit representation for the Laplace-Beltrami operator. Many discrete Laplace-Beltrami operators are proposed to approximate it (please refer to [134] for comparison). Here, we adopt the most widely-used cotangent approximation scheme of the Laplace-Beltrami operator on the triangular meshes [124]. Specifically, the Laplace-Beltrami operator will be defined in order to satisfy the Eq. 3.1.

$$L(\mathbf{p}_i) = \frac{1}{2A(\mathbf{p}_i)} \sum_{\mathbf{p}_j \in N_1(\mathbf{p}_i)} (\cot \alpha_{ij} + \cot \beta_{ij})(\mathbf{p}_j - \mathbf{p}_i), \tag{3.1}$$

where $A(\mathbf{p}_i)$ is the area of the Voronoi region of vertex $\mathbf{p}_i$; $N_1(\mathbf{p}_i)$ is the neighbours of vertex $\mathbf{p}_i$ and α_{ij} and β_{ij} are the angles opposite the edge $(\mathbf{p}_i, \mathbf{p}_j)$.

For computation, we use matrix to encode the Laplacian. First, given vertices with index $1, 2, \ldots, |\mathbf{P}|$, the matrix $\mathbf{W} \in \Re^{|\mathbf{P}| \times |\mathbf{P}|}$ represents the weight part in the Eq. 3.1 where W_{ij} and W_{ii} are the values in the follwing:

$$\begin{aligned} W_{ij} &= \frac{1}{2}(\cot \alpha_{ij} + \cot \beta_{ij}) \quad for \quad j \in N_1(\mathbf{p}_i), \\ W_{ii} &= -\sum_{j \in N_1(i)} W_{ij}. \end{aligned} \tag{3.2}$$

All other entries are zero. It can be seen that the matrix $\mathbf{W}$ only encodes part of Laplacian Beltrami operator. The Voronoi region area A should be also incorporated.

Thus, a diagonal matrix $\mathbf{M} \in \Re^{|\mathbf{P}| \times |\mathbf{P}|}$ is built.

$$\mathbf{M} = diag(A(\mathbf{p}_1), A(\mathbf{p}_2), \ldots, A(\mathbf{p}_{|\mathbf{P}|})). \tag{3.3}$$

Laplacian-Beltrami operator is then: $\mathbf{L} = \mathbf{M}^{-1}\mathbf{W}$. Let $\mathbf{\Lambda}$ be the diagonal matrix of the eigenvalues of $\mathbf{L} \in \Re^{|\mathbf{P}| \times |\mathbf{P}|}$, and $\mathbf{\Phi}$ be the matrix with the corresponding eigenvectors, the heat kernel of the mesh is computed as Eq. 3.4:

$$\mathbf{K}_t = \mathbf{\Phi} \exp(-t\mathbf{\Lambda})\mathbf{\Phi}^T. \tag{3.4}$$

Each entry in $k_t(i, j)$ represents the heat diffusion between points $\mathbf{p}_i$ and $\mathbf{p}_j$. The diagonal elements of this matrix is composed of HKS. Thus, a HKS feature is a vector whose entry $k_{t_j}(\mathbf{p}_i, \mathbf{p}_i)$ is the heat at point $\mathbf{p}_i$ at time of t_j:

$$\{k_{t1}(\mathbf{p}_i, \mathbf{p}_i), k_{t2}(\mathbf{p}_i, \mathbf{p}_i), ..., k_{tn}(\mathbf{p}_i, \mathbf{p}_i)\}. \tag{3.5}$$

When the dissimilarity of HKS between the template point and target point in Eq. 3.6 is less than a threshold t, the target point is selected as candidate for the template point.

$$\Delta s = ||HKS(\mathbf{p}_t) - HKS(\mathbf{p}_s)||, \tag{3.6}$$

where $HKS(\mathbf{p})$ is the heat kernel signature at point $\mathbf{p}$, $\mathbf{p}_t$ and $\mathbf{p}_s$ are the points on the template and target respectively. Here, the scale-invariant HKS (si-HKS) [23] is applied to detect features for meshes.

After the initial correspondence is achieved by si-HKS, an expanded set of candidate points are obtained as shown in Figure 3.2 (a). As it can be observed, the expanded candidates for the point on the right foot of the target model distribute on both feet of the target model, presenting the symmetric property.

3.3.2 Skeleton-filtered Correspondence Candidates

To locate the single correspondence for template point, the next step is to remove those symmetric flipping points. Skeleton is an important clue for filtering flipped correspondences. As shown in Figure 3.3, skeleton divides mesh into 17 parts and each mesh part attached a segment has a unique label and the right extremity and its left counterpart have different labels. Therefore, the proposed method is able to discriminate the right points and their counterparts on the left, addressing symmetric flip problems. When the template point and candidate points are on the same skeleton segment, they are kept; otherwise, the candidates are removed. The filtered candidate set for template point is shown in Figure 3.2 (b).

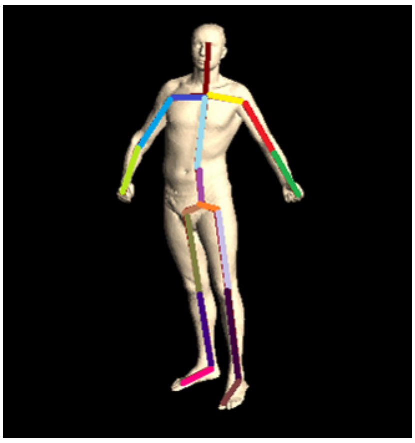


Figure 3.3: Mesh division by skeleton; each colour represents a skeleton segment.

3.3.3 One-to-one Correspondence

After the symmetric flip problem is solved, the remaining candidates need to be further filtered to find the one-to-one correspondence pair. Therefore, the next step in the proposed method uses the sum of relative distances from candidates to the base vertex set to filter invalid candidates.

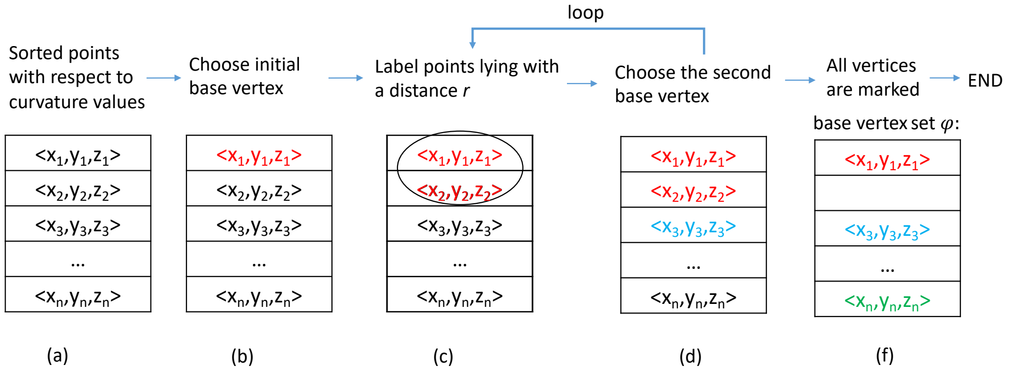

loop
Sorted points with respect to curvature values
Choose initial base vertex
Label points lying with a distance r
Choose the second base vertex
All vertices are marked
END
base vertex set φ:
<x1,y1,z1>
<x2,y2,z2>
<x3,y3,z3>
...
<xn,yn,zn>
(a)
(b)
(c)
(d)
(f)

Figure 3.4: The process of base vertex set selection.

The base vertex set [144] is selected based on the Gaussian curvature which is the product of the two principal curvatures at a vertex on the surface. Thus, the Gaussian curvature of the points on the shape of a cylinder is zero while mean curvature is not. As the human body can be approximated by a set of cylinders [154], the Gaussian curvature can well reflect the locally salient points in the case of the human body. This process of base vertex selection is illustrated in Figure 3.4. Initially, at each vertex of the original mesh, the Gaussian curvatures are computed using a simple way proposed in [139] with E.q. 3.7.

$$gc(\mathbf{p}) = 3(2\pi - \sum \alpha_i)/\sum A(f_i), \tag{3.7}$$

where $A(f_i)$ is the area of the face f_i that is adjacent to the vertex $\mathbf{p}$ and the angle α_i is the angle of f_i at the vertex. Then the vertices are sorted into a list in descending order with respect to their curvature values like in Figure 3.4 (a) and the top vertex is chosen as the first base vertex, e.g. marked point (x_1, y_1, z_1) in Figure 3.4 (b). Then, as shown in Figure 3.4 (c), the geodesic distance is computed from this vertex and all its neighbouring points lying within a radius r are marked. In the experiment, the Dijkstra's shortest path algorithm is adopted to compute the geodesic distance between two vertices as E.q. 3.8. The weight of each edge of Dijkstra's path is the Euclidean distance between neighbouring vertices and is defined by E.q. 3.9.

$$g(\mathbf{p}_i, \mathbf{p}_j) = \sum_{\mathbf{p}_i, \mathbf{p}_j \in \mathbf{P}} \omega_i, \tag{3.8}$$

$$\omega_i = \min_{\mathbf{p}_k \in N_i} ||\mathbf{p}_i - \mathbf{p}_k||, \tag{3.9}$$

where N_i is the neighbours of point $\mathbf{p}_i$. The next base vertex is the first unmarked vertex in the list like (x_3, y_3, z_3) in Figure 3.4 (d). This process is repeated until all points are marked and based vertex set is built. The final base vertex set is illustrated in Figure 3.4 (f). Given base vertex set ϕ, the relative surface distance is calculated from each candidate to ϕ with E.q. 3.10. The candidate $\mathbf{C}$ with the minimum relative distance to ϕ is regarded as the final correspondence as shown in Figure 3.2 (c).

$$D_{iso}\left(\mathbf{c}_i, \phi\right) = \sum_{(\mathbf{v}_j \in \phi, \mathbf{c}_i \in \Theta)} g(\mathbf{c}_i, \mathbf{v}_j), \tag{3.10}$$

$$\mathbf{C} = arg \min_{\mathbf{c}_i \in \Theta} \left(D_{iso}\left(\mathbf{c}_i, \phi\right) - D_{iso}\left(\mathbf{p}, \phi\right)\right). \tag{3.11}$$

Here, $g(.,.)$ is the geodesic distance between two vertices. $\mathbf{p}$ is the vertex on the source that wants to find its correspondence on the target. $D_{iso}(\mathbf{p}, \phi)$ is the distance from the point $\mathbf{p}$ to the base vertex set. The distances from the candidates to *base vertex set* are acquired using the Eq. 3.10 and the errors between $D_{iso}(\mathbf{c}_i, \phi)$ for candidate $\mathbf{c}_i$ and $D_{iso}(\mathbf{p}, \phi)$ for vertex $\mathbf{p}$ on the source is computed with Eq. 3.11. The candidate with minimum error is regarded as the final correspondence.

To be more clearly, the proposed algorithm is further illustrated. For detecting the correspondences for each point $\mathbf{p}_i$ on the source mesh, the process of locating the final accurate correspondence on the target is shown in Figure 3.5.

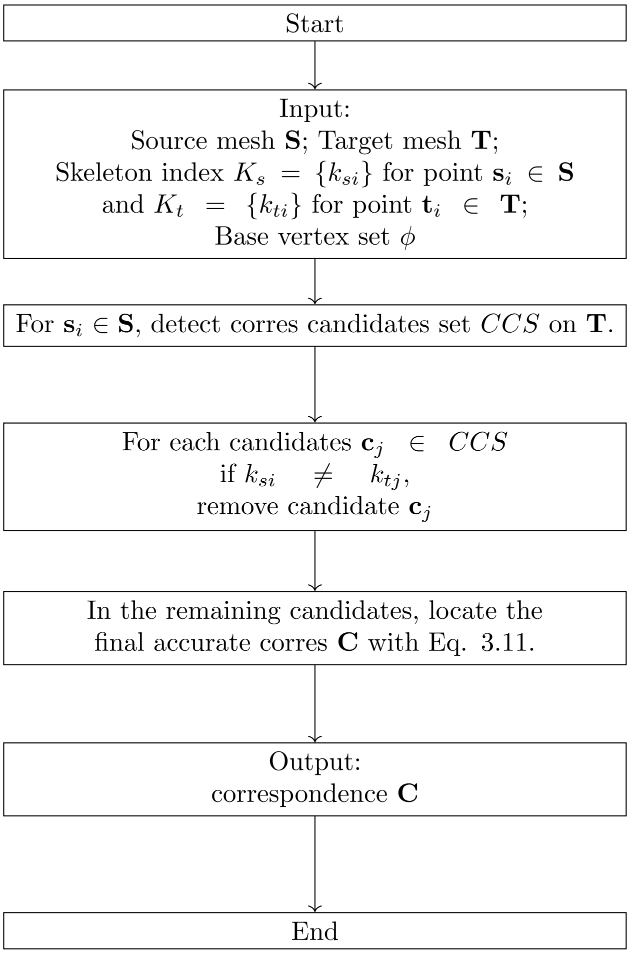

Start
Input:
Source mesh $\mathbf{S}$; Target mesh $\mathbf{T}$;
Skeleton index $K_s = \{k_{si}\}$ for point $\mathbf{s}_i \in \mathbf{S}$
and $K_t = \{k_{ti}\}$ for point $\mathbf{t}_i \in \mathbf{T}$;
Base vertex set ϕ
For $\mathbf{s}_i \in \mathbf{S}$, detect corres candidates set CCS on $\mathbf{T}$.
For each candidates $\mathbf{c}_j \in CCS$
if $k_{si} \neq k_{tj}$,
remove candidate $\mathbf{c}_j$
In the remaining candidates, locate the
final accurate corres $\mathbf{C}$ with Eq. 3.11.
Output:
correspondence $\mathbf{C}$
End

Figure 3.5: The flowchart of skeleton-based symmetry-aware correspondence algorithm.

3.4 Experiments

The proposed method is evaluated with the state-of-the-art in the dataset of SCAPE [6]. The SCAPE model contains both symmetric and various deformed shapes. For a detailed description of SCAPE dataset, please refer to Section 2.2 in Chapter 2.

3.4.1 Performance Evaluation

Table 3.1: The quantitative comparison of the proposed method against the C2FCM method.

GeoErr %Corr Method		10	20	30	40	50	60	70	80	90	100
C2FCM	**Mean**	0.122	0.103	0.093	0.093	0.092	0.097	0.098	0.097	0.097	0.098
Ours		0.097	0.074	0.070	0.074	0.081	0.084	0.082	0.081	0.080	0.083
C2FCM	**SD**	0.020	0.025	0.022	0.021	0.024	0.017	0.024	0.022	0.021	0.023
Ours		0.022	0.019	0.020	0.021	0.019	0.018	0.020	0.020	0.021	0.020

The average geodesic error (GeoErr) is quantitatively compared with C2FCM algorithm in Table 3.1 which shows the average geodesic error and the standard deviation (SD) for different proportion of the dataset. For each proportion of the dataset, we adopt the simple random sampling without replacement to obtain the sub-dataset. It is assumed that the correspondences of two isometric shapes should have the same geodesic distance from the base vertex set. Thus, we define the average geodesic error using Eq. 3.12 to evaluate the performance of the proposed method.

$$avg\_Geo\_Err = \frac{\sum_{i=1}^{n}(D_{iso}(\mathbf{S}_i, \phi) - D_{iso}(\mathbf{T}_i, \phi))}{n}, \tag{3.12}$$

where D_{iso} is defined in Eq. 3.10; ϕ is the based vertex set; $\mathbf{S}_i$ and $\mathbf{T}_i$ are the correspondence i on the source and target meshes; and n is the number of correspondences. It can be seen that for different proportions of correspondences, the geodesic errors of the proposed method are less than those of C2FCM. The average of geodesic error of all correspondences, shown in the column of 100% correspondences, the proposed method outperforms C2FCM, which means the proposed method is able to find the correspondences more accurately. The comparison of the standard deviation shown in Table 3.1 displays the distribution of geodesic error for each proportion of dataset. It is shown that our method is more stable than the C2FCM method.

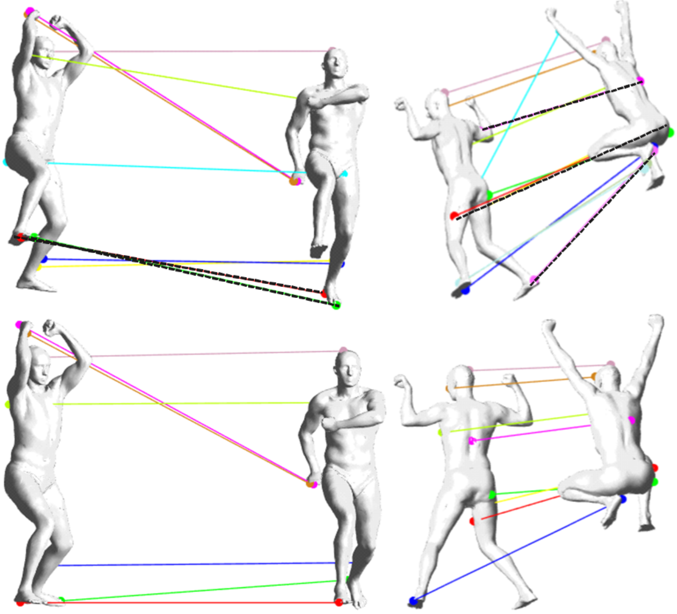


Figure 3.6: The qualitative comparison between two methods: Top: results obtained using the method in [145]; Bottom: results of the proposed method. Matched point pairs are connected by lines. Symmetry flips are connected by black dash lines.

More results are shown in Figure 3.6. It can be seen that in coarse-to-fine algorithm correspondences which are shown in the top line present symmetry properties (both on the left and right foot) for template point. The proposed method is able to find the unique correspondences correctly compared to C2FCM [145] in terms of semantics. Moreover, it successfully removes the symmetric flipped invalid correspondences and achieves accurate one to one mapping from template to target meshes.

3.5 Summary

For robust registration, the intuitive way is to find dense correspondences. Therefore, in this chapter, a robust method is proposed to locate the shape correspondences for human body meshes. It is able to address the nontrivial symmetric flip problem in shape correspondence research area and improve the accuracy of the detected correspondences. This approach effectively removes the flipped correspondences by introducing skeleton information and minimizing distortion error. With the introduction of based vertex set, the proposed method can locate the one-to-one semantically similar correspondence more accurately. Experimental results indicate that the proposed approach outperforms traditional approaches that rely only on surface information. This approach addresses the symmetry flip problem in finding correspondences for closed meshes.

However, the proposed method still requires the input shapes to satisfy some strict conditions such as being watertight and no holes. In addition, in a practical scene, two human body meshes in different poses are hard to be perfectly isometric because of missing data, distortion parts and topological noise. These imperfections make the feature based methods infeasible in real user-generated scans. Therefore, in the next chapter, user-generated scans will be considered and robust registration approaches which can align the template shape to those user-generated target scans will be proposed.

Chapter 4

Building High-fidelity Human Body from User-scanned Point Clouds

When it comes to human body mesh registration, one approach is to establish the dense correspondences between two meshes of different identities and poses. As discussed in the Chapter 3, most current methods that extract mesh features still suffer from the inaccurate representation of meshes, especially those are with holes, topological noise, in different shapes and self-symmetric. Another type of traditional approaches employs nonrigid ICP to perform registration. However, nonrigid ICP is sensitive to local optima. Therefore, a set of correspondences are usually established manually or automatically for a good initialization [6, 184]. Textures and markers are commonly used to generate reliable registration results [15]. Moreover, the majority of current methods are tailored for high-quality scans from expensive scanners [67, 15, 6, 127].

The recent advances of 3D scanning techniques have enabled the convenient acquisition of 3D scans. Nowadays 3D human body meshes can be straightforwardly acquired with affordable scanners within 2 minutes [109, 110]. More and more portable devices are compatible with 3D scanning techniques, making it popular to scan objects whenever and wherever possible. Although the technology development prompts an extremely large amount of user-generated 3D data that can be regarded as a valuable source for building large-scale 3D datasets. To be useful in vision research, these scans must be registered, i.e., aligned with a common topology. However, in practice, user-generated 3D data is usually full of noise, holes (missing data), self-contact and distorted parts. Due to non-ideal light condition, inevitable movements of the subjects and untrained operations during scanning, some body parts are severely obscured and may not even keep their shapes. In these cases, a complicated and delicate registration method designed for high quality mesh often produce noisy and meaningless results. Missing data which causes holes on the surface is prevalent, especially in the parts of head top, armpit and crotch. Particularly, missing data heavily degrades the performance of non-rigid registration, which is particularly evident for those closest-point based methods [2, 5], as they tend to get trapped in the local minimum. To make the matters worse, the large variation among different body poses could heavily affect their robustness, as they usually require a good initialization.

Therefore, in this chapter, firstly, a capturing platform is shown which is composed of a single Kinect and a running turntable. Based on this platform, a dataset of human body scans captured with a commodity 3D scanner - Kinect - by general users is built. The collected dataset contains 55 identities in 5 different poses. Distorted parts, holes and irregular poses are presented in the scans. These imperfections of source data pose challenges to build high-fidelity human body meshes.

After the user-generated scans are acquired with the proposed platform and setup, a robust approach is presented aiming at registering meshes with unneglectable flaws. A fully automatic active registration method is proposed to deform a high-resolution template mesh to match the low-quality human body scans with similar poses. The proposed registration method operates on two levels of statistical shape models: (1) the first level is a holistic body shape model that defines the basic figure of a human; (2) the second level includes a set of shape models for every body part, aiming at capturing more body details. The fitting procedure follows a coarse-to-fine approach that is robust and efficient. Experiments show that the proposed method is comparable with the state-of-the-art methods for high-quality meshes in terms of accuracy and it outperforms them in the case of low-quality scans where noise, holes and distorted parts are prevalent.

Besides the case where the target scan has the similar pose with the template, the approach is extended to address the registration problem in different poses with the help of a sparse set of manually-annotated landmarks.

4.1 3D Data Acquisition

In this section, the scanning platform and setup which are used to capture our user-generated point clouds are introduced. As summarized in Chapter 2, almost all the publicly available human body model datasets are acquired with expensive and delicate 3D scanners. The user-generated data poses more challenges to human body modelling. Any user can acquire their scans. There is no special training required for the subjects or the scanning operators regarding scanning skills to achieve the best scans. Unlike the high-quality scans, the scanning devices are family-affordable but the accuracy of the devices is rather low, as introduced in Section 2.1. Here, to straightforwardly acquire the user-generated 3D scans for experimental purpose, a scanning platform is built with which the dataset- Kinect based human body dataset (K3D-hub)- is captured. Some examples of K3D-hub dataset are given in Figure 4.3.

4.1.1 Offline 3D Capturing Platforms with Single Hand-held Scanner

In this part, the 3D scanning platform with a single Microsoft Kinect for Xbox 360 is introduced. The setup is shown in Figure 4.1 where the whole scanning platform is composed of the scanning subject, a running turntable and a standing Microsoft Kinect for XBOX 360. The distance from the Kinect to the subject is around 1 meter and the Kinect is connected to a personal computer for reconstructing each scan in real-time. The platform is built upon the ReconstructMe application [1] which is based on Kinect Fusion [109]. During scanning, the Kinect keeps still at three different heights while the subject is standing on a running turntable at a certain speed (30 seconds per round). After scanning one round at the first height, the height of the Kinect is adjusted to the second height and the second round is scanned around the subject. We scan one body part once

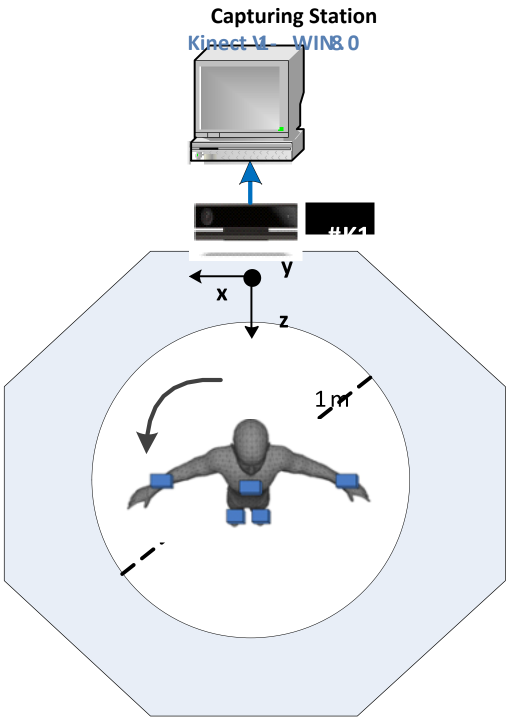


Figure 4.1: The top view of the spatial arrangement of the offline 3D capturing platform.

[1] http://reconstructme.net/

and then move Kinect to the next part. This application is also capable of capturing and processing the colour information of the object being scanned but in experiments of this research, only geometry is considered. For each pose, from our experience, it takes about 90 seconds to build. In the data capturing, the participants are required to wear tight clothes. Each person is captured in 5 poses which include a natural A pose, and other 4 poses (The pose examples are shown in Figure 4.3). The capturing process is displayed in Figure 4.2 where the subject is standing on the running turntable and performing one of the predefined poses. The scanned parts are reconstructed in real time.

