Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

From scans to models: Registration of 3D human shapes exploiting texture information

Ph.D. Candidate
Federica Bogo

Advisor
Prof. Enoch Peserico

School Director
Prof. Matteo Bertocco

Coordinator
Prof. Carlo Ferrari

2015

Ph.D. School in
Information Engineering

Series XXVI

University of Padova

Dept. of Information Engineering

Sede amministrativa: Università degli Studi di Padova
Dipartimento: Ingegneria dell'Informazione

Scuola di dottorato di ricerca in: Ingegneria dell'Informazione
Indirizzo: Scienza e Tecnologia dell'Informazione
Ciclo: XXVI

From scans to models: Registration of 3D human shapes exploiting texture information

Direttore della Scuola
Prof. Matteo Bertocco

Coordinatore
Prof. Carlo Ferrari

Supervisore
Prof. Enoch Peserico

Dottoranda
Federica Bogo

PUBLICATIONS

This thesis presents scientific results published in:

- F. Bogo, J. Romero, M. Loper, M. J. Black. *FAUST: Dataset and evaluation for 3D mesh registration*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3794–3801, 2014. [34]
- F. Bogo, J. Romero, E. Peserico, M. J. Black. *Automated detection of new or evolving melanocytic lesions using a* 3D *body model*. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Lecture Notes in Computer Science, vol. 8673, pp. 593-600, 2014. [35]

In the last years I worked on several research topics. The main results obtained led to the following publications, that are not included in the corpus of this thesis:

- F. Bogo, F. Peruch, A. Belloni Fortina, E. Peserico. *Where's the lesion? Variability in human and automated segmentation of dermoscopy images of melanocytic skin lesions*. Book chapter in Dermoscopy Image Analysis, M. E. Celebi, T. Mendonca, J. S. Marques eds., CRC Press/Taylor & Francis. To appear.
- F. Peruch, F. Bogo, M. Bonazza, V. Cappelleri, E. Peserico. *Simpler, faster, more accurate melanocytic lesion segmentation through MEDS*. IEEE Transactions on Biomedical Engineering, 61(2), pp. 557–565, 2014.
- F. Peruch, F. Bogo, M. Bonazza, M. Bressan, V. Cappelleri, E. Peserico. *Simple, fast, accurate melanocytic lesion segmentation in* 1D *colour space*. 8^{th} International Conference on Computer Vision Theory and Applications (VISAPP), pp. 191–200, 2013.
- F. Bogo, E. Peserico. *Optimal throughput and delay in delay-tolerant networks with ballistic mobility*. ACM 19^{th} Annual International Conference on Mobile Computing & Networking (MobiCom), pp. 303–314, 2013.
- F. Bogo, M. Samory, A. Belloni Fortina, S. Piaserico, E. Peserico. *Psoriasis segmentation through chromatic regions and Geometric Active Contours*. 34^{th} Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5388–5391, 2012.

ABSTRACT

New scanning technologies are increasing the importance of 3D mesh data, and of algorithms that can reliably *register* meshes obtained from multiple scans. Surface registration is important e.g. for building full 3D models from partial scans, identifying and tracking objects in a 3D scene, creating statistical shape models.

Human body registration is particularly important for many applications, ranging from biomedicine and robotics to the production of movies and video games; but obtaining accurate and reliable registrations is challenging, given the articulated, non-rigidly deformable structure of the human body.

In this thesis, we tackle the problem of 3D human body registration. We start by analyzing the current state of the art, and find that: a) most registration techniques rely only on geometric information, which is ambiguous on flat surface areas; b) there is a lack of adequate datasets and benchmarks in the field. We address both issues.

Our contribution is threefold. First, we present a model-based registration technique for human meshes that combines geometry and surface texture information to provide highly accurate mesh-to-mesh correspondences. Our approach estimates scene lighting and surface albedo, and uses the albedo to construct a high-resolution textured 3D body model that is brought into registration with multi-camera image data using a robust matching term.

Second, by leveraging our technique, we present FAUST (Fine Alignment Using Scan Texture), a novel dataset collecting 300 high-resolution scans of 10 people in a wide range of poses. FAUST is the first dataset providing both real scans and automatically computed, reliable "ground-truth" correspondences between them.

Third, we explore possible uses of our approach in dermatology. By combining our registration technique with a melanocytic lesion segmentation algorithm, we propose a system that automatically detects new or evolving lesions over almost the entire body surface, thus helping dermatologists identify potential melanomas.

We conclude this thesis investigating the benefits of using texture information to establish frame-to-frame correspondences in dynamic monocular sequences captured with consumer depth cameras. We outline a novel approach to reconstruct realistic body shape and appearance models from dynamic human performances, and show preliminary results on challenging sequences captured with a Kinect.

SOMMARIO

Lo sviluppo di nuove tecnologie di scansione sta accrescendo l'importanza dei dati tridimensionali (3D), e la necessità di algoritmi di registrazione adeguati per essi. Registrare accuratamente superfici 3D è importante per identificare oggetti ed effettuarne il tracking, costruire modelli completi a partire da scansioni parziali, creare modelli statistici.

La registrazione di scansioni 3D del corpo umano è fondamentale in molte applicazioni, dal campo biomedico a quello della produzione di film e videogiochi; ottenere registrazioni accurate e affidabili è però difficile, poiché il corpo umano è articolato, e si deforma in maniera non rigida.

In questa tesi, affrontiamo il problema della registrazione di scansioni 3D del corpo umano. Iniziamo la nostra analisi considerando lo stato dell'arte, e rilevando che: a) la maggior parte delle tecniche di registrazione 3D usa solo informazione geometrica, che è ambigua in zone in cui le superfici sono lisce; b) c'è una mancanza di adeguati dataset e benchmark nel settore. L'obiettivo di questa tesi è quello di risolvere questi problemi.

In particolare, portiamo tre contributi. Primo, proponiamo una nuova tecnica di registrazione per scansioni 3D del corpo