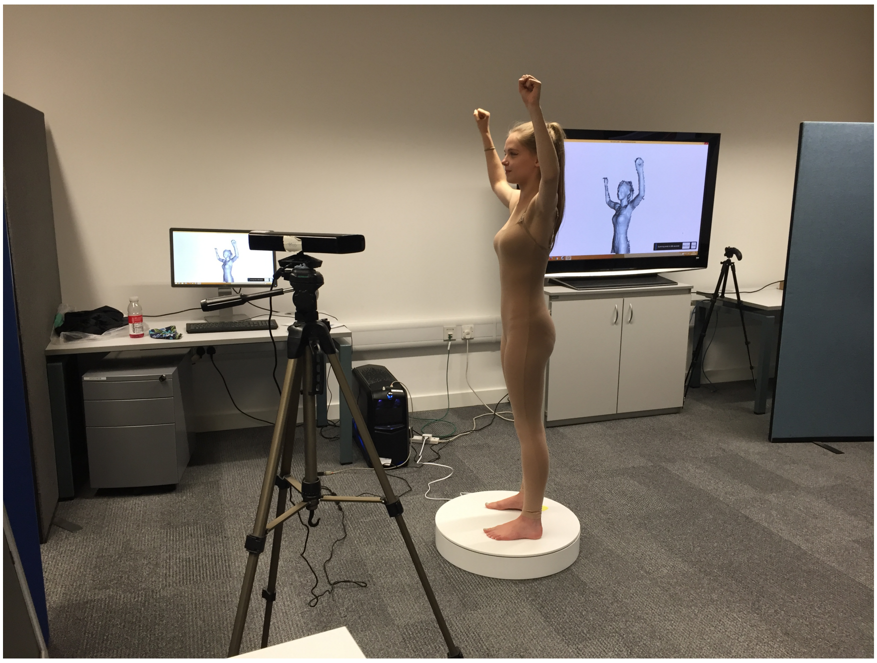


Figure 4.2: The illustration of the capturing process.

Big holes exist on head top and soles of the feet which the kinect cannot “see”. To the best of our knowledge, there is no public Kinect based human body mesh dataset. Therefore, the platform is used to build a low-quality mesh dataset, named **K**inect-based **3D Hu**man **B**ody (K3D-hub) Dataset to facilitate the development of human body modelling from low-quality but real scans. So far, the K3D-hub dataset contains 55 different identities and 5 poses for each person. Some examples are shown in Figure 4.3.

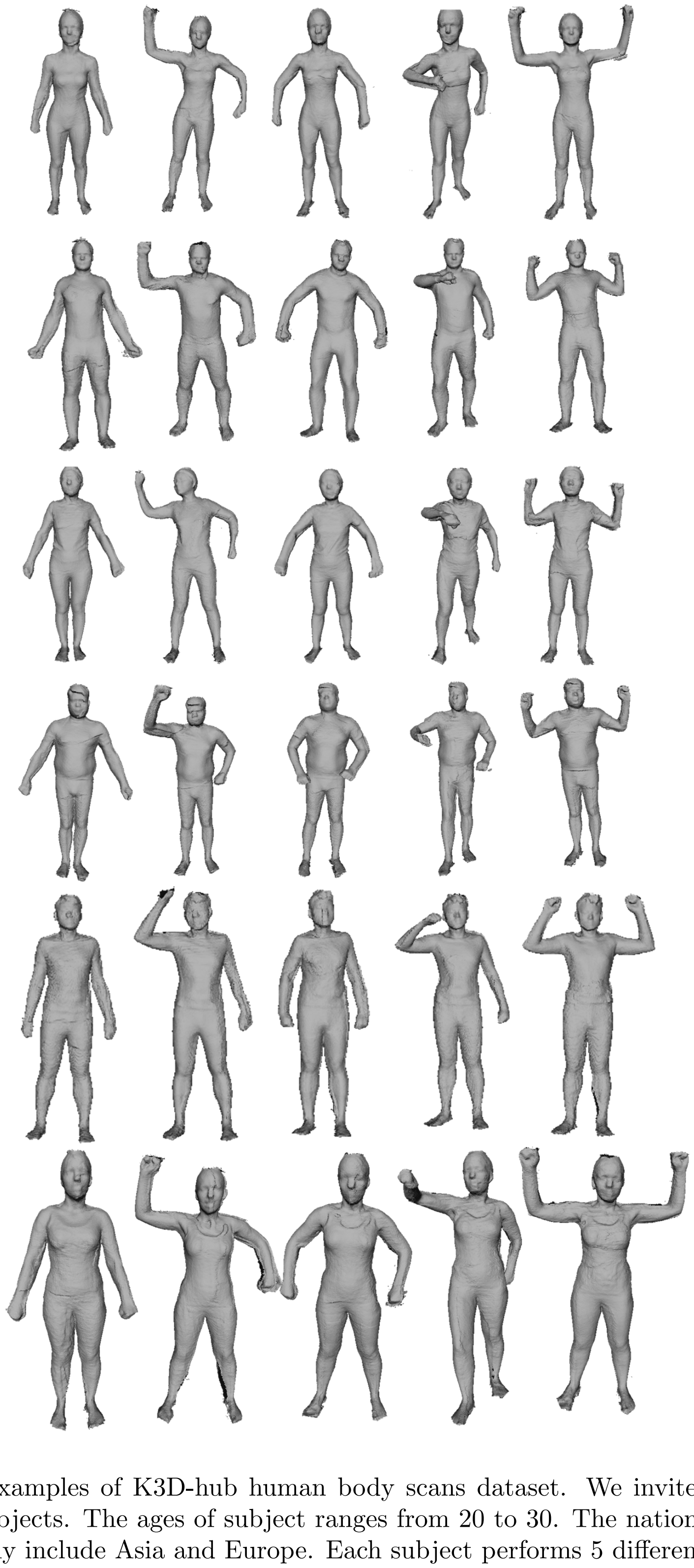


Figure 4.3: Examples of K3D-hub human body scans dataset. We invited both male and female subjects. The ages of subject ranges from 20 to 30. The nationalities of the subjects mainly include Asia and Europe. Each subject performs 5 different poses.

4.2 Multilevel Active Registration for Scans in Same Poses

Based on the user-generated scans acquired with the above platform and setups introduced in the Section 4.1, this part of the research aims at addressing the problem of low-quality human body scan registration in the same pose with the template. To faithfully register the body mesh from Kinect scans, a multilevel active body registration (**MABR**) approach is presented to build a watertight and high fidelity virtual human body in an automatic way. The goal is to align a template mesh to the target point clouds from Kinect scans as closely as possible. Inspired by [32], the statistical shape model is combined with non-rigid iterative closest point algorithm. Two levels of shape models are trained. The holistic shape model describes the whole human body and a set of local shape models are trained to describe the details of each part of the body. Here, a template mesh is the mean shape of the holistic shape model which is learned from an existing human dataset.

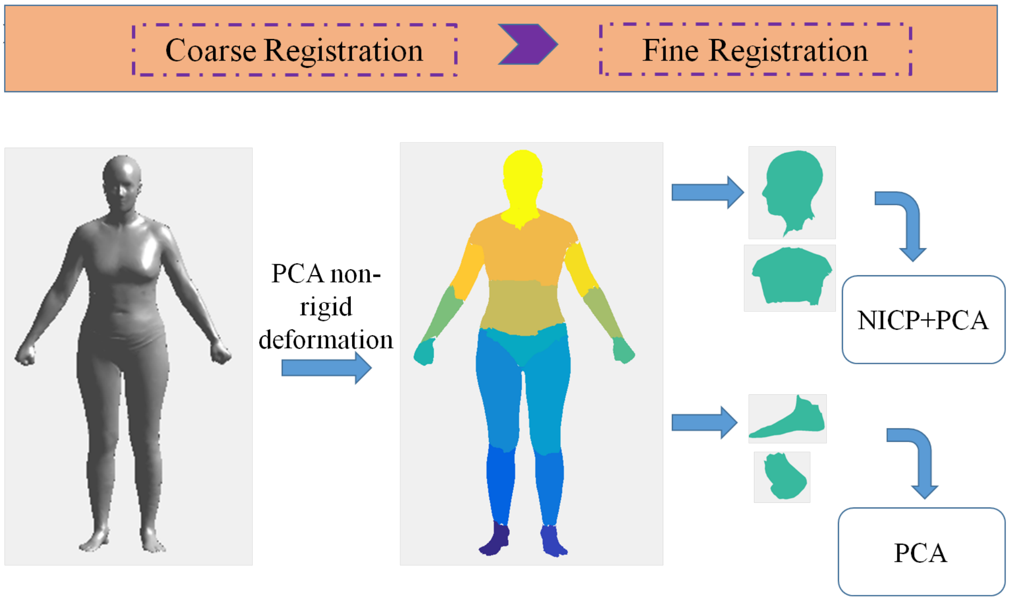


Figure 4.4: The work flow of the proposed multilevel active body registration method (MABR) for scans with same poses.

In MABR, to robustly and accurately register the Kinect scans with our trained shape model, we follow a coarse-to-fine process.

- In the coarse level, the template (the mean shape of the holistic shape model) and target are roughly aligned for the basic description of the human body shape.
- In the fine level, a region-based registration is performed where the template is divided into 12 parts and each part is fitted to the target point clouds separately. For the main body parts where the scan is complete and full of details such as torso, legs and arms, the local affine transformation for each vertex is computed by combining the local shape models with ICP methods.
- In the hand/foot parts, direct application of ICP based fitting method to low-cost, noisy and incomplete Kinect scans could lead to inconsistent and erroneous results. This happens particularly often when it comes to the hands and feet fitting. The main reason is that Kinect scan of the feet can barely be separated from the stand; while during the data capturing, a small movement of hands is inevitable, causing serious artefacts in hand scan. Even if the coarse level registration is performed, the distance of these parts between the source and the target might be large, the nearest neighbours tend to be incorrect and non-rigid ICP easily gets trapped in local minima [72]. Therefore, a PCA based fitting method is proposed that takes special care of foot and hand modelling.

The pipeline of the proposed MABR method is shown in Figure 4.4. As it can be seen in Figure 4.4, in the coarse registration level, the template mesh is deformed non-rigidly into the target, making the template overlap with the target in most parts. In the fine registration level, a region-based deformation is applied to deform the template more accurately. Before nonrigid registration, the rigid transformation is firstly performed to rotate and translate the template, making the template and target meshes share a common coordinate system. Then the MABR registration method is applied to the main body parts and hand/foot parts respectively with the trained morphable shape model.

4.2.1 Related Work

Although shape matching has been deeply researched, finding full correspondences for non-rigid and articulated meshes is still challenging. Geometry information is usually used to extract local features. Histogram of Oriented Normal Vectors [165] and Local Normal Binary Patterns (LNBPs) [147] are descriptors presented based on surface normal. Colour information cannot represent the unique feature in 3D mesh domain and it is usually used as an auxiliary information to other features [15]. Besides using the local geometric features, many works extend the existing 2D features to the 3D domain [37, 156, 130]. 3D-Harris [130] is the 3D extension of the 2D corner detection method with Harris operator. Local depth SIFT (LD-SIFT) [37] is a version of extended image SIFT feature that represents the vicinity of each interest point as a depth map and estimates its dominant angle using the principal component analysis to achieve rotation invariance. MeshSIFT [156] characterizes the salient points neighbourhood with a feature vector consisting of concatenated histograms of shape indices and slant angles. All the feature vectors are used to 3D face shape matching. MeshSIFT is robust to expression variations, missing data and outliers. Clearly, both of these methods rely on local shape features such as curvature or angles. Since they are not pose independent, they cannot be used for shapes undergoing affine transformation, like human shape with different poses.

If two shapes are perfectly isometric, then there exists an isometry i.e., a distance-preserving mapping, between these shapes such that the geodesic distance between any two points on one shape is exactly the same as the geodesic distance between their correspondences on the other [144]. The human body can be regarded as isometric shapes. For example, given two meshes of a human body in two different poses, the geodesic distance from the finger point to the nose point on one mesh should be the same with as the geodesic distance between their counterparts on the other mesh. Different approaches are proposed to exploit the isometry of human body meshes for shape correspondences [45, 75, 143, 102, 40]. One way is to embed shape into a different domain where geodesic distances are replaced by Euclidean distance so that isometric deviation can be measured and optimized in the embedding space [45]. Euclidean embedding can be achieved by using various techniques such as classical MDS (Multidimensional Scaling) [75, 143], least-squares MDS [45], and spectral analysis of the graph Laplacian [102] or of the Laplace-Beltrami operator [40]. Although Isometry-based methods are more accurate than geometry-based ones, they usually require watertight meshes and suffer from self-symmetry of human body shape.

Many approaches are proposed to fit a common template meshes to noisy scans based on the nearest neighbour search. Once fitted, these scans share the common topology of the template and are fully registered. Traditional methods tend to rely on auxiliary modelling tools, such as Maya [2], Blender [3], manual markers and texture information. For automatic registration, the ICP framework is usually taken. Various non-rigid ICP [5, 88, 46, 58, 86] are proposed to register 3D mesh. They usually combine the traditional ICP with regularization terms to make the surface deformation smooth. However, the ICP-based methods are sensitive to missing data and outliers. When they are used in noisy Kinect scans, the hand/foot parts and top of the head are usually distorted severely.

Another way is to infer the body shapes from noisy, incomplete and ambiguous scans using statistical shape models. The statistical modelling of an accurate 3D human body is a fundamental problem for many applications such as mesh deformation, animation, and reconstruction. It can be used to infer the human body models from landmarkers [94], images [17] and videos [1, 79]. However, building a statistical model is also a challenging task. The modelling of human body meshes requires accurately registering a corpus of 3D scans with a common 3D template. Therefore, the robust registration of human body meshes needs the regularization from the statistical model while the building of statistical model requires hundreds of registered meshes. To tackle this chicken-and-egg problem, either synthetic data is used (like TOSCA [22]) or manual assistances (e.g. makers, texture or professional tools) are made, like SCAPE [6], CAESAR [136] and FAUST [15].

[2] http://www.autodesk.co.uk/products/maya/overview
[3] https://www.blender.org/

In addition, the current human body statistical models are built with the close-to-be-naked scans where the subjects usually wear least clothes. Thus, the estimated human body models cannot simulate the human body shape wearing clothes. Even though the subjects wear very thin and tight clothes, the muscles and the tissues will be covered or squeezed to deform. The traditional statistical human body models are hard to describe the accurate body shapes of user-generated scans captured by the applications which are prompted by the appearance of low-cost scanners [109, 168, 36, 40, 119]. Furthermore, human body meshes could be captured for different identities in different poses. Thus, the meshes may be various in shapes, heights and poses. The statistical shape model can only describe the shapes of the human body. Purely shape model based registration is hard to process rigid, articulated and nonrigid transformation robustly. In the below, the registration methods that combined with the statistical model is described in detail.

In order to register the 3D scan, several 3D fitting methods are proposed [125, 15, 207, 40, 2, 5]. The invertible finite volume method [40] is used to control the template tetrahedral mesh to the target point clouds. The stitched puppet model [207] adopts the DPMP algorithm which is a particle-based method to align a graphical model to target meshes. More efforts are made to perform the nonrigid iterative closest point (ICP) [2, 5] which computes the affine transformation at each vertex of the template to allow non-rigid registration of template and scans. Although these ICP based nonrigid registration methods demonstrate high accuracy, it is sensitive to missing data, which might lead to an erroneous fitting result. For Kinect-like scanners, due to self-occluded parts like crotch and armpit, holes and distortion on the mesh are inevitable.

Statistical shape models are employed in non-rigid registration to improve the smoothness and robustness, as the prior knowledge is embedded. SCAPE [6] learns a PCA shape model to describe the body shape variations using 45 instances in a similar pose. It also builds a pose model which is a mapping from posture parameters to the body shape with a dataset that includes 70 poses of one subject. With the learned model, it builds a human body dataset but only the pose dataset is released which contains meshes of 70 different poses of a particular person. The SCAPE model does not consider the differences in shape changes caused by the same pose but different identities. The changes of body shapes for different people performing the same pose vary greatly. For example, considering the same pose of arm lifting, the muscle variations of normal people and the athlete are definitely different. To address this variance, TenBo [31] considers the shape and pose variances in a combination way. Tenbo model is built with the dataset from [61]. The model is used to estimate shape and pose parameters with the depth map and skeleton provided by Microsoft Kinect sensors. The FAUST [15] makes use of the texture information to assist the nonrigid registration of human body meshes and provides a registered dataset that contains 300 scans of 10 people in 30 different poses. The registered mesh has 6890 vertices and 13776 faces. Compared with SCAPE dataset, the resolution is lower but the mesh is still realistic. Nonetheless, its registration method is not fully automatic for the reason that it is based on the texture information which is added by hand. The CAESAR dataset [135] contains 2,400 male and female laser scans with texture information and hand-placed landmarks. Each range scan in the dataset has about 150,000 - 200,000 vertices and 73 markers. Unfortunately, this dataset does not provide correspondences and contains many holes. The MPI [61] captures 114 subjects in a subset of 35 poses using a 3D laser scanner. All the aforementioned methods are working on the scans from expensive scanners, which means the trained statistical model might not be directly applicable on low quality scans from cheap scanners, like Kinect, or cannot achieve satisfactory results. Region-based modelling technique has been prevalent in face [166, 32] and human body [207] modelling, as it allows for a richer shape representation and enables the fitting of different parts to be specifically tailored.

4.2.2 Rigid Registration

The target mesh is captured from Kinect scanners and the template mesh is the mean shape from the public dataset. The rigid transformation is first performed to align these two meshes rigidly without shape deformation. In traditional rigid transformation, correspondences are needed to compute the rigid transformation matrix. Some works use markers to establish correspondences manually. Some 3D mesh features like Heat Kernel Signature [162], are based on surface properties, like geodesic distance, curvature, or face normals. These features work well on public human mesh dataset as they are processed to have the same number of faces and vertices and high-quality without noise or folding faces so that the geometry distance is measurable. However, in the case of locating correspondences between two scans captured from different types of scanners, the number of vertices varies while the template mesh has a fixed number of vertices. Moreover, it is obvious that the physique such as height and muscle properties of the template is different from those of scans. Lastly, noise and holes exist in our target scans. These above-mentioned problems make a full correspondence establishment based on surface measurements infeasible. Therefore, the features which work on the high-quality surface cannot be applied on registering user-generated 3D scans.

Without using correspondences like being presented in Chapter 3, a shape-aware coordinate system is built for each model and transform the source to align its origin and axes with the target. Principal Component Analysis (PCA) is used to identify the most important parts among the vertex set. PCA based alignment is to align the principal directions of the vertex set. Firstly, given the centroid location $c = (c_x, c_y, c_z)$ and vertex i defined by $p_i = (p_{ix}, p_{iy}, p_{iz})$, axes are computed with the matrix in Eq. 4.2 of vertex set in each mesh.

$$\mathbf{P} = \begin{bmatrix} p_{1x} - c_x, p_{2x} - c_x, ..., p_{nx} - c_x \\ p_{1y} - c_y, p_{2y} - c_y, ..., p_{ny} - c_y \\ p_{1z} - c_z, p_{2z} - c_z, ..., p_{nz} - c_z \end{bmatrix}. \tag{4.1}$$

The covariance matrix M is formulated as:

$$\mathbf{M} = \mathbf{P}\mathbf{P}^T. \tag{4.2}$$

The eigenvectors of the covariance matrix represent the principal directions of mass variation. They are orthogonal to each other while the eigenvalues indicate the amount of variation along each eigenvector. Therefore, the eigenvector with the largest eigenvalue indicates the direction where the mesh mass varies the most. In the human body mesh, the principal direction should be along the height direction. The next two directions should be along the width and thickness of the human body respectively. As in this part, we consider that the target and the template have the same pose. Thus in rigid registration, we align them by aligning their three principal directions.

Let **A** and **B** be two matrices where columns are the axes, in order to align these two orthogonal matrices, the rotation **R** is computed such that

$$\begin{aligned} \mathbf{RA} &= \mathbf{B} \\ \mathbf{R} &= \mathbf{B} \setminus \mathbf{A}. \end{aligned} \tag{4.3}$$

The translation vector **t** is computed as:

$$\mathbf{t} = \mathbf{c_T} - \mathbf{c_S}, \tag{4.4}$$

where $\mathbf{c_T}$ and $\mathbf{c_S}$ are the centroids of target and template vertex sets correspondingly. Finally, given the vertex set **P**, the transformed vertex set $\mathbf{P}'$ can be calculated by performing the PCA based alignment as follows:

$$\mathbf{P}' = \mathbf{c_T} + \mathbf{R}(\mathbf{P} - \mathbf{c_S}). \tag{4.5}$$

4.2.3 Multilevel Active Registration for Human Body Scans in A Pose

Morphable Shape Models In this part, the statistical body shape model trained from 200 entire human body scans using PCA technique is introduced. Given a set of training shapes, the statistical shape model can be represented as:

$$\mathbf{v} = \mathbf{Bc} + \mathbf{m}, \tag{4.6}$$

where $\mathbf{v} \in \Re^{4N \times 1}$ are the 3D coordinates (x, y, z) plus corresponding homogeneous coordinates of all N vertices; $\mathbf{B} \in \Re^{4N \times k}$ are the eigenvectors of the PCA model, $\mathbf{m} \in \Re^{4N \times 1}$ is the mean shape, and $\mathbf{c} \in \Re^{k \times 1}$ contains the non-rigid parameters for shape deformation.

Apart from a holistic body shape model, to further describe the large amount of shape variability in human body, the model of each region of the body is built with its own PCA model. In this research, the body segmentation model provided by the SCAPE [6] dataset is employed. Assume that there are p independent parts in the segmented template $\mathcal{V} = \{\mathbf{v}^i\}_{i=1}^{p}$, and the i^{th} part $\mathbf{v}_i$ can also be modelled using Eq. 4.7:

$$\mathbf{v}^i = \mathbf{B}^i \mathbf{c}^i + \mathbf{m}^i. \tag{4.7}$$

Here, $\mathbf{v}^i$, $\mathbf{B}^i$ and $\mathbf{m}^i$ are the shape coordinates, eigen basis and mean shape of the model for the i^{th} region respectively, and $\mathbf{c}^i$ is the latent variable controlling the deformation of the model. As a result, two levels of shape model are trained: the first level is a holistic model for the entire body and the second one is a region-based model that models each body part separately.

Coarse level registration

The main goal of this registration is to overlap the template and target scan, while minor details of body can be ignored in this level. After the rigid transformation, the ICP algorithm with the holistic PCA model are combined to derive the deformed template that would place closer to the target point clouds. Here, with target point clouds $\mathbf{u}$ retrieved by the nearest neighbours search using the k-d tree algorithm which is a widely-recognised efficient way in nearest neighbour search problem, the cost function to be minimized can be formulated as:

$$E(\mathbf{c}) = ||\mathbf{v} - \mathbf{u}||^2 = ||(\mathbf{B}\mathbf{c} + \mathbf{m}) - \mathbf{u}||^2. \tag{4.8}$$

To solve this equation, the partial derivative is taken with regard to $\mathbf{c}$ and achieves the minimum when it approaches to zero:

$$\mathbf{B}^T\mathbf{B}\mathbf{c} + \mathbf{B}^T(\mathbf{m} - \mathbf{u}) = \mathbf{0}, \tag{4.9}$$

and get the closed-form solution,

$$\mathbf{c} = -(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\mathbf{m} - \mathbf{u}). \tag{4.10}$$

Fine Level Registration

After the coarse level registration, to capture the non-rigid nature of body surface and provide a more accurate fitted mesh, the region-based statistical shape model is combined with the Non-rigid Iterative Closest Points (**NICP**) algorithm [5]. Note that during scanning, the subject is unlikely to hold the exact pose like the template, especially in the parts of arms and legs, thus the hand and foot parts could easily appear as outliers. Although the first level fitting alleviates this effect, the original NICP algorithm still might not generate satisfactory fitting results. Therefore, for the parts of hand and foot, only the non-rigid parameters are used to control the deformation in a coarse grained level and a regularization term is added to make it smooth on the boundary. In this way, the hand and foot parts which are impaired in the scanning process can be recovered. The clear and semantically correct hands and feet allow for the shape statistical modelling in the next stage.

Main Body Registration In this research, the body parts that exclude feet and hands are regarded as *main body*. For the *main body* parts, the statistical shape model is combined with the NICP algorithm. The goal is to find a set of affine matrices $X = \{\mathbf{X}^i\}_{i=1}^p$ and non-rigid parameters $C = \{\mathbf{c}^i\}_{i=1}^p$ such that the sum of Euclidean distances between pair of points of each region is minimal. Here, $\mathbf{X}^i$ is a $3 \times 4n_i$ matrix that consists of the affine matrix for every template vertex in the i^{th} part where n is the number of vertices. Figure 4.5 describes the developed technique for fitting a template S to target mesh T. Each of these surfaces is represented as a triangle mesh. Each vertex v_i is influenced by a 3×4 affine matrix $\mathbf{X}_i$ and non-rigid parameter $\mathbf{C}_i$. The data error, indicated by the arrows in Figure 4.5, is a weighted sum of the squared distances between template surface S' and target surface T. Besides data error, in order to deform the template smoothly, a stiffness term is defined to constrain the vertices not to move directly towards the target, but to move parallel along it. These error terms are summarized in Figure 4.5 and described in details in the following.

Distance term: The distance term is used to minimize the Euclidean distances between the source and the target. It is assumed that each i_{th} part has n_i points and the cost function is denoted as the sum of error of each pair of vertices:

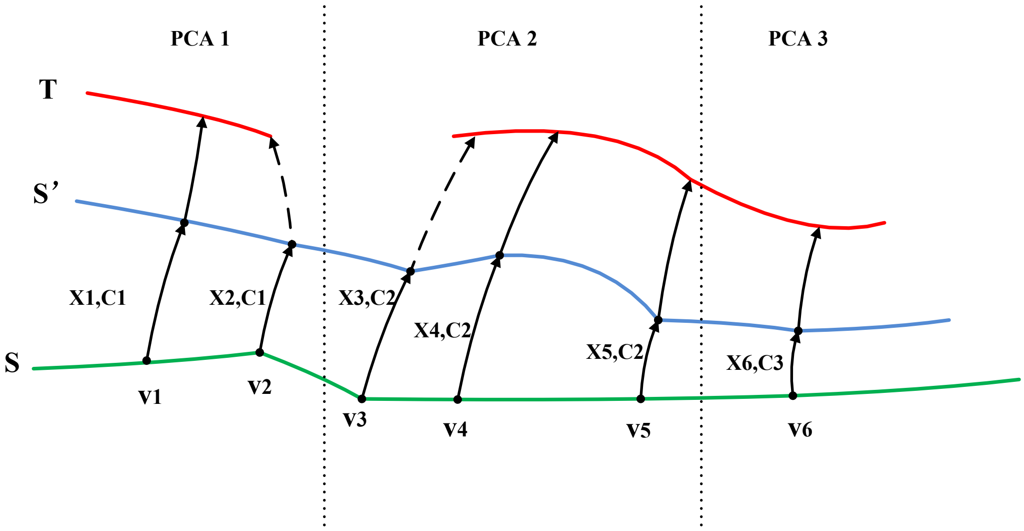


Figure 4.5: The summary of the matching framework. Our target is to find a set of affine transformations $\mathbf{X}_i$ and local PCA parameters $\mathbf{C}_i$, that, when applied to the vertices v_i of the template mesh S, result in a new surface S' that matches the target surface T. This diagram shows the match in progress; S' is moving towards the target but has not reached it. The whole vertices are divided into three parts which are controlled by three local PCAs. The transformation of each vertex is controlled by affine transformation as well as the local parameters of the part which it belongs to.

$$E_d(\mathbf{X}) = \sum_{i=1}^{p}\sum_{j=1}^{n_i} ||\mathbf{X}_j^i \mathbf{v}_j^i - \mathbf{u}_j^i||_F^2, \tag{4.11}$$

where X_j^i is the transformation matrix for vertex j in the part i.

Since each part is modelled by the shape model $\mathbf{v}_j^i = \mathbf{B}_j^i \mathbf{c}_j^i + \mathbf{m}_j^i$, based on Eq.4.7, the distance term could be rewritten and rearranged as:

$$\begin{aligned} E_d(\mathbf{X}) &= \sum_{i=1}^{p}\sum_{j=1}^{n_i} ||\mathbf{X}_j^i(\mathbf{B}_j^i \mathbf{c}_j^i + \mathbf{m}_j^i) - \mathbf{u}_j^i||_F^2 \\ &= \sum_{i=1}^{p} \left\| \begin{bmatrix} \mathbf{X}_1^i & & \\ & \ddots & \\ & & \mathbf{X}_{ni}^i \end{bmatrix} \begin{bmatrix} \hat{\mathbf{v}}^{\mathbf{i}}_{\mathbf{1}} \\ \vdots \\ \hat{\mathbf{v}}^{\mathbf{i}}_{\mathbf{ni}} \end{bmatrix} - \begin{bmatrix} \mathbf{u}_1^i \\ \vdots \\ \mathbf{u}_{ni}^i \end{bmatrix} \right\|_F^2 . \end{aligned} \tag{4.12}$$

It can be seen that the above equation is not in the standard linear form. In order to differentiate, the position of the unknown $\mathbf{X}$ and $\mathbf{V} = [\hat{\mathbf{v}}^{\mathbf{i}}_{\mathbf{1}}, ..., \hat{\mathbf{v}}^{\mathbf{i}}_{\mathbf{n_i}}]^T$ are swapped to obtain the following form.

$$E_d(\mathbf{X}) = \sum_{i=1}^{p} ||\mathbf{D}^i \mathbf{X}^i - \mathbf{U}^i||_F^2, \tag{4.13}$$

where the term $\mathbf{D}^i = diag({\mathbf{v}_1^i}^T, {\mathbf{v}_2^i}^T, ..., {\mathbf{v}_{n_i}^i}^T)$, and the set of closest points $\mathbf{U}^i = [\mathbf{u}_1^i, \mathbf{u}_2^i, ..., \mathbf{u}_{n_i}^i]^T$.

Stiffness term: The stiffness term penalizes the difference between the transformation matrices of neighbouring vertices. Similar to the method described in [5], it is defined as:

$$E_s(\mathbf{X}) = \sum_{i=1}^{p} ||(\mathbf{M}^i \otimes \mathbf{G}^i)\mathbf{X}^i||_F^2. \tag{4.14}$$

Here, for the i^{th} body part, the weighting matrix $\mathbf{G}^i = diag(1, 1, 1, \gamma^i)$, where γ^i is used to balance the scale of rotational and skew factor against the translational factor. It depends on the units of the data and the deformation type to be expressed. $\mathbf{M}^i$ is the node-arc incidence matrix of the template mesh topology [5] This matrix is defined for directed graphs. It contains one row for each arc (edge) of the graph and one column per node (vertex). To construct a node-arc incidence matrix from the source topology, the edges and vertices of the mesh are numbered and its edges are directed from the lower numbered vertex to the higher numbered. If edge r connects the vertices (v_i, v_j) the nonzero entries of M in row r are $\mathbf{M}_{ri} = 1$ and $\mathbf{M}_{rj} = 1$.

Complete cost function for Main Body: Eq. 4.13 and Eq. 4.14 are combined to obtain the complete cost function:

$$\begin{aligned} E(\mathbf{X}) &= E_d(\mathbf{X}) + E_s(\mathbf{X}) \\ &= \sum_{i=1}^{p} ||\mathbf{D}^i\mathbf{X}^i - \mathbf{U}^i||_F^2 + \sum_{i=1}^{p} ||(\mathbf{M}^i \otimes \mathbf{G}^i)\mathbf{X}^i||_F^2 \\ &= \sum_{i=1}^{p} \left\| \begin{bmatrix} \mathbf{D}^i \\ \mathbf{M}^i \otimes \mathbf{G}^i \end{bmatrix} \mathbf{X}^i - \begin{bmatrix} \mathbf{U}^i \\ \mathbf{0} \end{bmatrix} \right\|_F^2 \\ &= \sum_{i=1}^{p} \left\| \mathbf{A}^i\mathbf{X}^i - \mathbf{U}^{*i} \right\|_F^2 . \end{aligned} \tag{4.15}$$

Eq. 4.15 is not a quadratic function and it is difficult to obtain the optimal local affine transformation $\mathbf{X}^i$ and non-rigid parameters $\mathbf{c}^i$ simultaneously. Similar to [32], an alternating optimization scheme is employed to find the optimal set of parameters.

Given the initilization $\mathbf{c_0} = \mathbf{0}$ for the non-rigid parameter $\mathbf{c}^i$, the $E_d(\mathbf{X})$ and $E_s(\mathbf{X})$ are combined into the standard form of least square problem in Eq. 4.15. To obtain the minimun of this objective, we take the derivative with respect to $\mathbf{X}$ and let the derivative be zero.

$$[\partial E(\mathbf{X})/\mathbf{X}^1; \partial E(\mathbf{X})/\mathbf{X}^2; ...; \partial E(\mathbf{X})/\mathbf{X}^p] = \mathbf{0}. \tag{4.16}$$

For the local affine transformation $\mathbf{X}^i$ for part i of the body, we have:

$$E(\mathbf{X}^i) = ||\mathbf{A}^i\mathbf{X}^i - \mathbf{U}^{*i}||^2. \tag{4.17}$$

To solve this least square problem, the derivative with respective to $\mathbf{X}^i$ should be zero as below:

$$\partial E(\mathbf{X}^i)/\partial \mathbf{X}^i = \partial(||\mathbf{A}^i\mathbf{X}^i - \mathbf{U}^{*i}||^2)/\partial \mathbf{X}^i = 0. \tag{4.18}$$

With the matrix differentiation [11], Eq. 4.18 can have a closed-form solution:

$$\mathbf{X}^i = ((\mathbf{A}^i)^T\mathbf{A}^i)^{-1}(\mathbf{A}^i)^T\mathbf{U}^{*i}. \tag{4.19}$$

We represent $((\mathbf{A}^i)^T\mathbf{A}^i)^{-1}$ with $\mathbf{A}$. To efficiently acquire the inverse of $\mathbf{A}$, we can firstly perform the SVD decomposition on the matrix $\mathbf{A}$, obtaining:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Sigma}\mathbf{V}^T, \tag{4.20}$$

where $\mathbf{Q}$ and $\mathbf{V}$ are the orthogonal matrix and the inverse equals to the transpose; $\mathbf{\Sigma}$ is the diagonal matrix with the elements being singular values of matrix $\mathbf{A}$. Accordingly, the inverse of $\mathbf{A}$ can be efficiently acquired as below.

$$inv(\mathbf{A}) = \mathbf{V}\mathbf{\Sigma^{-1}}\mathbf{Q^T}, \tag{4.21}$$

in which $\mathbf{\Sigma^{-1}}$ can be easily computed by replacing each element in the diagonal with its reciprocal.

After $\mathbf{X}^i$ is obtained, we then optimize $\mathbf{c}^i$. The partial derivative is taken to be zero. Similarly, the optimal $\mathbf{c}^i$ is obtained using:

$$\mathbf{c}^i = ((\mathbf{X}^i\mathbf{B}^i)^T\mathbf{X}^i\mathbf{B}^i)^{-1}(\mathbf{X}^i\mathbf{B}^i)^T(\mathbf{X}^i\mathbf{m}^i - \mathbf{u}^i). \tag{4.22}$$

This iteration will end when the error of two rounds of affine transformations is smaller than a threshold. In the alternative way, the affine transformation and the non-rigid parameters for each part of the main body are acquired.

For better understanding, the probabilistic interpretation of Eq. 4.15 is illustraed as follows. According to the Bayes' Rule, given the observations $\mathbf{O}$ (in this context, the target human body meshes are regarded as observations $\mathbf{O}$), the possible transformation matrix $\mathbf{X}$ can be regarded as the parameters that maximize the posterior possibilities:

$$\begin{aligned}\mathbf{X} &= \underset{\mathbf{X}}{\operatorname{argmax}}\ p(\mathbf{X}|\mathbf{O}) \\ &= \underset{\mathbf{X}}{\operatorname{argmax}}\ \frac{p(\mathbf{O}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{O})} \\ &= \underset{\mathbf{X}}{\operatorname{argmax}}\ p(\mathbf{O}|\mathbf{X})p(\mathbf{X}),\end{aligned} \tag{4.23}$$

where $p(\mathbf{O}|\mathbf{X})$ is the possibilities of $\mathbf{O}$ given the parameter $\mathbf{X}$; $p(\mathbf{O})$ is a constant; and $p(\mathbf{X})$ is the prior distribution.

Taking the negative logarithm of $p(\mathbf{O}|\mathbf{X})p(\mathbf{X})$, the Eq. 4.23 can be rewritten as:

$$\mathbf{X} = \underset{\mathbf{X}}{\operatorname{argmin}}\ [-\log p(\mathbf{O}|\mathbf{X}) - \log p(\mathbf{X})] \tag{4.24}$$

The prior $p(\mathbf{X}) \sim e^{-\frac{\bar{\mathbf{X}}}{2\boldsymbol{\delta}^2}}$, thus, we can have:

$$\begin{aligned}\mathbf{X} &= \underset{\mathbf{X}}{\operatorname{argmin}}\ [-\log p(\mathbf{O}|\mathbf{X}) + \frac{1}{\boldsymbol{\delta}^2}(\mathbf{X} - \bar{\mathbf{X}})^2] \\ &= \underset{\mathbf{X}}{\operatorname{argmin}}\ [-\log p(\mathbf{O}|\mathbf{X}) + \frac{1}{\boldsymbol{\delta}^2}\mathbf{X}^2],\end{aligned} \tag{4.25}$$

where $\boldsymbol{\delta}$ is the standard deviation of the prior distribution. Here, we set $\bar{\mathbf{X}} = \mathbf{0}$ for simplicity. As we can see, the part of $-\log p(\mathbf{O}|\mathbf{X})$ in Eq. 4.25 can be regarded as the data energy that measures the negative of likelihood that the data $\mathbf{O}$ is observed given the transformation parameter $\mathbf{X}$ and the second part of $\mathbf{X}^2$ is the prior energy.

Hands & Feet Registration Although the main body parts are roughly aligned after the first level registration, the distance between the source hands/feet and corresponding target is large in most cases. In this situation, the ICP-based methods can easily be trapped in local minima [72]. To tackle this problem, a PCA-based fitting is performed for the individual part of hand/foot. Given one particular part model of a hand/foot that has eigenbases $\mathbf{B}^*$ and mean shape $\mathbf{m}^*$, the objective function that consists of a distance term and a regularization term is defined, and the optimal non-rigid parameters $\mathbf{c}^*$ is obtained by minimizing the objective function.

Distance term: It is defined similarly to Eq. 4.36, but without the affine transformation matrix,

$$E_d(\mathbf{c}^*) = ||(\mathbf{B}^*\mathbf{c}^* + \mathbf{m}^*) - \mathbf{u}^*||_F^2. \tag{4.26}$$

Boundary smoothness term: In order to stitch hand/foot with its neighbouring part smoothly, a boundary smoothness term is defined as follows:

$$E_b(\mathbf{c}^*) = ||\mathbf{S}^*(\mathbf{B}^*\mathbf{c}^* + \mathbf{m}^*) - \mathbf{F}^*||_F^2, \tag{4.27}$$

where $\mathbf{S}^*$ is the selection matrix of hand/foot parts that picks out the boundary points. $\mathbf{F}^*$ is the boundary points of the neighbouring part. Enforcing the boundary constraints between the two parts regulates the part fitting process to avoid erroneous result caused by outliers.

Complete cost function for Hands/Feet Part: The fitting objective function can be formulated as:

$$\begin{aligned} E(\mathbf{c}^*) &= \alpha E_d(\mathbf{c}^*) + (1-\alpha)E_b(\mathbf{c}^*) \\ &= \left\| \begin{bmatrix} \alpha\mathbf{I} \\ (1-\alpha)\mathbf{S}^* \end{bmatrix} (\mathbf{B}^*\mathbf{c}^* + \mathbf{m}^*) - \begin{bmatrix} \alpha\mathbf{u}^* \\ (1-\alpha)\mathbf{F}^* \end{bmatrix} \right\|_F^2 \\ &= \left\| \mathbf{A}^*(\mathbf{B}^*\mathbf{c}^* + \mathbf{m}^*) - \mathbf{U}^* \right\|_F^2, \end{aligned} \tag{4.28}$$

where α is the weighting factor between the two terms, $\mathbf{A}^* = [\alpha\mathbf{I}, (1-\alpha)\mathbf{S}^*]^T$ and $\mathbf{U}^* = [\alpha\mathbf{u}^*, (1-\alpha)\mathbf{F}^*]^T$. Similar to the optimization of Eq.4.17, the minimum occurs where the gradient vanishes, that is $\partial E_{\mathbf{c}^*}/\partial\mathbf{c}^* = \mathbf{0}$. Thus Eq. 4.43 has closed-form solution:

$$\mathbf{c}^* = -[(\mathbf{A}^*\mathbf{B}^*)^T(\mathbf{A}^*\mathbf{B}^*)]^{-1}(\mathbf{A}^*\mathbf{B}^*)^T(\mathbf{A}^*\mathbf{m}^* - \mathbf{U}^*). \tag{4.29}$$

The proposed fitting method for hands and feet has a nice convergence property, demonstrated by an example of residual error curve for all iterations of fitting in Figure 4.12.

In this part, we describe a robust registration approach to align a template mesh to a target scan in A pose. Compared with the classic NICP proposed in [5], the proposed method merges a body shape model trained with PCA from a set of body meshes during optimization. The shape model can regularize the movements of the template vertices so that they retrieve the reasonable closest target points. In this way, the shape model effectively helps address the local minima problem when the situation of missing data and bad initialization occurs. We also take special care of the hand and foot parts to greatly improved the robustness.

4.2.4 Performance Evaluation

To evaluate the performance of the MABR method, experiments are conducted on both high and low quality meshes, and showed the shape root mean square (RMS) error curve as well as some fitting results for visualization purpose.

High Quality Mesh Evaluation For the evaluation on good quality data, the SPRING [189] dataset is used that contains 3038 meshes with various human body shapes. These good quality meshes are complete and points are evenly distributed. Furthermore, it has point-to-point correspondences with the SCAPE model, which means the segmentation of the SCAPE model can be directly used on the SPRING data. It is also used as the ground truth in this research. Also, the SPRING dataset is divided into male and female subsets. In order to train a model whose muscle and tissue properties are gender specific, male and female shape models are separately trained. For each gender, 200 meshes from SPRING dataset are used as the training set and the remaining meshes are regarded as the testing set.

To show the superior performance of the proposed method, the MABR method is compared with **NICP** in [5], **ANICP** in [32], and **PCA** deformation on SPRING dataset. The 3D shape root mean square error (RMS Error) is computed with the Eq. 4.45 to measure the accuracy of four methods.

$$RMSError = \sqrt{\frac{\sum_{i=1}^{n}(p_i - \hat{p_i})^2}{n}}, \tag{4.30}$$

where points p_i and $\hat{p_i}$ are corresponding points of the ground truth and the fitted results. n is the number of the points in 3D template mesh. As we can see, RMS Error shows how much the predicted body shapes vary around the ground-truth body shapes . It is widely-recognised effective matrix to evaluate the accuracy of the model in the problem of registration [188, 203]. As shown in Figure 4.6, the accuracy of the proposed MABR is comparable with ANICP and is much higher than NICP and PCA. In the PCA method, the whole model is only controlled by the trained orthogonal basis which cannot cover all the shape variations. Consequently, the accuracy of PCA is the lowest.

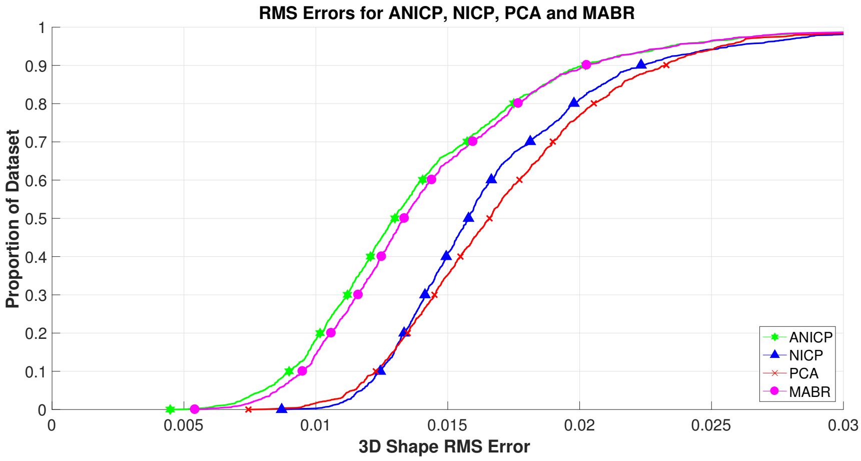


Figure 4.6: The comparison of 3D shape RMS error of ANICP, NICP, PCA and the proposed MABR.

In order to show the error distribution over the whole body, we compare the average error of each body parts among MABR, ANICP, NICP and PCA in Figure 4.7. A pink line is added for better visualization. It is obvious to see that the foot and hand parts in MABR are much better than the other three methods. These parts are easy to get trapped into the local minima so that the nearest neighbours are searched erroneous. The combination of the PCA model reduces the accuracy of body registration, especially where good initialization is obtained, like $(8)-(12)$ parts. This can be seen by comparing the MABR, PCA against NICP and ANICP with respective to the average errors of parts of $(8)-(12)$. Therefore, the proposed methods are better suitable for the registration of user-generated human body meshes which are captured by Kinect-like scanners and full of flaws.

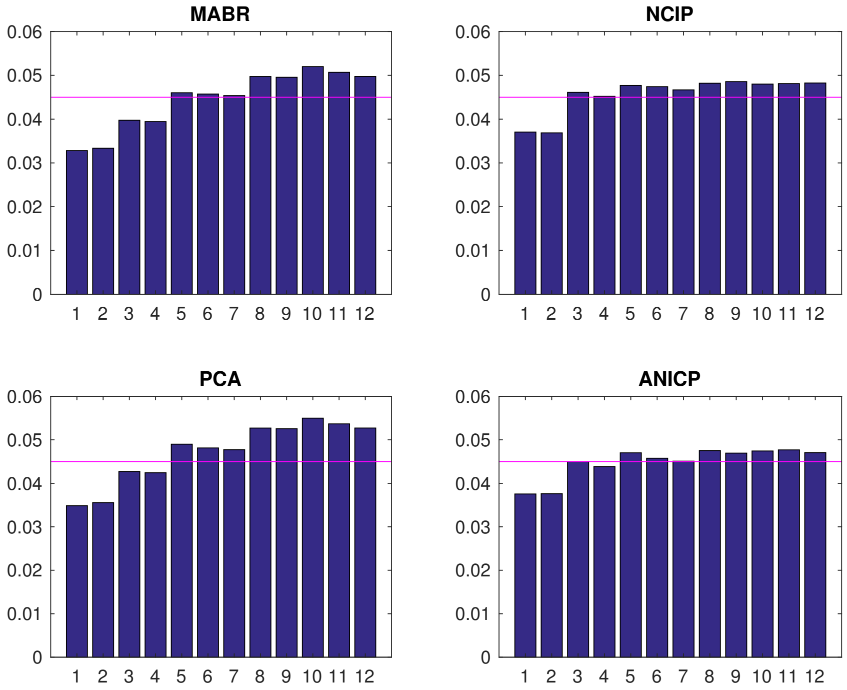


Figure 4.7: The average error distribution on meshes over SPRING dataset. In each subfigure, from left to right, each bar represents the error of: (1) right foot, (2) left foot, (3) right thigh, (4) left thigh, (5) bottom, (6) right forearm, (7) left forearm, (8) torso, (9)breast, (10) right upper arm, (11) left upper arm and (12) head.

Local minima problem happens often in the case that the shape of the template is largely different from the target shape and the nearest neighbour is searched by fault. Figure 4.8 shows that MABR is able to present better fitting results when the local minima happen. The target meshes differs a lot from the template with respect to the arm parts, and it can be seen that neither ANICP nor NICP can obtain a complete arm for the given target while the proposed MABR method can achieve a meaningful arm. This is because ICP-based methods, ANICP and NICP, heavily reply on good initialization. This figure verifies the robustness of the proposed method that is able to alleviate the local minima problem.

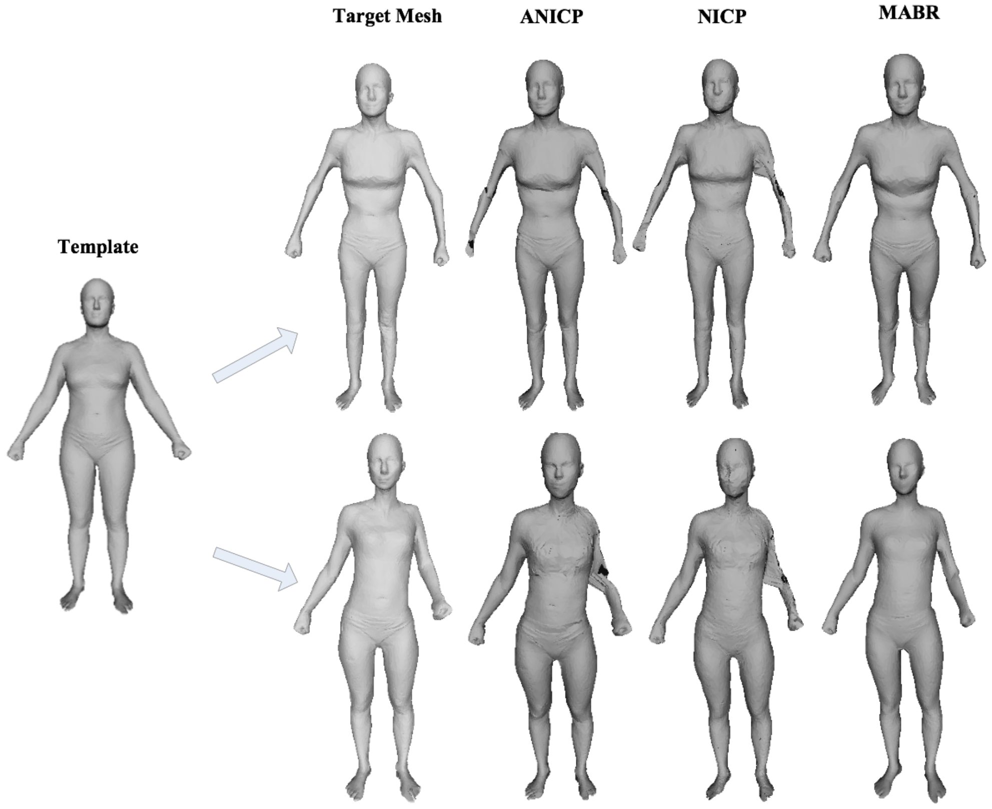


Figure 4.8: The front view of fitted results of ANICP, NICP and MABR in the case that the shape of general template differs a lot from the target mesh.

The details of fitted results from the above four methods are compared in Figure 4.9. It can be seen that for the hand and elbow parts, the PCA method and the proposed MABR method are much better than the other two approaches. In Figure 4.9, compared with PCA and MABR, the hand parts of ANICP distorted severely and the fitted hand of NICP is obscure. As for the elbow, the results of ANICP and NICP are broken while

PCA and MABR are able to preserve the continuity of the fitted mesh. Although PCA can get meaningful results, MABR outperforms it in terms of accuracy, which is reflected by the fitted results in face parts. It can be seen obviously that the face of MABR is much more similar with the raw scan than the face of PCA. MABR successfully recovers the shape of the target mesh. In the elbow part, it is obvious that the curvature of the mesh from MABR is much closer to the target than PCA's result.

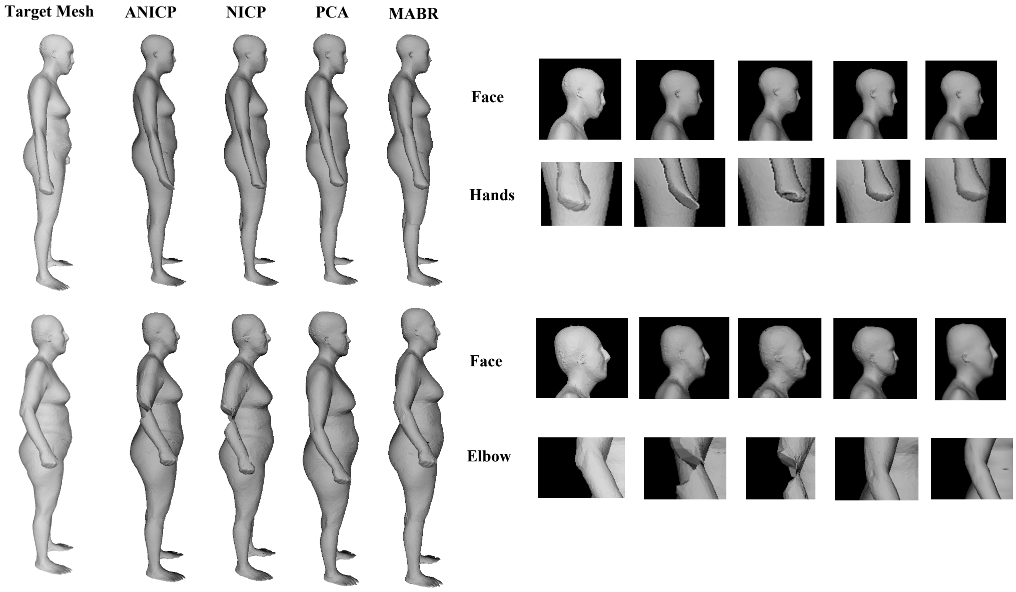


Figure 4.9: The detail comparison of fitting results on SPRING dataset. The side of the raw scan and the fitted results of ANICP (column 2), NICP (column 3), PCA (column 4) and MABR (column 5). Besides the comparison of the full body, the details of the face, hand and elbow from each method are also compared successively.

Low Quality Scans Evaluation The proposed method is evaluated on low quality scans which are captured by Microsoft Kinect for XBOX 360 following the procedure described in the subsection 4.1. The scans are pre-processed to remove background. The proposed **MABR** method is compared with **NICP** [5] and **ANICP** in [32].

Fitting results of these three methods are shown in Figure 4.10. It is obvious to see that the proposed MABR method is the only one that models the hand and foot parts completely and, meanwhile, keeps high accuracy of the fitting results. It is obvious to see that the raw sans have a lot of noise which is close to the surface. Large holes exist on top of the head. All these challenges require the registration method to be robust to noise, outliers and holes at the same time. From the results, it can be seen that neither ANICP nor NICP is robust enough to obtain complete and accurate registered mesh. The hand parts of ANICP and NICP tend to be distorted and incomplete while the MABR method enables meaningful and complete hands. The reason for the failure of NICP and ANICP is that, in real scans, the human pose is hard to control and consequently the limbs are usually not completely overlapped with the template.

Therefore, the shape of the closest points of the limbs cannot keep the limb shape of scans, resulting in unexpected fitted shapes. Since this fitting procedures are active, the limb parts of the template can be stretched along with the direction of PCA basis before performing non-rigid ICP, recovering the size of the hand and foot roughly. In this way, the MABR method is able to keep a good shape of the scan and also to be robust to noises. The robustness to holes of MABR is also shown in Figure 4.11. Even though there exist big holes on top of the head in the raw scan, MABR and ANICP can fill the hole smoothly, benefiting from the training with the prior knowledge. NICP merely relies on finding the nearest points on the target, which is sensitive to holes. Therefore, as illustrated in Figure 4.11, the fitted result of NICP is unsmooth.

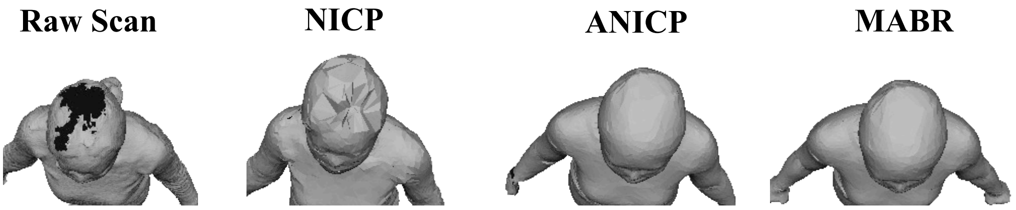


Figure 4.11: The comparison of hole tolerance.

In addition, the proposed MABR method shows superior convergence properties. Figure 4.12 shows one example of the residual error changes as the fitting of left hand progresses. As it can be seen, the residual error monotonically decreases and gradually converges to a minimum value.

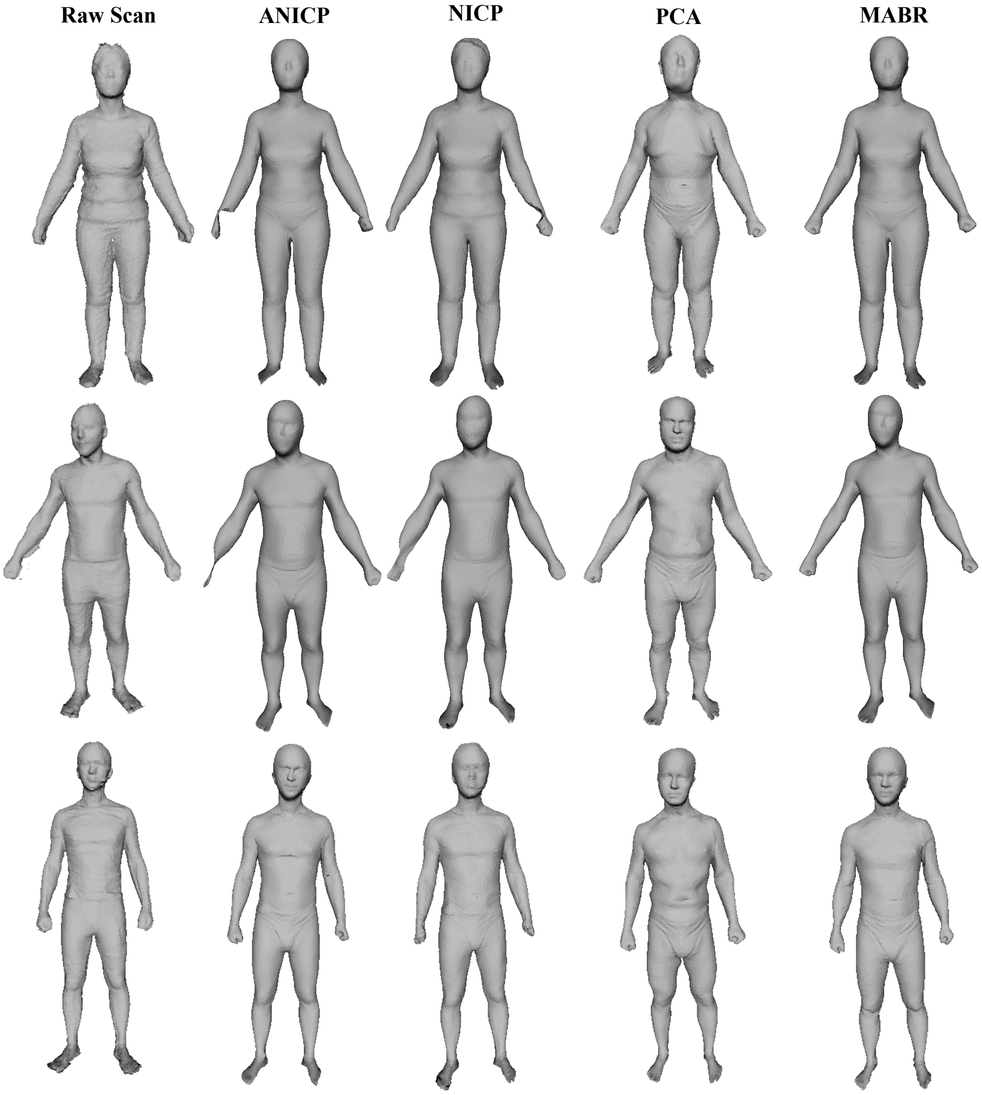


Figure 4.10: Fitting results on Kinect scans. Column 1 shows the raw body scans, the second to the last columns illustrate the shapes from NICP, ANICP and the proposed MABR method respectively.

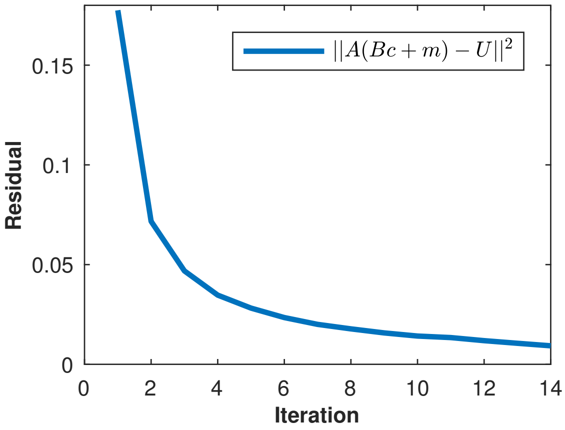


Figure 4.12: The example of residual error changing as the fitting of left hand in the second level progresses.

4.3 Landmark based MABR for Scans in Various Poses

The last section addresses the registration problem of user-generated scans when the template and the target scans have similar poses. However, the human body presents a large range of poses. Therefore, a problem arises on how to tackle the registration problem when the template and the target scans present two different poses, i.e, the template is in A pose while the target scan presents other poses. To this end, based on the approach proposed in the last section, a landmark based registration approach is proposed in this section.

Recent model-based methods [128, 18, 95] have been developed for human body modelling. In [94], 67 markers are used to constrain pose changing. Human body models, like SMPL and SCAPE, are trained with high quality human body meshes. The scanned subjects are close to naked to allow the scanners to capture the shape of muscle and soft tissue. Thus, the above models are designed with the constraint to describe the nearly naked human body shapes and poses only. However, in the everyday scanning for the ordinary people which happens in a much broader range of applications such as online shopping and virtual try-on, the requirement for close-to-naked scans of models is hard to be met. Even though subjects are asked to wear tight clothes, some muscle and soft tissue details are squeezed to change or shielded by the clothes. Therefore, pure model based registration can hardly describe the human body shapes captured by users in routine life.

Here, the problem of human body registration from user-generated scans in different poses is addressed. In this scenario, it is hard to acquire close-to-naked body scans. Instead, it is assumed that scans with tight clothes can be obtained. Human body with loose clothes is not considered as more body features are covered.

To address this problem, a region based human body registration approach is proposed that builds a watertight and high fidelity virtual human body in an accurate and robust way. Firstly, sparse landmarks are used to pose the template to perform in the same way with target scans. During mesh registration, a statistical body shape prior is learned from the publicly available dataset and combine it with the Iterative Closest Point registration method. In this method, two groups of statistical template models are trained: a holistic shape model for the whole human body and a group of local shape models for describing each body part to improve the quality in some body parts that are more error-prone than the whole body model. In the first level of registration, the holistic template and target are roughly aligned based on the pre-trained holistic PCA shape model. In the second level, a region-based registration is applied with a combination of NICP and local PCA models. Extensive experiments on data scanned using devices from professional to low-cost types verify that our approach is both accurate and robust to incomplete and noisy data.

4.3.1 Landmark-based MABR for Scans in Different Poses

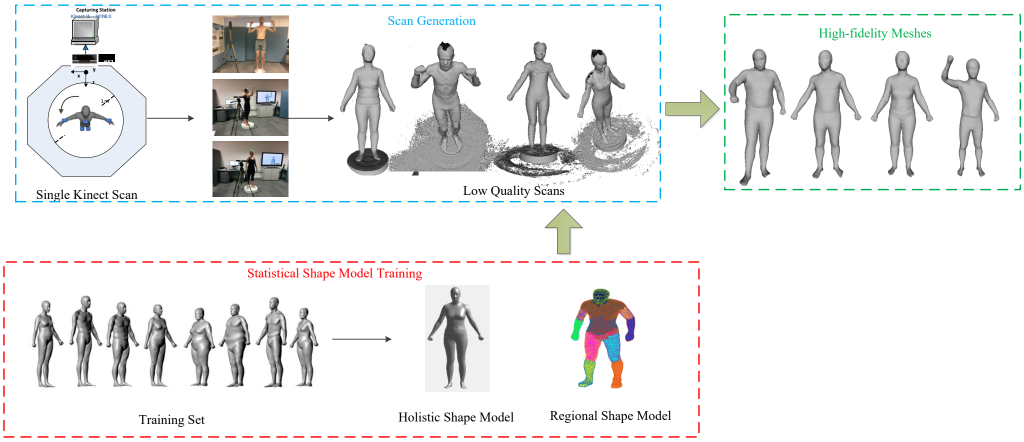


Figure 4.13: The pipeline of the proposed method. First of all, two statistical models are trained from a dataset of human body meshes. The holistic statistical shape model is trained using all the vertices of the body. The local statistical shape model is composed of a collection of PCA models which are trained with vertices from 12 regions of the body. Secondly, the trained models can be used to register the low quality scans which are obtained from our single Kinect scanning platform. Finally, watertight, smooth and re-topologised meshes are acquired.

The pipeline of the region based body registration method is shown in Figure 4.13.

The method involves three main stages: The statistical shape models for both the holistic and the regional are trained; the template is reposed to perform a similar pose with the target; finally the registration is employed to acquire the closed and high-fidelity meshes. The input is the Kinect scans of ordinary people wearing tight clothes. The goal is to register these scans with a common template which is trained from publicly available datasets.

4.3.2 Pose template

The shape model only describes the body shape variation. Here, the proposed method firstly deforms the pose of the template before performing registration. In the tradition, object deformation with linear blending scheme (LBS) is regarded as the fastest approach for deforming geometric models by transforming each point on the object by a linear combination of several affine transformations with different weights. However, it is designed for skeletons and is usually required to manually assign the influence of the skeleton to the surface points. Another issue is that, as shown in Figure 4.14, linearly blending rotation causes well-known artefacts: the loss of volume in bent areas and a "candy wrapper" effect happens when skeleton joints are twisted.

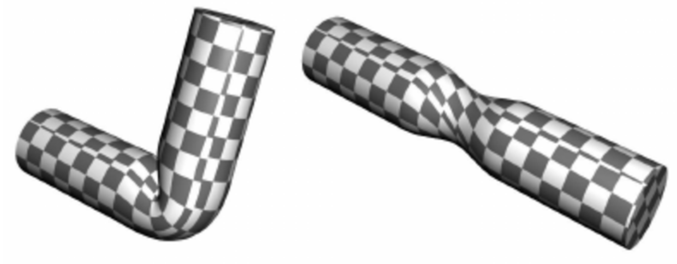


Figure 4.14: The loss of volume in the bent area and the "candy wrapper" effect when the joints are twisted.

Our goal is to perform smooth deformation of the template to have the similar poses with the target scans. To ensure that the affine transformations of each control point (i.e. annotated landmarks) are blended to produce intuitive deformation, the bounded biharmonic weight (BBW) [74] is computed. In this case, the weights w_j are defined as minimizers of the Laplacian energy. The computed weights allow for smooth and intuitive deformation. The weights are computed in the same way as defined in [74].

$$\begin{aligned} & argmin_{w_j, j=1,...,m} \sum_{j=1}^{m} \frac{1}{2} \int_{\Omega} ||\Delta w_j||^2 dV \\ & subject \quad to : w_j|_{H_k} = \delta_{jk} \\ & \qquad w_j|_F \quad is \quad linear \\ & \qquad \sum_{j=1}^{m} w_j(p) = 1 \qquad \forall p \in \Omega \\ & \qquad 0 \leqslant w_k(p) \leqslant 1, j = 1, ..., m, \forall p \in \Omega, \end{aligned} \tag{4.31}$$

where δ_{jk} is the Kronecker delta which is 1 when the $j = k$, and 0 otherwise; Ω denotes the human body surface. These constraints promise the locality of the weights. H_k is the k_{th} controller. With the E.q. 4.31, the bounded biharmonic weights are computed and displayed in Figure 4.15 where the salient parts are the distribution of the weights. It is obvious to see that the bounded biharmonic weights are local and sparse. Each control point only influences the shape in its vicinity and each vertex on the surface is only controlled by a few closest control points. The influence of each control point distributes smoothly from the control point to the vertices.

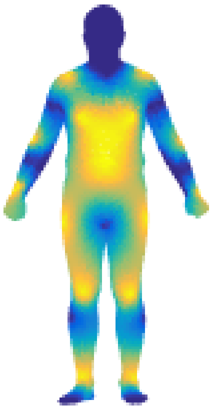


Figure 4.15: The weight influences distributed on the surface of the template mesh.

In order to drive the pose, given the bounded biharmonic weights, linear blending scheme is applied. Compared to [127], only 16 landmarks are used to change the pose of the template. Experiments show that these sparse landmarks provide enough control to pose the template. All points $\mathbf{p}$ are deformed by their weighted combinations:

$$\mathbf{p}' = \sum_{j=1}^{m} w_j(\mathbf{p}) T_j \mathbf{p}, \tag{4.32}$$

where w_j is the weight associated with bone j and T_j is the transformation of the control point j.

4.3.3 Landmark based Multilevel Active Body Registration

After the posing step, the template is deformed by the sparsely manual-annotated landmarks and performs the same pose with the target scan. Then, the proposed MABR approach in Section 4.2 is applied for the mesh registration.

With the morphable model trained in Section 4.2.3, the first level of registration is performed to overlap the trained template model and target scan. With target point clouds $\mathbf{u}$ retrieved by nearest neighbours search using the kd-tree algorithm, the cost function to be minimized is formulated as:

$$E(\mathbf{c}) = ||\mathbf{v} - \mathbf{u}||^2 = ||(\mathbf{B}\mathbf{c} + \mathbf{m}) - \mathbf{u}||^2. \tag{4.33}$$

The partial derivative is taken with regard to $\mathbf{c}$ and the minimum is achieved when it approaches zero and get the closed-form solution,

$$\mathbf{c} = -(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\mathbf{m} - \mathbf{u}). \tag{4.34}$$

For capturing the details of the body surface, the whole registration is composed of: the distance term and the smooth term.

Distance term: The distance term is used to minimize the Euclidean distances between the source and the target. Each part has n_i points and the cost function is denoted as:

$$E_d(\mathbf{X}) = \sum_{i=1}^{p} \sum_{j=1}^{n_i} ||\mathbf{X}_j^i \mathbf{v}_j^i - \mathbf{u}_j^i||^2. \tag{4.35}$$

Since each part is modelled by the shape model $\mathbf{v}_j^i = \mathbf{B}_j^i \mathbf{c}_j^i + \mathbf{m}_j^i$, based on Eq.4.7, the distance term can be rewritten and rearranged as:

$$\begin{aligned} E_d(\mathbf{X}) &= \sum_{i=1}^{p} \sum_{j=1}^{n_i} ||\mathbf{X}_j^i (\mathbf{B}_j^i \mathbf{c}_j^i + \mathbf{m}_j^i) - \mathbf{u}_j^i||^2 \\ &= \sum_{i=1}^{p} ||\mathbf{D}^i \mathbf{X}^i - \mathbf{U}^i||^2, \end{aligned} \tag{4.36}$$

where the term $\mathbf{D}^i = diag({\mathbf{v}_1^i}^T, {\mathbf{v}_2^i}^T, ..., {\mathbf{v}_{n_i}^i}^T)$, and the set of closest points $\mathbf{U}^i = [\mathbf{u}_1^i, \mathbf{u}_2^i, ..., \mathbf{u}_{n_i}^i]^T$.

Stiffness term: The stiffness term penalizes the difference between the transformation matrices of neighbouring vertices. Same with Eq. 4.37, it is defined as:

$$E_s(\mathbf{X}) = \sum_{i=1}^{p} ||(\mathbf{M}^i \otimes \mathbf{G}^i)\mathbf{X}^i||_F^2, \tag{4.37}$$

Complete cost function for Main Body: The terms defined in Eq. 4.36 and Eq. 4.37 are combined into the complete cost function that is minimized:

$$\begin{aligned} \underset{\mathbf{X},\mathbf{c}}{\operatorname{argmin}} E(\mathbf{X}) &= \underset{\mathbf{X},\mathbf{c}}{\operatorname{argmin}} [E_d(\mathbf{X}) + \alpha E_s(\mathbf{X})] \\ &= \underset{\mathbf{X},\mathbf{c}}{\operatorname{argmin}} \sum_{i=1}^{p} \left\| \begin{bmatrix} \mathbf{D}^i \\ \alpha \mathbf{M}^i \otimes \mathbf{G}^i \end{bmatrix} \mathbf{X}^i - \begin{bmatrix} \mathbf{U}^i \\ \mathbf{0} \end{bmatrix} \right\|_F^2 \\ &= \underset{\mathbf{X},\mathbf{c}}{\operatorname{argmin}} \sum_{i=1}^{p} \left\| \mathbf{A}^i \mathbf{X}^i - \mathbf{U}^{*i} \right\|_F^2, \end{aligned} \tag{4.38}$$

which is a well-known linear least square problem. The minimum can be obtained when the gradient vanishes as below.

$$[\partial E(\mathbf{X})/\mathbf{X}^1; \partial E(\mathbf{X})/\mathbf{X}^2; ...; \partial E(\mathbf{X})/\mathbf{X}^p] = \mathbf{0}. \tag{4.39}$$

Similar to the optimization of Eq.4.15, the Eq. 4.38 can get the closed-form solution for each part:

$$\mathbf{X}^i = ((\mathbf{A}^i)^T \mathbf{A}^i)^{-1} (\mathbf{A}^i)^T \mathbf{U}^{*i}. \tag{4.40}$$

Similarly, the hands/feet are processed separately. Only the pre-trained PCA models are used to perform the deformation.

Distance term: It is defined similar to Eq. 4.36, but without the affine transformation matrix,

$$E_d(\mathbf{c}^*) = ||(\mathbf{B}^* \mathbf{c}^* + \mathbf{m}^*) - \mathbf{u}^*||_F^2. \tag{4.41}$$

Boundary smoothness term: In order to stitch hand/foot with its neighbouring part smoothly, a boundary smoothness term is defined as follows:

$$E_b(\mathbf{c}^*) = ||(\mathbf{B}_b \mathbf{c}^* + \mathbf{m}_b) - \mathbf{F}_b||_F^2 \tag{4.42}$$

where $\mathbf{B}_b$ is the statistical model of boundary points of template and $\mathbf{F}_b$ is their corresponding boundary points on the neighbouring part. By enforcing the boundary constraints between two parts, the part fitting process can be regulated to avoid erroneous result caused by outliers.

Complete cost function for Hand/Foot Part: The fitting objective function can be formulated as:

$$\begin{aligned} E(\mathbf{c}^*) &= \alpha E_d(\mathbf{c}^*) + \beta E_b(\mathbf{c}^*) \\ &= \left\| \begin{bmatrix} \alpha \mathbf{B}^* \\ \beta \mathbf{B}_b \end{bmatrix} \mathbf{c}^* + \begin{bmatrix} \alpha \mathbf{m}^* - \alpha \mathbf{u}^* \\ \beta \mathbf{m}_b - \beta \mathbf{F}_b \end{bmatrix} \right\|_F^2 \\ &= \left\| \mathbf{A}^* \mathbf{c}^* - \mathbf{U}^* \right\|_F^2, \end{aligned} \tag{4.43}$$

where α is the weighting factor between two terms, $\mathbf{A}^* = [\alpha \mathbf{B}^*, \beta \mathbf{B}_b]^T$ and $\mathbf{U}^* = [-\alpha \mathbf{m}^* + \alpha \mathbf{u}^*, -\beta \mathbf{m}_b + \beta \mathbf{F}_b]^T$. The minimum of this linear least square problem occurs where the gradient approaches to zero, that is $\partial E_{\mathbf{c}^*} / \partial \mathbf{c}^* = \mathbf{0}$. Thus Eq. 4.43 has closed-form solution:

$$\mathbf{c}^* = (\mathbf{A}^{*T} \mathbf{A}^*)^{-1} \mathbf{A}^{*T} \mathbf{U}^*. \tag{4.44}$$

4.3.4 Evaluation

Evaluation on scans from professional devices *SPRING* [189] dataset contains 3038 high-quality meshes with a wide range of human body shapes. The meshes have the same topology with SCAPE dataset [6]. The source meshes of SPRING dataset are laser scans from the CAESAR dataset. The segmentation of 16 body parts is also provided as shown in Table 4.1. A hand and the associated forearm are integrated into one part as the number of points for a hand is too small. Similarly, a foot and the connected calf are combined together. By doing this, the segmentation of 12 parts for each mesh is formed. In the experiment of high quality meshes, 200 meshes are randomly selected as the training set and the remaining meshes in SPRING dataset are the test set. Holistic and local shape models are trained for the male and female separately. During the training of the shape model, 99% of the variations are kept in the experiment.

Evaluation on the SPRING dataset To show the superior performance of the proposed method, it is compared with classic **NICP** in [5] and the latest work of **ANICP** in [33] which also combines PCA and ICP but it is designed for face alignment. The RMS Error is computed with the Eq. 4.45 to measure the accuracy of three methods.

Table 4.1: The human body mesh segmentation

Index	Part	#Points	Proportion
0	right_foot&right_calf	1324	10.58%
1	left_foot&left_calf	1326	10.60%
2	right_thigh	1123	8.98%
3	left_thigh	1086	8.68%
4	bottom	881	7.04%
5	right_hand&right_forearm	856	6.84%
6	left_hand&left_forearm	804	6.43%
7	low_torso	1177	9.41%
8	up_torso	1531	12.25%
9	right_up_arm	443	3.54%
10	left_up_arm	363	2.9%
11	head	1586	12.69%

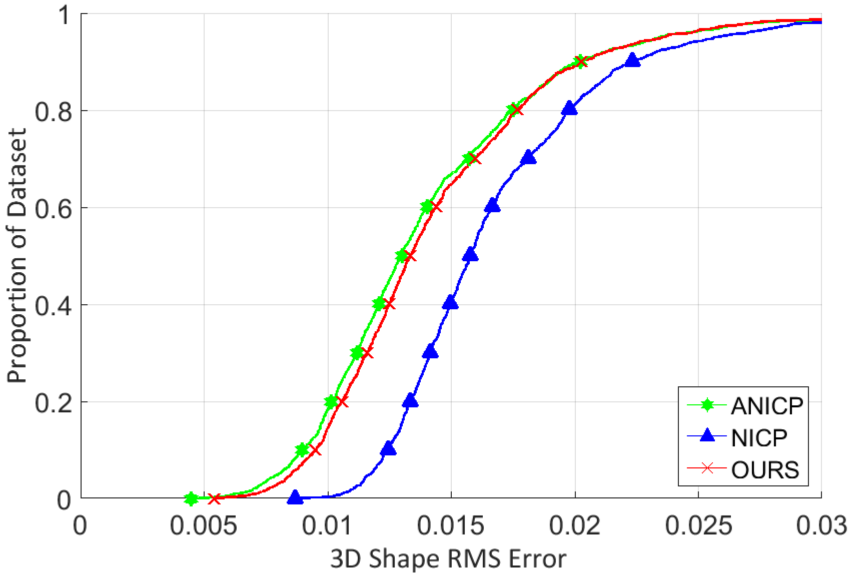


Figure 4.16: The comparison of 3D shape RMS error of ANICP, NICP and our methods on SPRING dataset.

$$RMSError = \sqrt{\frac{\sum_{i=1}^{n}(p_i - \hat{p_i})^2}{n}}, \tag{4.45}$$

where points p_i and $\hat{p_i}$ are ground truth and the fitted results; n is the number of the points of 3D template mesh. As shown in Figure 4.16, the accuracy of the proposed method is comparable with ANICP and higher than NICP. Although the human body meshes have only standing pose, the overlapping degree in limb parts varies from different meshes. It is easy for limbs to be regarded as outliers in ICP-based algorithms. As a result, the fitting results of NICP and ANICP in the hand and foot parts are distorted and erroneous. The accuracy of ANICP and NICP is also impaired. In the proposed method, the hands and feet are deformed by the trained local shape model. The shapes of the hands and feet are well kept by the proposed method.

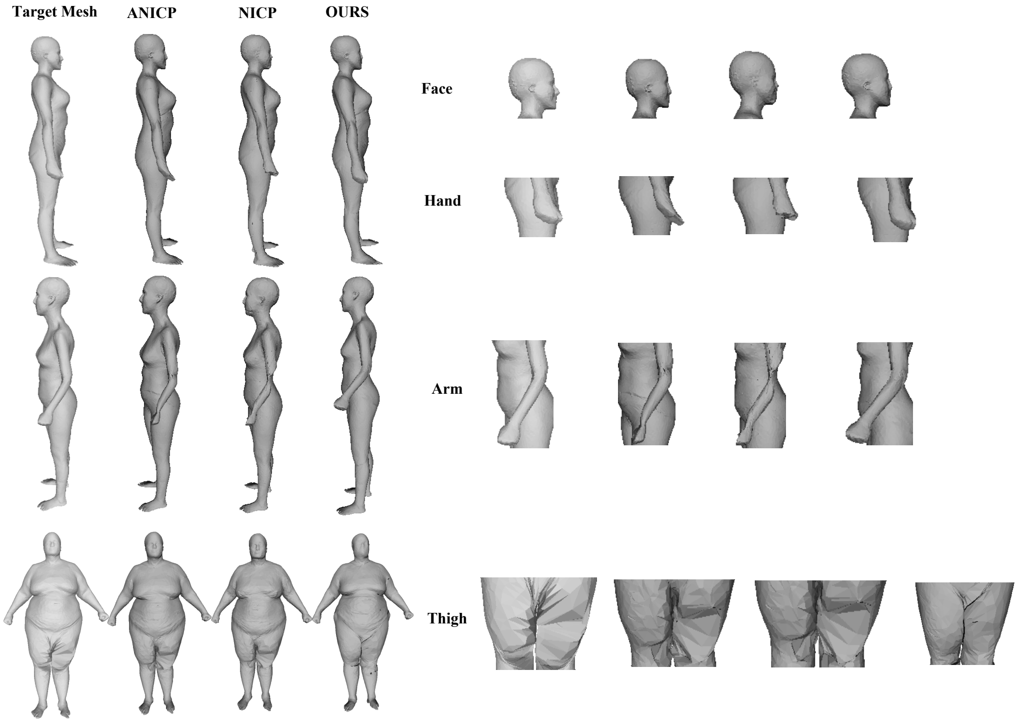


Figure 4.17: Registration results on SPRING dataset.

In order to compare the robustness to outliers, the details of fitted results from ANICP, NICP and our method are also displayed in Figure 4.17. In the first row, compared with the proposed method, the hand parts of ANICP distort severely and the fitted hand of NICP is not clear. As for deformed arms in the second row, the results of ANICP and NICP are broken while our method is able to preserve the continuity of the fitted mesh and the deformed arm keeps the original arm shape of the target.

Moreover, in the last line when the target mesh has irregular points on the thigh part, the proposed method manages to recover the reasonable thigh shape, benefiting from the process of learning prior knowledge of body shape from the training set. In comparison, in ANICP and NICP the transformation of each vertex is only controlled by the closest points. Therefore, ANICP and NICP are not able to predict the reasonable positions in the deformed meshes. The example in the third line also verifies that our method is robust to noisy data.

Evaluation on the MPI dataset The MPI dataset as described in [61], is captured using a professional Vitronic laser scanner. As the number of captured poses for each subject is random in MPI, they are categorized into 4 classes based on their visual similarity. 5-fold cross validation for each class is performed to compare the accuracy of the proposed method with ANICP and NICP for each class. For each round, 4 folds of the data are used as the training set to build the holistic and local models, the remaining one fold is used as the test set. The average of RMS Error is used to measure the accuracy. Figure 4.18 shows the RMS errors of three methods for each class over different proportions of the dataset. Clearly, the proposed method outperforms the ANICP and NICP for the four classes. Although the meshes in 35 poses are classified into four classes, meshes still vary largely in pose within the class. When NICP and ANICP are performed on those data which varies much in poses, some target points are easily regarded as outliers, resulting in erroneous fitting.

The proposed method is also compared with ANICP and NICP visually in Figure 4.19. It can be observed that the proposed method is able to register the limb parts robustly and accurately. However, in ANICP and NICP, the arm parts in the first two rows do not present the target both in terms of arm shapes and the lifting degrees. This is because both ANICP and NICP rely on a good initialization. The prior knowledge of human body shape increases the robustness of our algorithm, as it can be seen in the last column of results in Figure 4.19. In the third row, some vertices on the left leg are registered to the right. Bad initializations make the nearest point in ANIP and NICP confusing, which causes adhesion of legs.

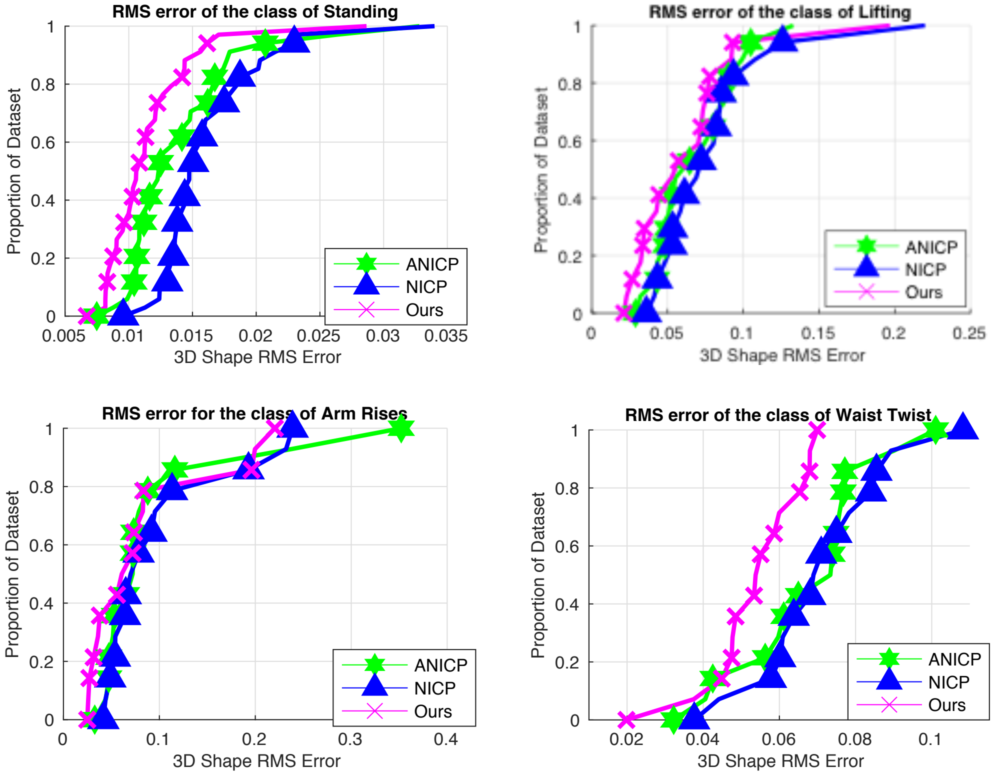


Figure 4.18: The comparison of ANICP, NICP and the proposed method on MPI.

Evaluation on Low quality scans from Kinect More importantly, besides the test on scans of professional scanners, the proposed method is also evaluated on K3D-hub dataset which is captured by untrained users with a single Kinect which provides relatively low resolution data (320×240) with a high noise level. All the meshes in K3D-hub are used as the test data and the holistic and local shape models are trained with 200 meshes randomly selected from the SPRING dataset. Fitting results of three methods are shown in Figure 4.20. It is clear to see that the proposed method is the only one that models the hand and foot parts completely and keeps high accuracy of the fitted results. In real scans, the slight movement of limbs is inevitable, resulting in a partial overlap with the template limbs. Therefore, the points retrieved by the nearest neighbour search algorithm cannot keep the limb shape of scans, leading to some unexpected fitted shapes while our method is able to keep a good shape of the scan and also robust to noises.

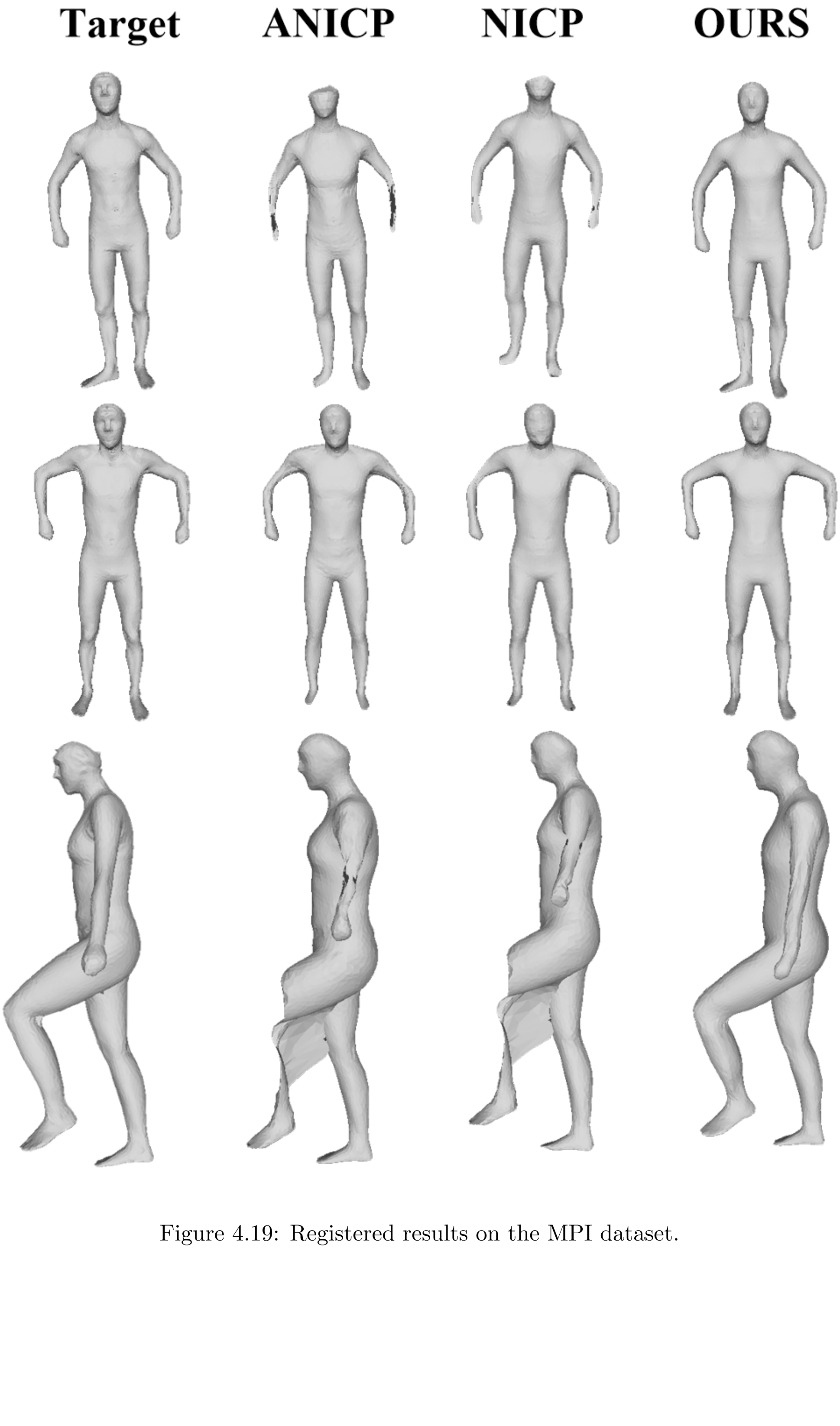


Figure 4.19: Registered results on the MPI dataset.

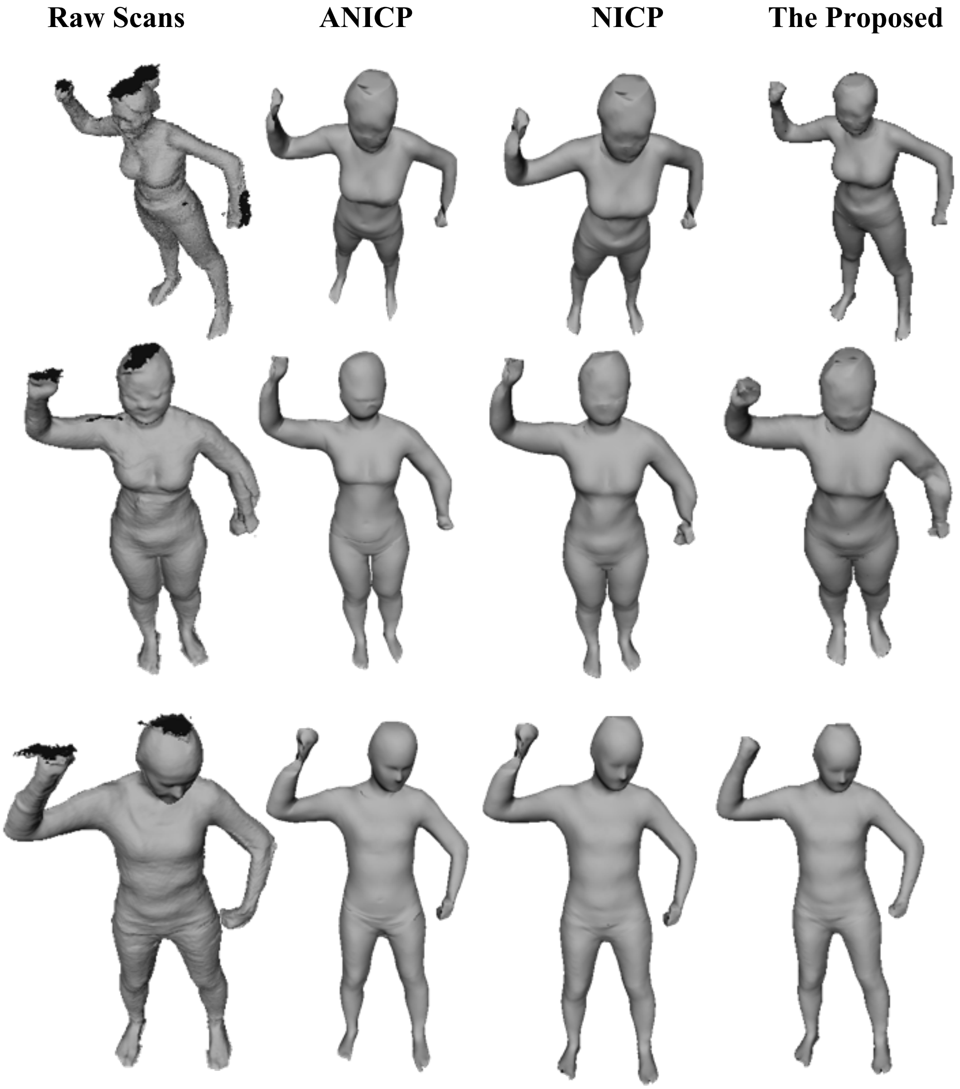


Figure 4.20: Fitting results from Kinect scans.

4.4 Summary

In this chapter, the approach of building human body models from user-generated point clouds which are acquired with Kinect is presented. In the Section 4.2, a multilevel active registration (MABR) method is proposed which combines the non-rigid ICP with the statistical shape model to automatically fit the body template model to the target point clouds. Since the PCA shape model is trained with 200 registered mesh, the combination of PCA makes our method robust to noise, outliers and holes. It is shown that the performance of the proposed algorithm is comparable with the state-of-the-art non-rigid registration methods and outperforms them when it comes to the alignment of hands/foot parts. Experiments verify that the proposed approach is robust to both noisy Kinect scans and high-quality meshes.

The second Section 4.3 extends the MABR method to different poses by posing the template with the help of a sparse set of landmarks. Compared with the popular NICP and ANICP, the proposed method takes specific care of particular body parts by defining different energy functions. The experiments demonstrate that the landmark based MABR approach successfully registers the scans with different poses from the template. The registration results are comparable with the state-of-the-art in the high quality meshes and outperform them in the case of medium and low quality meshes.

Besides the robust registration method, a Kinect based human body dataset, named K3D-hub, is collected which is the first publicly available low-quality human body scans dataset.

Chapter 5

Estimation of 3D Human Body Models from User-captured 2D Images

In Chapter 4, the problem of building high-fidelity human body models from user-generated point clouds is discussed. With the proposed robust registration methods, the challenges of registration caused by noises, holes and obscure parts are successfully addressed. Beside point clouds, a more common kind of user-generated data is images. A number of traditional methods of 3D reconstruction from dozens of 2D multi-view images have been proposed. However, building human body models from a single image is still a challenging task. In this chapter, the focus is on the accurate estimation of human body models from a single 2D image.

Research findings have suggested that building realistic and personalized human avatar is tightly coupled with next-generation industry, such as virtual reality, digital gaming and movie effect et al. Accurate estimation of human body shapes and poses for real people is an important clue for virtual try-on. However, the current virtual try-on applications mainly use predefined virtual avatars which are created by professionals. The model is often unrealistic as it lacks the simulation of human body muscle and soft tissue.

To address this problem, this chapter presents an automatic method, given a single image, to estimate accurate 3D human body shapes and poses by using the state-of-the-art parametric human body model. The acquired 3D model can be used for improving the display of clothing on the consumer's body for on-line business. To achieve this goal, a parametric 3D human body model SMPL is fitted to 2D joints as well as the boundary of the human body which is segmented using CNN methods automatically. Considering the scenario of virtual dressing where people are in stable poses at most of the time, a stable pose prior is trained from the CMU motion capture (mocap) dataset [1] for further improving the accuracy. Experiments verify that the stable pose prior can provide more accurate pose estimation than general pose prior and boundary constraints improve the shape estimation significantly.

5.1 Overview

With the advances in online fashion business, novel recommendation technologies are essential for promoting sales, improving shopping experiences and reducing the high rate of return and realistic virtual dressing systems play an important role in personalised clothing recommendation. For consumers, it enhances the shopping experience by allowing customers to virtually but realistically 'try-on' apparel without being physically presented in the retail shop. For manufacturer and retailers, such technology facilitates real-time visualization of the garments' appearances on the human body in the design process and thus improves the fitting quality of garment designs. It also improves the efficiency and substantially reduces the cost. However, the premise of the above benefits is that a realistic human avatar of can be straightforwardly obtained.

[1] http://mocap.cs.cmu.edu/

In current virtual dressing applications, most virtual human body models are pre-made, i.e., designed by animators according to a set of pre-defined sizes based on a rough range of chest circumference, hip width and waist breadth. However, as muscle and soft tissue change with gender, ages or even occupations, human body shapes vary in each individual with many possibilities. The garment's fit on a template model does not necessarily reflect its look on the real customers. With the development of 3D scanning technology, many cost-effective scanners are accessible to ordinary people and many approaches [193, 44, 199] are proposed to use 3D scanning devices to capture 3D models of real people with one or several commodity 3D scanners. Although the 3D models can be obtained more easily than before, the scanning systems usually require people to be physically presented where the scanning systems are located. While in the next-generation of digital clothes retailers and fashion industry, consumers expect to remotely generate their realistic 3D human body models with observations of their body shapes and poses using mobile devices that are commonly accessible, like selfies taken by smart phones. However, due to the lack of depth information, estimation of human body shape and pose from a single image is very challenging. So far many CNN based methods have been widely applied to detect the 2D joints [25, 111, 176] but need a sufficient number of annotated images to train networks. Methods in [174, 106] estimate the 3D joint position in conjunction with an existing 2D joint detector. Guan et al. [55] estimate the human body shape and pose from a single image by fitting a parametric model of SCAPE [6] to silhouette overlap, edge distance and smooth shading. SMPLify [17] applies a simpler but effective parametric model of SMPL to infer body shape and pose by merely fitting 2D joints. However, aiming at practical applications, more accurate estimation of human body shape and pose from a single image is required.

Here, an automatic method is proposed to accurately estimate 3D human body meshes from single 2D images and hopefully push a step forward to virtual dressing applications. Rather than reconstructing human body avatar based on multiple 2D images from different views which needs careful camera calibration [77] or deep leaning methods [34, 161, 99] which need a large amount of training, in this research, the realistic human body avatar from a single 2D image is estimated by fitting the state-of-the-art 3D parametric human body model, SMPL [95] to the image to estimate human body shapes and poses. The SMPLify [17] method is taken as a base method and we go beyond it by exploiting boundary information of images to constrain the deformation of SMPL. Considering that people are usually in a stable status, for instance standing still or walking slowly in the scenario of virtual dressing, a stable pose prior is built to narrow the search space of human poses, making the pose estimation more precise. With the accurately estimated human body shape and poses, virtual dresses are fit on models to demonstrate the visualization of clothing on the human bodies.

5.2 Related Work

As the goal of this chapter is to build believable human body models towards the application of virtual try-on, the literatures on the acquirement of human body models using current popular reconstruction methods and the state-of-the-arts in virtual try-on applications are reviewed here.

3D human body reconstruction With the development of 3D scanning technique, many approaches are proposed to obtain human body models with one or several commodity 3D scanners. Li et al [89] use one Microsoft Kinect to capture 3D selfies for users without help from others. Body Fusion [193] proposes a real-time geometry fusion method with one Kinect for the reconstructing of human body models. Fusion 4D [44] enables high-quality reconstruction from noisy input. In [119], authors propose a calibration method for three Kinects to capture complete human body point clouds in real time. However, these approaches can acquire the human body scans but cannot estimate the close-to-naked human body shapes. Thus, the captured human body meshes can not be used for virtual try-on applications.

Recently, the availability of large-scale 3D datasets, like shapeNet [27], SURREAL [172], promotes 3D human related analysis with the help of deep learning method. In

[99], 2D sketches are input to a Convolutional Neural Network to generate corresponding point clouds. The 3D Generative-adversarial network (3D-GAN) is used in [182] to infer 3D models from a single image and the output is a voxel based representation. Wu et al. [183] propose a generative model based on a deep belief network trained on voxelized 3D shapes. Girdhar et al. [51] propose a TL-embedding network allowing 3D reconstruction from a 2D image. However, apart from the required large amount of training data, the above methods suffer from the low resolution of voxel representation that can only produce coarse 3D shapes and be lack of details.

Due to the specific data type of human body where surface elements can be consistently parameterized through correspondence techniques, the parametric models of human body has been extensively investigated in the past. Many parametric human body models [3, 2, 6, 95, 61, 31] have been proposed to capture human body shape of many people and the non-rigid deformation due to pose, facilitating image reshaping [204], pose estimation [172], anthropology measurements [169] and clothing modelling [129, 56]. Allen et al. [2] model human body shape space using PCA. The popular SCAPE model is a face-based human modelling approach which models the non-rigid deformation of triangle faces due to the pose. BlendSCAPE [67] improves SCAPE by defining the rotation of triangle faces to be a linear blend of the rotation of body parts. Allen et al. [3], Hasler et al. [61] and Loper et al. [95] model the human body shape and pose spaces in vertex way and experiments have proved that vertex based modelling methods are more accurate than face based approaches [95]. In order to describe the person-dependent pose deformation, Chen et al. [31] propose to jointly model human body shape and pose. Recent SMPL [95] is a learned linear model which is simpler than SCAPE but can produce equally or better results than SCAPE by adding pose blend shape to complement the shape changes due to pose. Thus, in this chapter, the human body parametric model - SMPL is used to estimate the naked human body models from a single image.

Virtual dressing systems With the advances e-commercial, people tend to go shopping online and virtual dressing system is indispensable for improving shopping experiences. Virtual try-on system usually has a synthetic human body model that is predefined and might be able to adjust the height, weight in a coarse way. Different costumes are put on the model and the system shows the fitting effect. In [41], a laser scanner is used to scan real people to acquire the 3D mesh model and after a series of pro-processing like purifying and smoothing, clothes are put on the avatar. In this work, clothes are modelled with scans of real clothes draping on a dressmaker's dummy. The TriMirro [2] systems simulate clothes on a predefined human avatar which can not describe various human body shapes and poses realistically. Fitnect [3] models both human body shape and clothes and allows animation of clothes with pose changing. Ye et al. [190] use RGBD data as input to reconstruct a personalized 3D avatar and adapt synthesized clothes. However, the above systems either use predefined avatar or reconstruct 3D human body models with depth sensors. These methods suffer from the inaccurate estimation of real human body shapes, which may recommend the wrong clothes sizes. In addition, consumers may make efforts to go to retail shops where virtual try-on systems are available to try clothes virtually. Therefore, it is desirable to have realistic and convenient simulation of human full body.

5.3 Accurate Human Body Shape Estimation in Stable Poses

In the proposed method, the first step is to extract the 2D body joints and boundary for images using CNN-based joint estimator, DeepCut [126] and DeepLab [30]. The estimated 2D joints and boundary for the image is illustrated in Figure 5.1. Here, the state-of-the-art parametric human body model SMPL is used to estimate the human body shape with the help of the detected 2D joints as well as the boundary. In the first level, as shown in the Figure 5.1(a), the SMPL model is deformed by minimizing the distance between the projection of the SMPL joints and the 2D detected joints of

[2] http://www.trimirror.com/en/

[3] http://www.fitnect.hu/

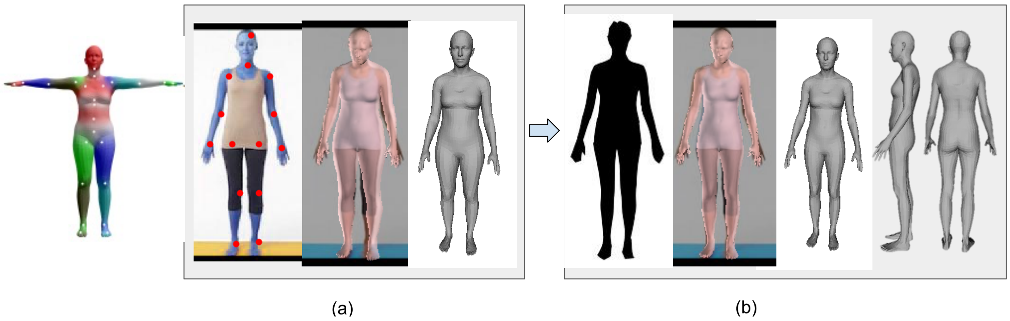


Figure 5.1: The pipeline of the proposed method. (a) A parametric human body shape and pose model is fitted to the 2D images with the guidance of the detected 2D joints. (b) Boundary is used to further improve the accuracy of the estimated human body shape and poses.

images. In the second level that is shown in Figure 5.1(b), the deformed SMPL obtained in the first level is further deformed by penalizing the distance between the projection of the SMPL boundary and the detected boundary. Note the shrinkage of the hips in Figure 5.1(b).

5.3.1 SMPL Human Body Model

SMPL is a human body model that is learned from thousands of registered human body meshes. It can generate a triangular mesh with 6890 vertices and 23 joints. SMPL is controlled by two sets of parameters: shape parameters to describe human body shape and pose parameters for pose. SMPL is defined as $M(\beta, \theta; \Phi)$, where shape parameters $\beta \in \Re^{1\times 10}$ are coefficients of a low-dimensional shape space and describe shape variances due to identity. A new body shape can be obtained by adding a linear combination of blend shapes S_i to the template mesh T:

$$T' = T + \sum_i \beta_i S_i. \tag{5.1}$$

The skeleton structure of SMPL model represented by $M(\beta, \theta; \Phi)$ has 23 joints. $\theta \in \Re^{1\times 72}$ is a vector of pose parameters where the first three elements are orientation and the following every three elements are the axis-angle representations of the relative rotation for each joint in kinematic tree. The axis-angle representation is defined in Eq. 5.2 where axis $[x, y, z]^T$ is normalised and angle θ is in radian representation. Rodrigues transformation is used to change the form of the axis-angle representation into the rotation matrix form while performing pose-dependent deformation.

$$< axis \cdot angle >= ([x, y, z]^T \cdot \theta). \tag{5.2}$$

Φ is the full set of model parameters which includes template mesh, skinning weight, joint location, shape blend shapes and pose blend shapes, learned from a large number of registered 3D human body meshes. Once learned, Φ is held fixed and new body shapes and poses are created and animated by varying β and θ. Besides pose related blend shape which simulates human body shape changes with regard to poses, SMPL also automatically estimates the joint location $J = [j_i^T ... j_n^T]^T$ as a function of body shape.

5.3.2 Stable Pose Prior

In the scenario of virtual try-on, people commonly stand or move slowly in front of the camera. The pose variance is limited. As the CMU dataset covers various human poses presented in daily life and sports for 144 subjects, a general pose prior cannot describe some specific poses accurately. Experiments show that the results of SMPLify present bent knees or stoop for the stable pose of "Standing". In order to provide more accurate pose prior for this case, the stable poses from CMU dataset are firstly identified. We define the stable poses to be those change slightly in a short period of time. For each frame, the error between its neighbouring frames is calculated as:

$$err = \frac{\sum_{k=-step}^{step} norm(\theta_i - \theta_{i+k})}{2 \times step}. \tag{5.3}$$

Here, the step is set to be 1 and θ is the pose parameter of the motion in each frame and when *err* is smaller than a *threshold*, the pose is regarded as stable poses. In the experiment, the threshold is set to be 0.001. The method proposed in [94] is applied to calculate the pose parameters θ for each frame of stable poses, which captures motion and shape from sparse markers provided by CMU mocap data.

Some selected stable pose samples are shown in Figure 5.2. As it can be seen, stable poses cover various kinds of poses, including stand, squat, leaning and sitting, which are common poses in a try-on scenario. With stable poses, the Gaussian Mixture Model

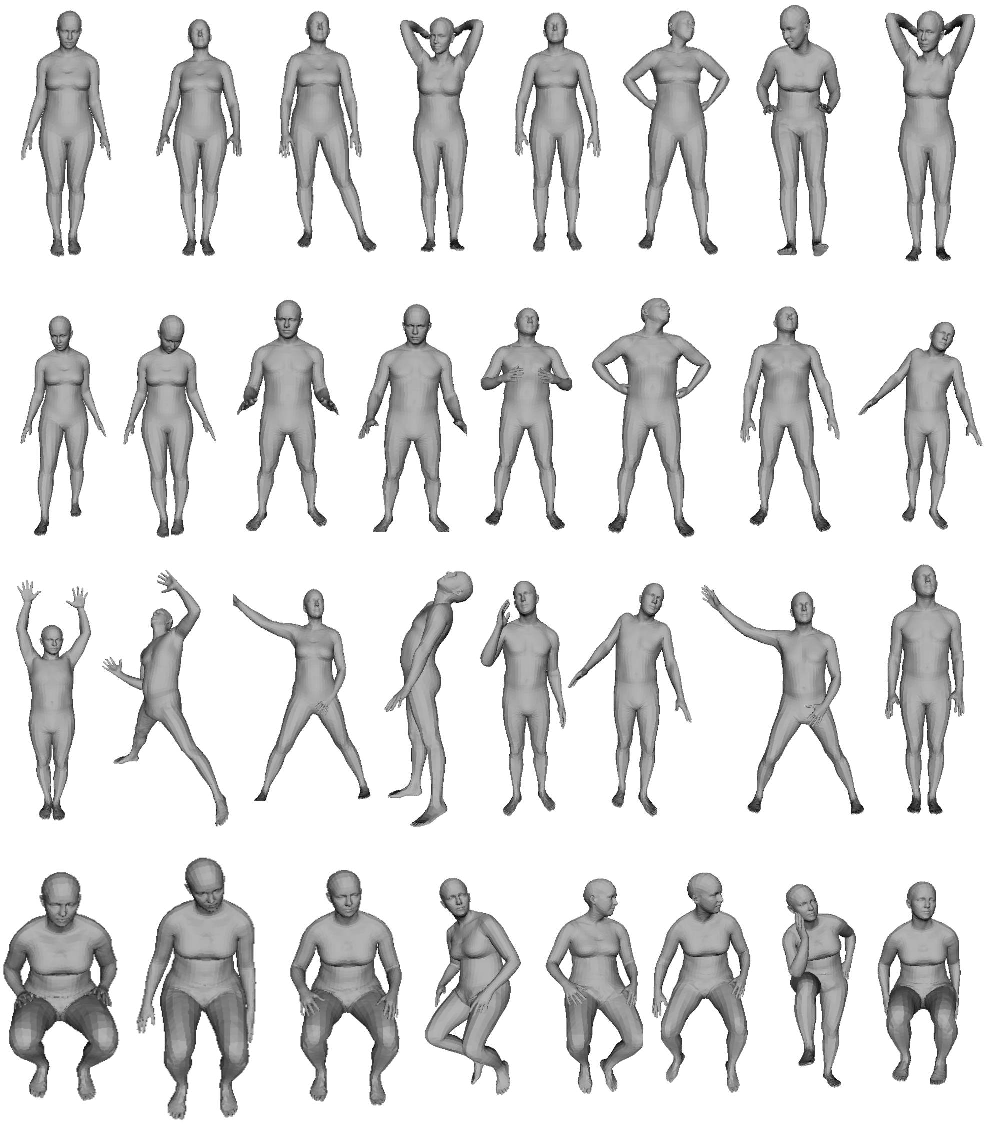


Figure 5.2: Sample stable poses.

(GMM) is used to describe the pose prior in this work.

5.3.3 Boundary Assisted Human Body Shape and Pose Estimation

Basic figure estimation The SMPL model is taken as the human body representation. Before using boundary information to improve the accuracy of body shape estimation, joints are used to estimate the basic figure and poses. Given the estimated 2D joints of the single image J_{est}, the energy function is formulated as:

$$E_M(\beta, \theta) = E_J(\beta, \theta; K, J_{est}) + \lambda_\theta E_{S\theta}(\theta) + \lambda_\alpha E_\alpha(\theta) + \lambda_\beta E_\beta(\beta), \tag{5.4}$$

where E_J is the data term and $E_{S\theta}$, E_α and E_β are priors. These terms are explained in details as follows.

Data term: this term encourages the template M to be close to target image. For each joint i of SMPL template, J_{est_i}, the distance between its projection position and the corresponding image joint is minimized. The joint fitting term is formulated as follows:

$$E_J(\beta, \theta; K, J_{est}) = \sum_{jointi} \omega_i \kappa(\Pi_K(R_\theta(J(\beta)_i)) - J_{est,i}), \tag{5.5}$$

where Π is the projection function from 3D to 2D that is defined by the camera parameter K; J is the joint estimation function, which returns joint locations; R is the rotation function; ω_i is the confidence that is gained from the extraction of joints with DeepCut and its value depends on the confidence of its estimation.

Shape prior term: $E_\beta(\beta)$ is shape prior learned from the SMPL body shape training set. The shape parameters β are PCA coefficients of a low-dimensional shape space, learned from thousands of registered scans. In this research, the number of used coefficients is 10.

$$E_\beta(\beta) = \beta^T \Sigma_\beta^{-1} \beta, \tag{5.6}$$

where Σ_{β}^{-1} is a diagonal matrix computed with PCA from the shape in the SMPL training set.

Stable pose prior term: $E_{S\theta}(\theta)$ and $E_{\alpha}(\theta)$ are pose priors which are learned from precomputed stable poses. Here, $E_{\alpha}(\theta)$ is only performed on the knees to avoid unnatural bending and is defined as:

$$E_{\alpha}(\theta) = \sum_{i} \exp(\theta_i), \tag{5.7}$$

where θ_i represents the pose parameters that correspond to the bending of knees or arms. When the joint is not bent, θ_i is zero. The function of exp() is a monotonically increasing function so that when it approximates to zero, the θ_i is negative which is natural bending and the penalty is not heavy while positive bending is unnatural and is penalized more. $E_{S\theta}(\theta)$ can favor probable stable poses over unstable ones. After training the stable pose prior in Section 5.3.2, $E_{S\theta}(\theta)$ is defined as the negative logarithm of a sum. As described in [113], a max mixture has much of the same character as a sum mixture and retains a similar expressivity but is well compatible with our optimization framework. Thus, we approximate the sum in the mixture of Gaussian by a max operator:

$$\begin{aligned} E_{S\theta}(\theta) &= -\log \sum_{j} (g_j \mathcal{N}(\theta; \mu_{\theta_j} \Sigma_{\theta_j})) \\ &\approx -\log(\max_{j} g_j \mathcal{N}(\theta; \mu_{\theta_j} \Sigma_{\theta_j})), \end{aligned} \tag{5.8}$$

where μ_{θ_j} and Σ_{θ_j} are trained with are trained with 30000 stable poses; g_i is the weight of each Gaussian mixture model component $\mathcal{N}(\cdot)$.

Boundary assisted shape estimation After the *Basic figure estimation* with stable pose prior described above, the initial pose and shape have been estimated. Boundary information is very important to enlarge or shrink the model to make the final estimated human body shape similar to the real person. To achieve this goal, the pre-trained model provided in [30] is adopted to predict the boundary of the human body images.

The optimization is defined as:

$$E(\beta, \theta) = E_M(\beta, \theta) + E_b(\beta, \theta; K, U). \tag{5.9}$$

Boundary term: this encourages the projected boundary of the human body to be close to the image boundary. After performing the optimization above, the camera position has been estimated, the boundary of SMPL model can be extracted from its projection in the camera. The boundary term is defined as follows:

$$E_b(\beta, \theta; K, U) = \sum_i^N ||(B_i - U_i(\Pi_K(M(\beta, \theta))))||^2, \tag{5.10}$$

where B_i is the i_{th} point on the boundary of images, $\Pi(\cdot)$ is the projection function and $U_i(\cdot)$ is the corresponding points of B_i on the boundary of the projected model. Combined with Eq. 5.4, the complete cost function defined in Eq. 5.9 is written as:

$$\begin{aligned} E(\beta, \theta) &= E_M(\beta, \theta) + E_b(\beta, \theta; K, U) \\ &= E_J(\beta, \theta; K, J_{est}) + \lambda_\alpha E_\alpha(\theta) + \lambda_\beta E_\beta(\beta) \\ &+ \lambda_\theta E_{S\theta}(\theta) + E_b(\beta, \theta; K, U) \end{aligned} \tag{5.11}$$

During the optimization, the camera parameter K which is composed of focal length, camera rotation and camera translation are firstly estimated. In this research, the camera focal length is required to be known and we do not consider the lens distortion. It is assumed that the person is standing parallel to the image plane. The camera translation is estimated via the ratio of similar triangles, defined by the torso length of the mean SMPL shape and the predicted 2D joints. The rotation of the camera is initialised to be a zero vector. Then we optimize the camera translation and rotation by minimising E_j during which β keeps fixed to the mean shape. The boundary of the projected model is updated for each round of optimization. For further accelerating the convergence of the cost function, the initial pose and shape are first obtained from Eq. 5.4. Starting from the optimized model, boundary assisted optimization is performed. Eq. 5.4 and Eq. 5.11

are minimized using $Powell's$ dogleg method, using OpenDR [96] and Chumpy [4]. The Optimization for a single image takes around 1 minute on a common desktop machine with 16 GB RAM and 4 cores.

5.4 Evaluation

5.4.1 Evaluation of Pose Estimation

A quantitative evaluation of the accuracy of 3D pose estimation is performed on CMU dataset. In order to show the superior performance of the proposed approach, the results are compared with SMPLify which predicts 3D pose from 2D joints.

According to the definition of stable poses in Section 5.3.2, the threshold is set to 0.001 to find frames of stable poses in the CMU mop dataset. Mosh [94] is firstly performed to get pose parameters for each frame which captures motion and shape from sparse markers provided by CMU mocap dataset. The acquired pose parameters are regarded as ground truth for evaluation. To evaluate the proposed stable pose prior, the human body meshes are synthesized by giving the ground-truth pose parameters and shape parameters are fixed to be zeros. Their joints are projected into 2D with a known camera.

Firstly, the comparison is conducted on stable poses. During experiments, 42797 stable poses are identified based on the Eq. 5.3 and 5-fold cross validation is performed for each class to compare the method against the state-of-the-art method SMPLify [17]. The error is calculated between the ground truth θ_{gt} and the estimated pose θ_{est} parameter based on the formula:

$$e = ||\theta_{gt} - \theta_{est}||^2 \tag{5.12}$$

The pose-to-pose Euclidean error for each round over different proportions of stable

[4] https://github.com/mattloper/chumpy

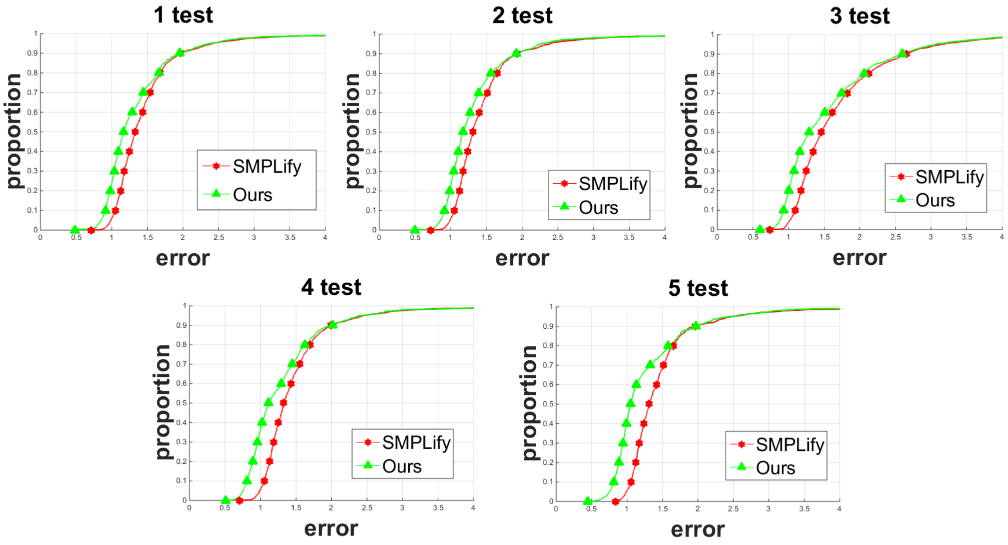


Figure 5.3: The comparison of the accuracy of the proposed method against SMPLify on stable poses.

pose dataset is shown in Figure 5.3 where the accuracy of the proposed method is stably higher than SMPLify in all rounds. One million pose data is used in SMPLify to train the GMM model while only 30000 pose data is used in this research, achieving more accurate pose estimation. SMPLify uses a more general pose prior to favor possible poses over impossible ones. However, given the 2D joints of images, there would be cases that several different 3D joints have the same 2D projection. Thus, the pose prior trained with a wide range of poses cannot prevent this case.

Besides comparison on stable poses, the performance of the proposed stable pose prior is also compared with the performance of general pose prior for random poses. 1000 poses are chosen from CMU dataset randomly for the comparison. As it can be seen in Figure 5.4, on random poses, the accuracy of results achieved with the stable pose prior is comparable with those achieved with random poses.

More results are visualized to show the comparison intuitively in Figure 5.5. Obviously, with the help of the stable pose prior, the estimated results are in a more natural "standing straight" pose. As it can be seen from the side views (columns 3, 6, 9 and 12 in

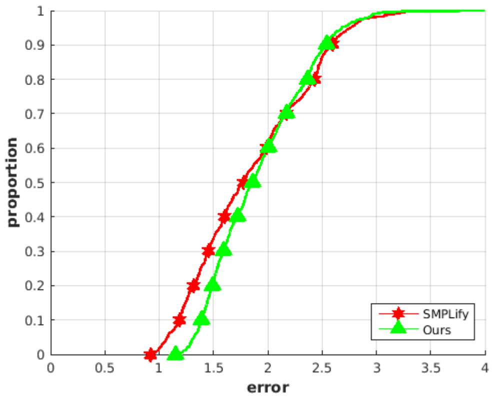


Figure 5.4: The comparison of the accuracy of our method and SMPlify on random poses.

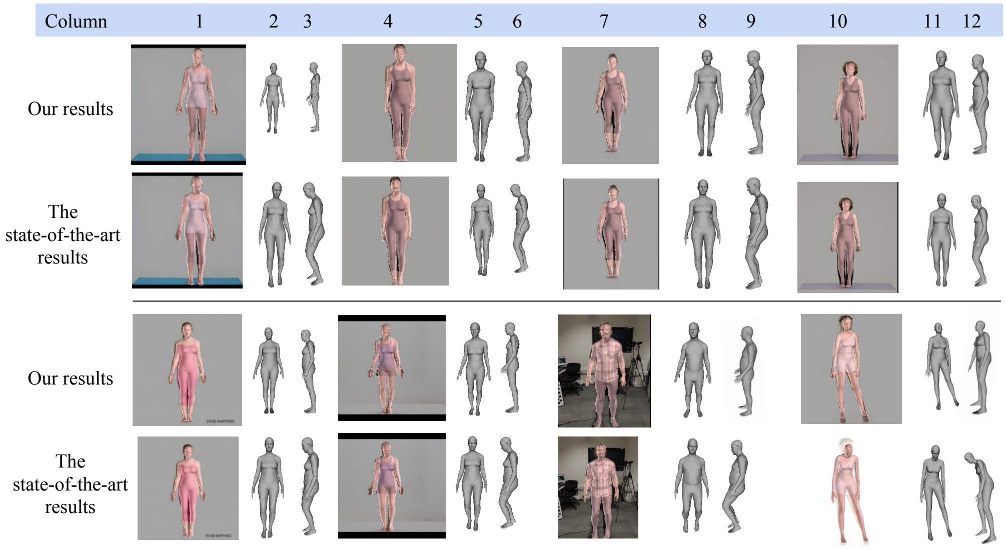


Figure 5.5: The qualitative comparison of Our results and SMPLify results.

Figure 5.5), results of the proposed method are more consistent with the original images while SMPLify results often present bent knees. In the experiments, it is shown that the both the SMPLify results with “bent knee” and our results that “stand straight” have joints projected to the same positions on 2D planes. The comparison presented in Figure 5.5 verifies that given general pose priors, only 2D joints of a single image cannot provide enough constraints on reasonable pose estimations in the stable states. The introduction of stable pose prior manages to favor possible stable poses over impossible ones (like bending knees and stoop in the state of standing). Moreover, due to the constraints of boundary, the estimated body shape from the proposed approach resembles the shape presented in the 2D images, which can be seen from the degrees of overlap of our results and SMPLify results (shown in column 1, 4, 7 and 10).

5.4.2 Evaluation on Human Body Measurement

In the above, the accuracy of pose estimation is evaluated. The shoulder breadth, breast, waist and hip circumferences of the estimated model are compared with those measured by the professionals. These four measurements are the key elements in the scenario of virtual try-on. Therefore, to quantitatively compare the accuracy of shape estimation, a dataset of real human is collected together with their actual measurements. This physical data is measured by professionals and used as the ground truth in the evaluation of the proposed methods. Each data sample contains three image of a real person in the front, side and back views, and the corresponding body measurements. In the experiment, only the image in the from view is exploited for human body estimation. The results of our proposed approaches are compared with other three methods: (1) SMPLify, (2) the method, JBG, that uses both joints and boundary cues under general pose prior and (3) the method, JS, that only exploits joints under stable pose prior. It is shown that the proposed method manages to provide more accurate estimation of human body shape and pose.

Human body measurement dataset The dataset contains 13 subjects (7 male and 6 female) with ages ranging from 18 to 50. All subjects wear tight clothes and pictures are taken from the front view, side view and back view. Besides, the designed virtual garments are also included in the dataset. An example of the dataset is given in Figure 5.6. As these garments are designed with known real human body measurements and fit on the models of real people. They can be used as the synthetic data for building the parametric clothing model in the future that can estimate cloth models given human body shape and pose parameters.

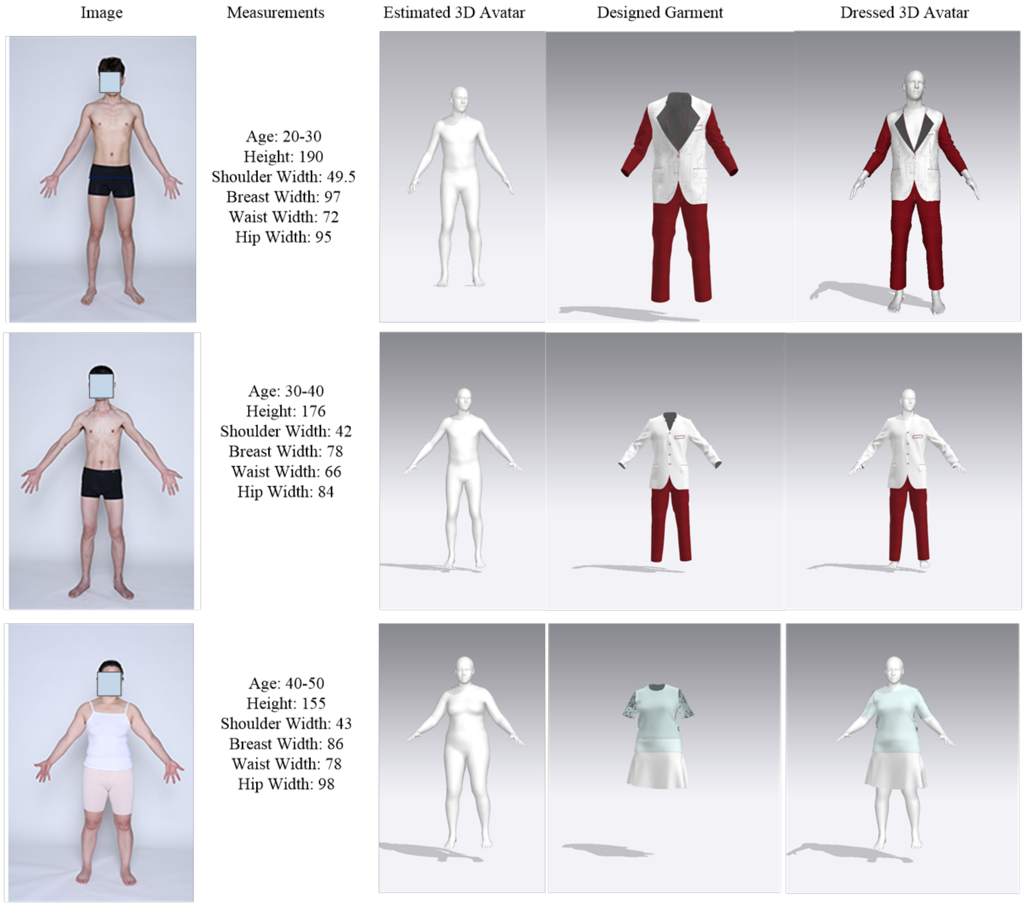


Figure 5.6: Some examples of the human body measurement dataset. It contains 2D image of real people, the actual body measurements, the estimated 3D avatar and the designed garments.

In the experiment, all the methods only use the single image with the front view. The same normal measurement means are adopted to measure the physical data for the subjects and the estimated models. Four elements are compared: Shoulder, Breast, Waist and Hip. For shoulder width, the distance between the right and left shoulder points is measured. For breast width, the circumference of the circle passing though the bustpoint is measured. The waist length is the circumference of the smallest circle in the waist part and the hip length is opposite, that is, the biggest circle in the hip part.

The actual measurements which are obtained from the subjects are used as ground truth and it is compared with the corresponding measurements of estimated meshes. The errors between the ground truth and the measured four elements for each method are shown in Table 5.1. The average errors of the four elements are calculated for comparing the overall accuracy.

Table 5.1: The quantitative comparison of accuracy of mesh estimation for each gender. For each gender, the first row is the measurement error of JBG; the second row is the error of the proposed approach; the third row is the error of SMPLify and the last row is JS measurement error.

Gender	Methods	Breast	Waist	Hip	Shoulder	Mean
Male	JBG	10.2418	13.1592	8.1640	6.4835	9.5121
	The proposed	7.7986	12.8607	6.9345	5.2428	**8.2091**
	SMPLify	5.0695	19.999	12.3443	15.7701	13.295
	JS	8.4562	14.9250	9.4884	13.3020	11.5429
Female	JBG	10.3607	13.4250	10.749	6.8181	10.3382
	The proposed	8.3977	11.9595	9.1784	7.6265	**9.2905**
	SMPLify	8.7196	15.9521	16.3209	12.0751	13.2669
	JS	8.8454	15.5266	13.1670	12.2902	12.4573

As it can be seen in the "Mean" column in Table 5.1, for both the male and female, boundary information manages to improve the accuracy of shape estimation significantly. For each gender, the mean errors of *JBG* and *"The proposed"* which add boundary constraints are lower than *SMPLify* and *JS* that only 2D joints are used. For the male, compared with *SMPLify* method, the average error of *"The proposed"* method is decreased by 38.25% and the error of the *JBG* method is decreased by 28.45%. For the female, the two corresponding errors are decreased by 29.97% and 22.07% respectively. It should also be noticed that the stable pose prior can help improve the shape estimation accuracy slightly by comparing *JBG* and *"The proposed"*. *JBG* uses general pose prior and *"The proposed"* uses stable pose prior.

In order to demonstrate the effect of the accurate estimation of human body shape on the virtual dress fitting, several types of clothes are predefined according to the human body measurements and fitted on the estimated human body models. The different looks on two estimated body shapes are shown in Figure 5.7. As the clothes are designed according to the true physical data measured by the professionals, the clothes should fit the body well if the body estimation is precise. However, it is obvious to see that the second column presents better fitting than SMPLify's results do. This suggests that boundary information is beneficial in providing a more accurate shape estimation and the resultant models are more closer to the actual body shape of real people and thus are more suitable for virtual dressing.

5.5 Virtual Dressing Applications

The human body shape and pose estimation plays an important role in many potential applications. One of the most typical scenarios is related to clothing. The approaches of accurate estimation of human body from a single image have been deeply investigated in this chapter. Here, two practical and pressingly needed applications are presented that require the accurate and personalized estimation of human body models.

Personalized Online Shopping Currently, most clothes are pre-made, i.e., manufactured and retailed according to a set of pre-defined sizes based on a rough range of chest circumference, hip circumference and waist measurements. However, due to the diversity of human body shapes, the garment's fit on a model does not necessarily reflect its look on the customer even in its best-fit size. Customers are often dissatisfied by a pre-made garment even if much time and efforts are investigated in searching and fitting.

Leveraging the capability of the proposed method, we can produce plausible estimation of human body models which resemble the human body in a 2D image and use them for virtual try-on in online shopping applications. In order to show the potential application in virtual clothing try-on application, the estimated model with the proposed method is dressed with clo3d and some samples are visualized in Figure 5.8. Clo3d [5] is a 3D fashion design software program that creates virtual, true-to life garment visualization for the fashion and apparel industries. The faces of images have been blurred for the reason of privacy protection.

Online Bespoke Clothing Design and Manufacture Traditionally, fashion consultancy,

[5] https://www.clo3d.com/

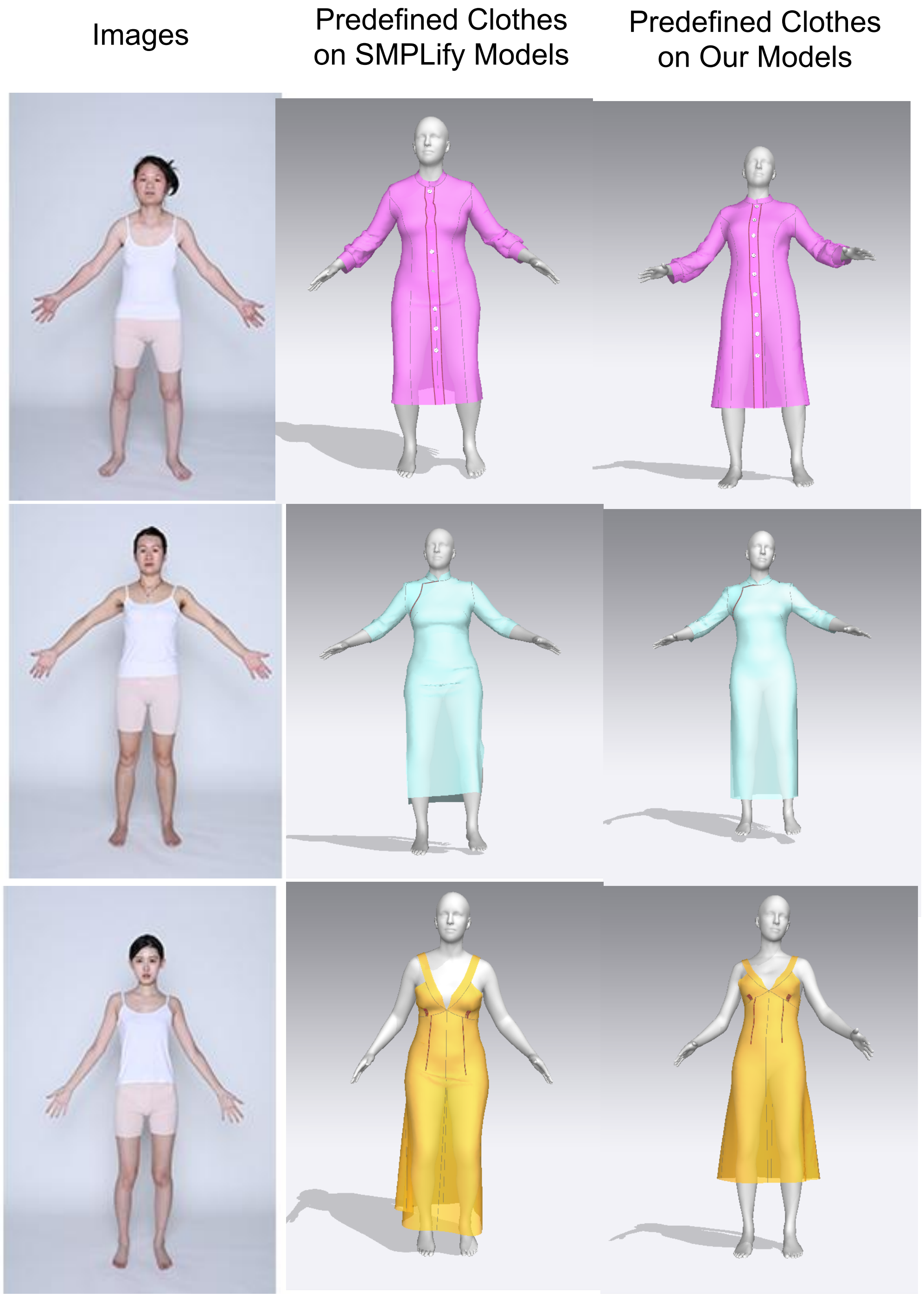

Images
Predefined Clothes on SMPLify Models
Predefined Clothes on Our Models

Figure 5.7: The visualization of predefined clothes model on estimated human body models with SMPLify and our method.

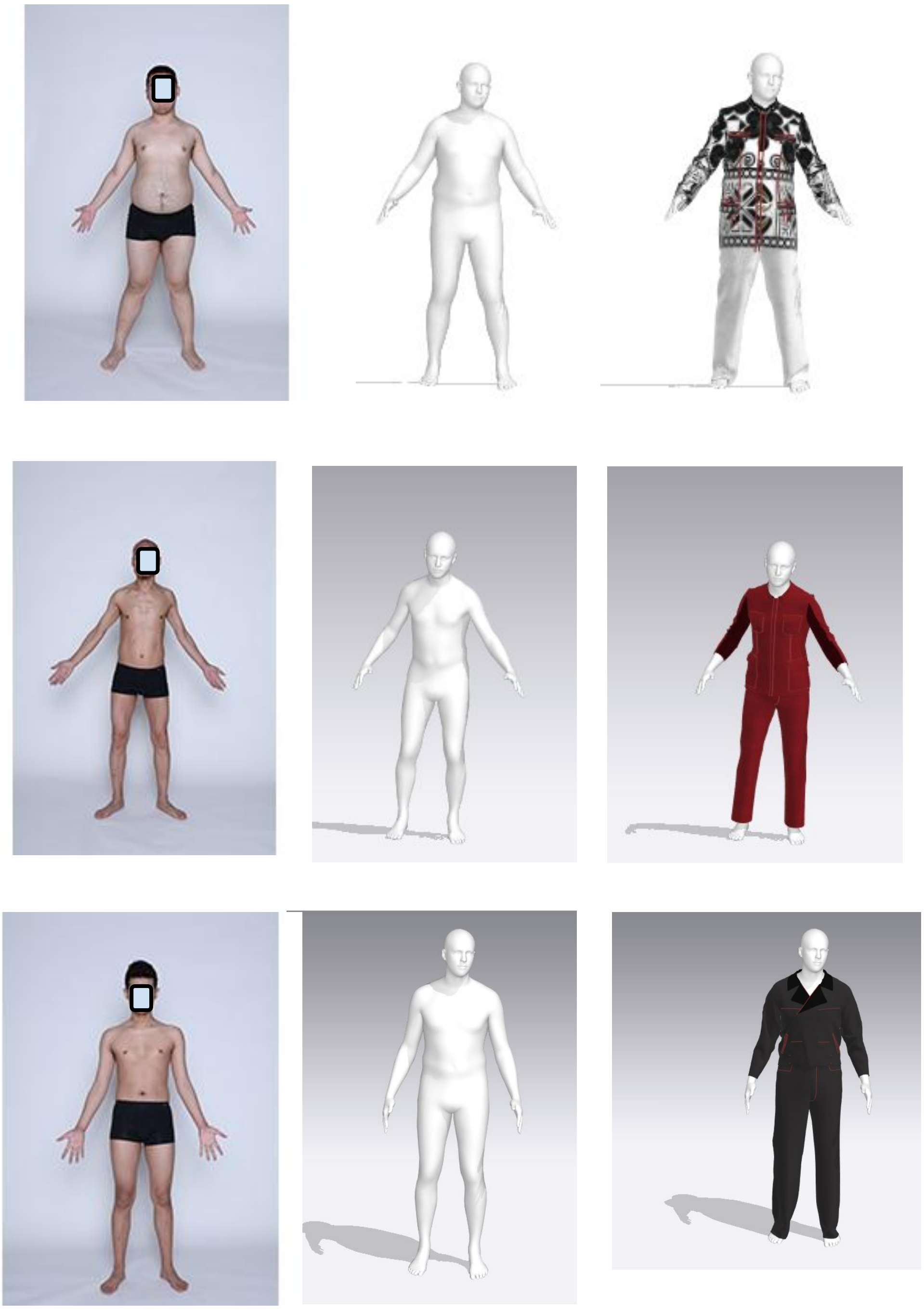


Figure 5.8: Examples of virtual on believable human body estimations.

bespoke apparel design and production can provide consumers with much more satisfying experience, but are exclusive for few people who can afford such premium services. To improve the satisfaction of more customers, the reduction of the cost of bespoke services is vital. Here, the precise estimation of body shape with the proposed method based on a single image can be very useful to reduce the cost of bespoken fitting, and furthermore, to collect body shape data for design and recommendation related applications using big-data and AI technologies. Moreover, having the realistically estimated 3D virtual model, it is then possible to fit the virtual garment on the model, and virtually adjust the tailoring, the colour, the pattern or even the design of styles, accordingly to achieve the best fitting result. With the number of consumers involved in this application increasing, more data regarding body shapes, personal preferences, measurements will be available for big data analysis for fashion recommendation and improvement of consuming experience.

5.6 Summary

In this chapter, an approach is presented for accurate human body shape and pose estimation using the boundary and joints of a single image. The stable pose prior is trained and boundary constraints are added to tackle this ill-posed problem for more accurate estimation. Experiments show that the pose estimation of the proposed approach is more accurate than the state-of-the-art with only one percent of the training set of SMPLify and the boundary information improves the accuracy of shape estimation significantly. It is also demonstrated that the proposed approach has various interesting applications, including personalised virtual try-on, or online inexpensive bespoke clothing design and manufacture.

Chapter 6

Conclusions and Future Work

In this chapter, the research work presented in this thesis is summarized and the important results are highlighted. This chapter is organised according to the overall thesis structure and finishes with a view of exploring a few potential future research directions in the field.

6.1 Summary of Contributions

In this thesis, the challenge of accurately estimating high-quality 3D human body models from user-generated data including low-quality point clouds and single images is addressed.

The 3D scanning technology has evolved in the past decades at an incredible pace and emerging commodity 3D scanners prompt the reconstruction, modelling and displaying of scenes, objects and human bodies. New introduced technologies and products make the acquisition of 3D human body models accessible to general people at a low cost. However, the convenience of 3D model acquisition comes with low quality of the captured scans. Big holes resulting from occlusion or self - occlusions, distortion caused by the movement of subjects or scanning turbulence, noises and outliers pose great challenges for building high-quality human body meshes from the raw scans.

Based on this observation, in chapter 3, a symmetry-aware approach is firstly proposed to find shape correspondences between two closed meshes. A major outcome of chapter 3 is the development of a novel approach for addressing the symmetric flips problem for closed human body shapes and improving the accuracy of the finally located correspondences. The contribution of the proposed shape-aware approach for establishing shape correspondence are listed below:

- skeleton information is integrated to robustly address symmetric flip problem which still exists in state-of-the-art techniques for closed mesh;
- the base vertex set is built to refine the final one-to-one correspondence and achieve higher accuracy.

It is shown that skeleton integrated with intrinsic shape features can effectively remove all the symmetric flipped correspondences. The base vertex set can also help locate the more accurate final shape correspondences in the extended correspondence candidates.

Although the proposed correspondence establishment approach improves the accuracy for closed meshes, it is challenging to extract accurate features for user-generated scans where holes, noises and distortion are prevalent. Therefore, a robust non-rigid shape registration method is proposed in Chapter 4. The proposed registration method combines the prior knowledge with traditional non-rigid iterative closest point algorithm, greatly improving the robustness and enabling the method to handle low-quality scanning data. Specifically, the contributions of the proposed human body mesh registration approach in Chapter 4 are summarized as follows.

- For low-quality scans in the same poses: A fully automatic registration method which performs well even on noisy low quality data is proposed. The proposed method follows the region-based approach to register the human body scans, which improves the accuracy of registration. According to the nature of different body parts, the approach adopts particular registration strategies, which makes the method robust to noisy Kinect scans.

- For low-quality scans in different poses: the above registration method is extended by using sparse landmarks to drive the template to perform in the same way as the target scans do. In this way, human body scans with different poses can be robustly registered. The number of landmarks used is 16, which is much sparser than the state-of-the-art works which use at least 44 landmarks.
- A dataset of 255 real human body scans with Kinect is built. 55 scans of people from Europe and Asia are captured and each of them are required to perform 5 predefined poses. To the best of our knowledge, K3D-Hub is the first dataset of 3D human body captured by Kinect which can be used to evaluate the robustness of registration algorithms in case of low quality scans.

Extensive experiments showed that the proposed novel multilevel non-rigid registration framework is robust to holes, noises and distorted parts presented in the low-quality human body scans. With the help of a sparse set of landmarks, the proposed registration approach is successfully extended into the case of various poses.

Apart from 3D scanning data captured with commodity scanners, another unignorable and more common type of user-generated data - 2D images - is also considered. With the rapid development of smart phone in the past decades, 2D images are the quickest and easiest data source we can acquire. Thus, besides building high-fidelity human body models from low-quality raw scans, in this research, the estimation of high-fidelity human body models from single 2D images is also investigated in Chapter 5.

The contributions of this research on human body shape and pose estimation from single 2D images are two-fold:

- First, an approach for automatic and accurate human body shape and pose estimation from a single image is proposed. Boundary information and stable pose prior are exploited to improve the accuracy of estimation.

- Second, a dataset is collected that contains 13 humans (7 male and 6 female) images and their body measurements and quantitative evaluation of the proposed method is performed on this dataset. Several types of clothes for each model are also provided. The virtual clothes are made by professionals with clo3d according to the actual measurements of models. To the best of our knowledge, there is no public dataset that contains 2D human body images, the corresponding human body measurements (i.e. height, weight and waist, hip and bust circumference) and clothing models. This dataset can be used for quantitatively evaluating the human body shape estimation from single images as well as for clothing related research work, like parametric clothing modelling or cloth shape analysis et al.

According to the above conclusions, it has become clear that the prior knowledge of vertex movement in the human body shape and pose space is the key point for robust high-fidelity human body modelling from low-quality raw scans or 2D images. Experiments have demonstrated that the combination of human shape model with traditional non-rigid ICP greatly improved the robustness. The proposed method is robust to the big holes, noises and artefacts on the surface and also tackles the local minimum problem commonly occurs in the hand / foot parts. The human body model can also be used to estimate the human body models given 2D joints and boundaries.

6.2 Future Work

It is believed that some possible future directions of the subsequent works can be investigated to improve the proposed methods and are summarized in the following areas.

- As can be seen in Chapter 4, landmarks are annotated manually to help the registration of target scans in various poses. Although only 16 landmarks are adopted to guide the deformation of template in the initial step, it is desirable to automatically establish the keypoints to replace manually-annotated landmarks based on extracting features. Currently emerged deep learning architectures like PointNet [132], PointNet++ [132] and 3DMatch [195] are proposed to process 3D data in the representation of point clouds and voxels. However, the architectures that can accept the triangular mesh format can be investigated further to extract features on the mesh end to end in the future.

- Currently, the approaches in this research are limited to the estimation of the naked shape of human body. This is caused by the different topology of human body and clothes. In the future, the accurate estimation of humans wearing clothes is worth to be further studied.

- Another interesting future direction is to model the clothes of different types like T-shirt, pants and coat et. al. according to the way it wears in different human body shapes and poses. The clothing model can be adjusted by the shape and pose parameters of human body. In this way, the size of the clothes can be adjusted continuously according to human body shape and poses. Although there are some works [56, 129] on clothing modelling, they are usually confined to limited types of clothes or synthetic clothes are used to build statistic model which can not reflect the realistic movements of clothes.

Bibliography

[1] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. *arXiv preprint arXiv:1803.04758*, 2018.

[2] 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.

[3] B. Allen, B. Curless, Z. Popović, and A. Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 147–156. Eurographics Association, 2006.

[4] A. Amamra and N. Aouf. Gpu-based real-time rgbd data filtering. *Journal of Real-Time Image Processing*, 14(2):323–340, 2018.

[5] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[6] 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, 2005.

[7] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point

sets. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):698–700, 1987.

[8] A. Aubel. Anatomically-based human body deformations. *Doktorarbeit an der Ecole Polytechnique Féderale de Lausanne*, 2002.

[9] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[10] I. Baran and J. Popović. Automatic rigging and animation of 3d characters. In *ACM Transactions on Graphics (TOG)*, volume 26, page 72. ACM, 2007.

[11] R. J. Barnes. Matrix differentiation. *Springs Journal*, pages 1–9, 2006.

[12] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*, volume 1611, pages 586–607. International Society for Optics and Photonics, 1992.

[13] V. D. Bhise. *Ergonomics in the automotive design process.* CRC Press, 2016.

[14] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[15] F. Bogo, J. Romero, M. Loper, and M. Black. Faust: Dataset and evaluation for 3d mesh registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3794–3801, 2014.

[16] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2300–2308, 2015.

[17] 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.

[18] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic faust: Registering human bodies in motion. In *Proc. the Conference on Computer Vision and Pattern Recognition*, 2017.

[19] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, Piscataway, NJ, USA, July 2017. IEEE.

[20] E. Bondi, P. Pala, S. Berretti, and A. Del Bimbo. Reconstructing high-resolution face models from kinect depth sequences. *IEEE Transactions on Information Forensics and Security*, 11(12):2843, 2016.

[21] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172, 2006.

[22] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. *Numerical geometry of non-rigid shapes.* Springer Science & Business Media, 2008.

[23] M. M. Bronstein and I. Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1704–1711. IEEE, 2010.

[24] A. Brunton, T. Bolkart, and S. Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *Computer Vision–ECCV 2014*, pages 297–312. Springer, 2014.

[25] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016.

[26] M. Camplani and L. Salgado. Efficient spatio-temporal hole filling strategy for kinect depth maps. In *Three-dimensional image processing (3DIP) and applications Ii*, volume 8290, page 82900E. International Society for Optics and Photonics, 2012.

[27] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[28] W. Chang and M. Zwicker. Global registration of dynamic range scans for articulated model reconstruction. *ACM Transactions on Graphics (TOG)*, 30(3):26, 2011.

[29] G. Chen, J. Li, B. Wang, J. Zeng, G. Lu, and D. Zhang. Reconstructing 3d human models with a kinect. *Computer Animation and Virtual Worlds*, 2015.

[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[31] 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.

[32] S. Cheng, I. Marras, S. Zafeiriou, and M. Pantic. Active nonrigid icp algorithm. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[33] S. Cheng, I. Marras, S. Zafeiriou, and M. Pantic. Statistical non-rigid icp algorithm and its application to 3d face alignment. *Image and Vision Computing*, 2016.

[34] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 628–644. Springer, 2016.

[35] S. Clarkson, J. Wheat, B. Heller, and S. Choppin. Assessment of a microsoft kinect-based 3d scanning system for taking body segment girth measurements: a comparison to isak and iso standards. *Journal of sports sciences*, 34(11):1006–1014, 2016.

[36] Y. Cui, W. Chang, T. Nöll, and D. Stricker. Kinectavatar: fully automatic body capture using a single kinect. In *Asian Conference on Computer Vision*, pages 133–147. Springer, 2012.

[37] T. Darom and Y. Keller. Scale-invariant features for 3-d mesh models. *Image Processing, IEEE Transactions on*, 21(5):2758–2769, 2012.

[38] J. Davis, S. R. Marschner, M. Garr, and M. Levoy. Filling holes in complex surfaces using volumetric diffusion. In *3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*, pages 428–441. IEEE, 2002.

[39] G. De Magistris, A. Micaelli, J. Savin, C. Gaudez, and J. Marsot. Dynamic digital human models for ergonomic analysis based on humanoid robotics techniques. *Int. J. Digital Human*, 1(1):81–109, 2014.

[40] T. K. Dey, B. Fu, H. Wang, and L. Wang. Automatic posing of a meshed human model using point clouds. *Computers & Graphics*, 46:14–24, 2015.

[41] A. Divivier, R. Trieb, A. Ebert, H. Hagen, C. Gross, A. Fuhrmann, V. Luckas, et al. Virtual try-on topics in realistic, individualized dressing in virtual reality. 2004.

[42] F. Dong, G. J. Clapworthy, M. A. Krokos, and J. Yao. An anatomy-based approach to human muscle modeling and deformation. *Visualization and Computer Graphics, IEEE Transactions on*, 8(2):154–170, 2002.

[43] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi. 3d scanning deformable objects with a single rgbd sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 493–501, 2015.

[44] 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.

[45] A. Elad and R. Kimmel. On bending invariant signatures for surfaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(10):1285–1295, 2003.

[46] P. Fechteler, A. Hilsmann, and P. Eisert. Kinematic icp for articulated template fitting. In *Proceedings of International Workshop on Vision, Modeling and Visualization*, pages 12–14, 2012.

[47] M. Ferrant, A. Nabavi, B. Macq, F. A. Jolesz, R. Kikinis, and S. K. Warfield. Registration of 3-d intraoperative mr images of the brain using a finite-element biomechanical model. *IEEE transactions on medical imaging*, 20(12):1384–1397, 2001.

[48] A. W. Fitzgibbon. Robust registration of 2d and 3d point sets. *Image and vision computing*, 21(13-14):1145–1153, 2003.

[49] R. Gal and D. Cohen-Or. Salient geometric features for partial shape matching and similarity. *ACM Transactions on Graphics (TOG)*, 25(1):130–150, 2006.

[50] T. Gatzke, C. Grimm, M. Garland, and S. Zelinka. Curvature maps for local shape comparison. In *Shape Modeling and Applications, 2005 International Conference*, pages 244–253. IEEE, 2005.

[51] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, pages 484–499. Springer, 2016.

[52] M. Goesele, B. Curless, and S. M. Seitz. Multi-view stereo revisited. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2402–2409. IEEE, 2006.

[53] S. Granger and X. Pennec. Multi-scale em-icp: A fast and robust approach for surface registration. In *European Conference on Computer Vision*, pages 418–432. Springer, 2002.

[54] E. S. Grood and W. J. Suntay. A joint coordinate system for the clinical description of three-dimensional motions: application to the knee. *Journal of biomechanical engineering*, 105(2):136–144, 1983.

[55] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1381–1388. IEEE, 2009.

[56] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black. Drape: Dressing any person. *ACM Trans. Graph.*, 31(4):35–1, 2012.

[57] 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.

[58] D. Haehnel, S. Thrun, and W. Burgard. An extension of the icp algorithm for modeling nonrigid objects with mobile robots. In *IJCAI*, volume 3, pages 915–920, 2003.

[59] E. P. Hanavan Jr. A mathematical model of the human body. Technical report, AIR FORCE AEROSPACE MEDICAL RESEARCH LAB WRIGHT-PATTERSON AFB OH, 1964.

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

[61] 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, 2009.

[62] S. Hauswiesner, M. Straka, and G. Reitmayr. Virtual try-on through image-based rendering. *IEEE transactions on visualization and computer graphics*, 19(9):1552–1565, 2013.

[63] Q. He, Y. Ji, D. Zeng, and Z. Zhang. Volumeter: 3d human body parameters measurement with a single kinect. *IET Computer Vision*, 12(4):553–561, 2018.

[64] Z. He, F. Zhu, and K. Perlin. Physhare: Sharing physical interaction in virtual reality. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 17–19. ACM, 2017.

[65] P. Henzler, V. Rasche, T. Ropinski, and T. Ritschel. Single-image tomography: 3d volumes from 2d x-rays. *arXiv preprint arXiv:1710.04867*, 2017.

[66] G. Hinton. Using relaxation to find a puppet. In *Proceedings of the 2nd Summer Conference on Artificial Intelligence and Simulation of Behaviour*, pages 148–157. IOS Press, 1976.

[67] D. A. Hirshberg, M. Loper, E. Rachlin, and M. J. Black. Coregistration: Simultaneous alignment and modeling of articulated 3d shape. In *Computer Vision–ECCV 2012*, pages 242–255. Springer, 2012.

[68] Y. Hong, P. Bruniaux, X. Zeng, K. Liu, Y. Chen, and M. Dong. Virtual reality-based collaborative design method for designing customized garment for disabled people with scoliosis. *International Journal of Clothing Science and Technology*, 29(2):226–237, 2017.

[69] B. K. Horn. Closed-form solution of absolute orientation using unit quaternions. *JOSA A*, 4(4):629–642, 1987.

[70] B. K. Horn, H. M. Hilden, and S. Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *JOSA A*, 5(7):1127–1135, 1988.

[71] X. Hu, Y. Wang, F. Zhu, and C. Pan. Learning-based fully 3d face reconstruction from a single image. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 1651–1655. IEEE, 2016.

[72] Q.-X. Huang, B. Adams, M. Wicke, and L. J. Guibas. Non-rigid registration under isometric deformations. In *Computer Graphics Forum*, volume 27, pages 1449–1457. Wiley Online Library, 2008.

[73] 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.

[74] A. Jacobson, I. Baran, J. Popovic, and O. Sorkine. Bounded biharmonic weights for real-time deformation. *ACM Trans. Graph.*, 30(4):78–1, 2011.

[75] V. Jain and H. Zhang. Robust 3d shape correspondence in the spectral domain. In *Shape Modeling and Applications, 2006. SMI 2006. IEEE International Conference on*, pages 19–19. IEEE, 2006.

[76] W. Jia, W.-J. Yi, J. Saniie, and E. Oruklu. 3d image reconstruction and human body tracking using stereo vision and kinect technology. In *Electro/Information Technology (EIT), 2012 IEEE International Conference on*, pages 1–4. IEEE, 2012.

[77] H. Jin, S. Soatto, and A. J. Yezzi. Multi-view stereo reconstruction of dense shape and complex appearance. *International Journal of Computer Vision*, 63(3):175–189, 2005.

[78] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):433–449, 1999.

[79] 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.

[80] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 38–44. IEEE, 1996.

[81] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE transactions on pattern analysis and machine intelligence*, 33(2):394–405, 2011.

[82] K. Kilteni, I. Bergstrom, and M. Slater. Drumming in immersive virtual reality: the body shapes the way we play. *IEEE Transactions on Visualization & Computer Graphics*, (4):597–605, 2013.

[83] V. G. Kim, Y. Lipman, and T. Funkhouser. Blended intrinsic maps. In *ACM Transactions on Graphics (TOG)*, volume 30, page 79. ACM, 2011.

[84] M. Körtgen, G.-J. Park, M. Novotni, and R. Klein. 3d shape matching with 3d shape contexts. In *The 7th central European seminar on computer graphics*, volume 3, pages 5–17, 2003.

[85] S. Kottner, L. C. Ebert, G. Ampanozi, M. Braun, M. J. Thali, and D. Gascho. A mobile, multi-camera setup for 3d full body imaging in combination with postmortem computed tomography procedures. In *Proceedings of the 7th International Conference on 3D Body Scanning Technologies*, 2016.

[86] Q. Kou, Y. Yang, S. Du, S. Luo, and D. Cai. A modified non-rigid icp algorithm for registration of chromosome images. In *International Conference on Intelligent Computing*, pages 503–513. Springer, 2016.

[87] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218, 2000.

[88] 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.

[89] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3d self-portraits. *ACM Transactions on Graphics (TOG)*, 32(6):187, 2013.

[90] W. Li, X. Li, M. Goldberg, and Z. Zhu. Face recognition by 3d registration for the visually impaired using a rgb-d sensor. In *European Conference on Computer Vision*, pages 763–777. Springer, 2014.

[91] X. Li and I. Guskov. Multiscale features for approximate alignment of point-based surfaces. In *Symposium on geometry processing*, volume 255, page 217. Citeseer, 2005.

[92] Z. Liu, J. Huang, S. Bu, J. Han, X. Tang, and X. Li. Template deformation-based 3-d reconstruction of full human body scans from low-cost depth cameras. *IEEE transactions on cybernetics*, 2016.

[93] Z. Liu, J. Huang, S. Bu, J. Han, X. Tang, and X. Li. Template deformation-based 3-d reconstruction of full human body scans from low-cost depth cameras. *IEEE transactions on cybernetics*, 47(3):695–708, 2017.

[94] M. Loper, N. Mahmood, and M. J. Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220, 2014.

[95] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6): 248, 2015.

[96] M. M. Loper and M. J. Black. Opendr: An approximate differentiable renderer. In *European Conference on Computer Vision*, pages 154–169. Springer, 2014.

[97] M. M. Loper, N. Mahmood, and M. J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, Nov. 2014. doi: 10.1145/2661229.2661273. URL `http://doi.acm.org/10.1145/2661229.2661273`.

[98] A. Lorusso, D. W. Eggert, and R. B. Fisher. *A comparison of four algorithms for estimating 3-D rigid transformations*. University of Edinburgh, Department of Artificial Intelligence, 1995.

[99] Z. Lun, M. Gadelha, E. Kalogerakis, S. Maji, and R. Wang. 3d shape reconstruction from sketches via multi-view convolutional networks. *arXiv preprint arXiv:1707.06375*, 2017.

[100] D. Lupton. Fabricated data bodies: Reflections on 3d printed digital body objects in medical and health domains. *Social Theory & Health*, 13(2):99–115, 2015.

[101] T. Masuda, K. Sakaue, and N. Yokoya. Registration and integration of multiple range images for 3-d model construction. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, volume 1, pages 879–883. IEEE, 1996.

[102] D. Mateus, R. Horaud, D. Knossow, F. Cuzzolin, and E. Boyer. Articulated shape matching using laplacian eigenfunctions and unsupervised point registration. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[103] D. McMakin, D. Sheen, T. Hall, J. Tedeschi, and A. M. Jones. New improvements to millimeter-wave body scanners. *Proceedings of 3DBODY. TECH*, 2017.

[104] C. Moore, T. Duckworth, R. Aspin, and D. Roberts. Synchronization of images from multiple cameras to reconstruct a moving human. In *2010 IEEE/ACM 14th International Symposium on Distributed Simulation and Real Time Applications*, pages 53–60. IEEE, 2010.

[105] T. M. Mora, J. Quelen, O. D. Escoda, C. F. Bernstrom, R. P. Scasso, D. B. Diez, S. A. Duart, and F. M. Burgos. Method and a system for generating a realistic 3d reconstruction model for an object or being, June 25 2015. US Patent App. 14/402,999.

[106] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1561–1570. IEEE, 2017.

[107] J. K. Murthy, G. S. Krishna, F. Chhaya, and K. M. Krishna. Reconstructing vehicles from a single image: Shape priors for road scene understanding. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 724–731. IEEE, 2017.

[108] P. Narayanan, P. W. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. In *Computer Vision, 1998. Sixth International Conference on*, pages 3–10. IEEE, 1998.

[109] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011.

[110] 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.

[111] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[112] M. Obentheuer, M. Roller, S. Björkenstam, K. Berns, and J. Linn. Human like motion generation for ergonomic assessment-a muscle driven digital human model using muscle synergies. In *Proceedings of the 8th ECCOMAS Thematic Conference on Multibody Dynamics*, pages 847–856, 2017.

[113] E. Olson and P. Agarwal. Inference on networks of mixtures for robust robot mapping. *The International Journal of Robotics Research*, 32(7):826–840, 2013.

[114] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 741–754. ACM, 2016.

[115] M. Ovsjanikov, J. Sun, and L. Guibas. Global intrinsic symmetries of shapes. In *Computer graphics forum*, volume 27, pages 1341–1348. Wiley Online Library, 2008.

[116] M. Ovsjanikov, Q. Mérigot, F. Mémoli, and L. Guibas. One point isometric matching with the heat kernel. In *Computer Graphics Forum*, volume 29, pages 1555–1564. Wiley Online Library, 2010.

[117] 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.

[118] M. Ovsjanikov, Q. Mérigot, V. Pătrăucean, and L. Guibas. Shape matching via quotient spaces. In *Computer Graphics Forum*, volume 32, pages 1–11. Wiley Online Library, 2013.

[119] P. Palasek, H. Yang, Z. Xu, N. Hajimirza, E. Izquierdo, and I. Patras. A flexible calibration method of multiple kinects for 3d human reconstruction. In *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*, pages 1–4. IEEE, 2015.

[120] S. Paulus, H. Schumann, H. Kuhlmann, and J. Léon. High-precision laser scanning system for capturing 3d plant architecture and analysing growth of cereal plants. *Biosystems Engineering*, 121:1–11, 2014.

[121] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901.

[122] M. Peters, E. Quadrat, A. Nolte, A. Wolf, J. Miehling, S. Wartzack, W. Leidholdt, S. Bauer, L. Fritzsche, and S. Wischniewski. Biomechanical digital human models: Chances and challenges to expand ergonomic evaluation. In *International Conference on Human Systems Engineering and Design: Future Trends and Applications*, pages 885–890. Springer, 2018.

[123] S. Pheasant. *Bodyspace: Anthropometry, Ergonomics And The Design Of Work: Anthropometry, Ergonomics And The Design Of Work*. CRC Press, 2014.

[124] U. Pinkall and K. Polthier. Computing discrete minimal surfaces and their conjugates. *Experimental mathematics*, 2(1):15–36, 1993.

[125] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3d human modeling. *arXiv preprint arXiv:1503.05860*, 2015.

[126] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016.

[127] 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.

[128] 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.

[129] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics,(Proc. SIGGRAPH)[to appear]*, 1, 2017.

[130] I. Pratikakis, M. Spagnuolo, T. Theoharis, and R. Veltkamp. A robust 3d interest points detector based on harris operator. In *Eurographics workshop on 3D object retrieval*, volume 5. Citeseer, 2010.

[131] C. Pu, N. Li, and R. B. Fisher. Robust rigid point registration based on convolution of adaptive gaussian mixture models. *arXiv preprint arXiv:1707.08626*, 2017.

[132] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1(2):4, 2017.

[133] M. P. Reed, U. Raschke, R. Tirumali, and M. B. Parkinson. Developing and implementing parametric human body shape models in ergonomics software. In *Proceedings of the 3rd international digital human modeling conference, Tokyo*, 2014.

[134] M. Reuter, S. Biasotti, D. Giorgi, G. Patanè, and M. Spagnuolo. Discrete laplace–beltrami operators for shape analysis and segmentation. *Computers & Graphics*, 33(3):381–390, 2009.

[135] K. M. Robinette, H. Daanen, and E. Paquet. The caesar project: a 3-d surface anthropometry survey. In *3-D Digital Imaging and Modeling, 1999. Proceedings. Second International Conference on*, pages 380–386, 1999. doi: 10.1109/IM.1999.805368.

[136] K. M. Robinette, H. Daanen, and E. Paquet. The caesar project: a 3-d surface anthropometry survey. In *3-D Digital Imaging and Modeling, 1999. Proceedings. Second International Conference on*, pages 380–386. IEEE, 1999.

[137] N. Robitaille, P. L. Jackson, L. J. Hébert, C. Mercier, L. J. Bouyer, S. Fecteau, C. L. Richards, and B. J. McFadyen. A virtual reality avatar interaction (vrai) platform to assess residual executive dysfunction in active military personnel with previous mild traumatic brain injury: proof of concept. *Disability and Rehabilitation: Assistive Technology*, 12(7):758–764, 2017.

[138] J. S. Roo and M. Hachet. Towards a hybrid space combining spatial augmented reality and virtual reality. In *3D User Interfaces (3DUI), 2017 IEEE Symposium on*, pages 195–198. IEEE, 2017.

[139] J. Rugis and R. Klette. Surface curvature extraction for 3d image analysis or surface rendering.

[140] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152. IEEE, 2001.

[141] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3d model acquisition. *ACM Transactions on Graphics (TOG)*, 21(3):438–446, 2002.

[142] R. M. Rustamov. Laplace-beltrami eigenfunctions for deformation invariant shape representation. In *Proceedings of the fifth Eurographics symposium on Geometry processing*, pages 225–233. Eurographics Association, 2007.

[143] Y. Sahillioğlu 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.

[144] Y. Sahillioglu and Y. Yemez. Minimum-distortion isometric shape correspondence using em algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2203–2215, 2012.

[145] Y. Sahillioğlu and Y. Yemez. Coarse-to-fine isometric shape correspondence by tracking symmetric flips. In *Computer Graphics Forum*, volume 32, pages 177–189. Wiley Online Library, 2013.

[146] Y. Sahilliolu and Y. Yemez. Coarse-to-fine combinatorial matching for dense isometric shape correspondence. In *Computer Graphics Forum*, volume 30, pages 1461–1470. Wiley Online Library, 2011.

[147] G. Sandbach, S. Zafeiriou, and M. Pantic. Local normal binary patterns for 3d facial action unit detection. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1813–1816. IEEE, 2012.

[148] F. Scheepers, R. E. Parent, W. E. Carlson, and S. F. May. Anatomy-based modeling of the human musculature. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 163–172. ACM Press/Addison-Wesley Publishing Co., 1997.

[149] F. Schmidt. The laplace-beltrami-operator on riemannian manifolds. In *Seminar Shape Analysis*, 2014.

[150] D. C. Schneider and P. Eisert. Fast nonrigid mesh registration with a data-driven deformation prior. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 304–311. IEEE, 2009.

[151] P. H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[152] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999.

[153] J. Shen and S.-C. S. Cheung. Layer depth denoising and completion for structured-light rgb-d cameras. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1187–1194. IEEE, 2013.

[154] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *International journal of computer vision*, 98(1):15–48, 2012.

[155] 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.

[156] D. Smeets, J. Keustermans, D. Vandermeulen, and P. Suetens. meshsift: Local surface features for 3d face recognition under expression variations and partial data. *Computer Vision and Image Understanding*, 117(2):158–169, 2013.

[157] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *The International Journal of Robotics Research*, 22(6): 371–391, 2003.

[158] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 175–184. ACM, 2004.

[159] R. P. Spicer, S. M. Russell, and E. S. Rosenberg. The mixed reality of things: emerging challenges for human-information interaction. In *Next-Generation Analyst V*, volume 10207, page 102070A. International Society for Optics and Photonics, 2017.

[160] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 951–958. IEEE, 2011.

[161] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.

[162] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.

[163] 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. *Visualization and Computer Graphics, IEEE Transactions on*, 19(7):1199–1217, 2013.

[164] Q. Tan, L. Gao, Y.-K. Lai, J. Yang, and S. Xia. Mesh-based autoencoders for localized deformation component analysis. *arXiv preprint arXiv:1709.04304*, 2017.

[165] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao. Histogram of oriented normal vectors for object recognition with a depth sensor. In *Computer Vision–ACCV 2012*, pages 525–538. Springer, 2012.

[166] J. R. Tena, F. De la Torre, and I. Matthews. Interactive region-based linear 3d face models. In *ACM SIGGRAPH*, pages 76:1–76:10, New York, NY, USA, 2011. ISBN 978-1-4503-0943-1. doi: 10.1145/1964921.1964971. URL http://doi.acm.org/10.1145/1964921.1964971.

[167] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt.

Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6): 183–1, 2015.

[168] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3d full human bodies using kinects. *IEEE transactions on visualization and computer graphics*, 18(4):643–650, 2012.

[169] A. Tsoli, M. Loper, and M. J. Black. Model-based anthropometry: Predicting measurements from 3d human scans in multiple poses. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 83–90. IEEE, 2014.

[170] G. Turk and M. Levoy. Zippered polygon meshes from range images. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 311–318. ACM, 1994.

[171] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (4):376–380, 1991.

[172] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from Synthetic Humans. In *CVPR*, 2017.

[173] M. W. Walker, L. Shao, and R. A. Volz. Estimating 3-d location parameters using dual number quaternions. *CVGIP: image understanding*, 54(3):358–367, 1991.

[174] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2361–2368, 2014.

[175] T. Weber, R. Hänsch, and O. Hellwich. Automatic registration of unordered point clouds acquired by kinect sensors using an overlap heuristic. *ISPRS Journal of Photogrammetry and Remote Sensing*, 102:96–109, 2015.

[176] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. *arXiv preprint arXiv:1602.00134*, 2016.

[177] S. Weik. Registration of 3-d partial surface models using luminance and depth information. In *3-D Digital Imaging and Modeling, 1997. Proceedings., International Conference on Recent Advances in*, pages 93–100. IEEE, 1997.

[178] T. Weise, B. Leibe, and L. Van Gool. Fast 3d scanning with automatic motion compensation. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[179] 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.

[180] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald. Kintinuous: Spatially extended kinectfusion. 2012.

[181] J. Wilhelms and A. Van Gelder. Anatomically based modeling. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 173–180. ACM Press/Addison-Wesley Publishing Co., 1997.

[182] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90, 2016.

[183] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[184] S. Wuhrer, P. Xi, and C. Shu. Human shape correspondence with automatically predicted landmarks. *Machine Vision and Applications*, 23(4):821–830, 2012.

[185] L. Xia, C.-C. Chen, and J. K. Aggarwal. Human detection using depth information by kinect. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 15–22. IEEE, 2011.

[186] Z. Xiao, F. Mengyin, Y. Yi, and L. Ningyi. 3d human postures recognition using kinect. In *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2012 4th International Conference on*, volume 1, pages 344–347. IEEE, 2012.

[187] K. Xu, J. Zhou, and Z. Wang. A method of hole-filling for the depth map generated by kinect with moving objects detection. In *Broadband Multimedia Systems and Broadcasting (BMSB), 2012 IEEE International Symposium on*, pages 1–5. IEEE, 2012.

[188] J. Yang, H. Li, and Y. Jia. Go-icp: Solving 3d registration efficiently and globally optimally. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1457–1464, 2013.

[189] Y. Yang, Y. Yu, Y. Zhou, S. Du, J. Davis, and R. Yang. Semantic parametric reshaping of human body models. In *3D Vision (3DV), 2014 2nd International Conference on*, volume 2, pages 41–48. IEEE, 2014.

[190] M. Ye, H. Wang, N. Deng, X. Yang, and R. Yang. Real-time human pose and shape estimation for virtual try-on using a single commodity depth camera. *IEEE transactions on visualization and computer graphics*, 20(4):550–559, 2014.

[191] Y. Yoshiyasu, E. Yoshida, and L. Guibas. Symmetry aware embedding for shape correspondence. *Computers & Graphics*, 60:9–22, 2016.

[192] R. Yu, C. Russell, N. D. Campbell, and L. Agapito. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 918–926, 2015.

[193] 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.

[194] A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud. Surface feature detection and description with applications to mesh matching. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 373–380. IEEE, 2009.

[195] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 199–208. IEEE, 2017.

[196] Y. Zeng, C. Wang, X. Gu, D. Samaras, and N. Paragios. Higher-order graph principles towards non-rigid surface registration. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2416–2429, 2016.

[197] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, July 2017. IEEE Computer Society. Spotlight.

[198] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, volume 2, page 3, 2017.

[199] Q. Zhang, B. Fu, M. Ye, and R. Yang. Quality dynamic human body modeling using a single low-cost depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–683, 2014.

[200] W. Zhang, B. Han, P. Hui, V. Gopalakrishnan, E. Zavesky, and F. Qian. Cars: Collaborative augmented reality for socialization. In *HotMobile*, pages 25–30, 2018.

[201] Z. Zhang, K. Yin, and K. W. Foong. Symmetry robust descriptor for non-rigid surface matching. In *Computer Graphics Forum*, volume 32, pages 355–362. Wiley Online Library, 2013.

[202] Q. Zheng, A. Sharf, A. Tagliasacchi, B. Chen, H. Zhang, A. Sheffer, and D. Cohen-Or. Consensus skeleton for non-rigid space-time registration. In *Computer Graphics Forum*, volume 29, pages 635–644. Wiley Online Library, 2010.

[203] Q.-Y. Zhou, J. Park, and V. Koltun. Fast global registration. In *European Conference on Computer Vision*, pages 766–782. Springer, 2016.

[204] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. In *ACM Transactions on Graphics (TOG)*, volume 29, page 126. ACM, 2010.

[205] H. Zhu, Y. Liu, J. Fan, Q. Dai, and X. Cao. Video-based outdoor human reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, 27 (4):760–770, 2017.

[206] 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.

[207] 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.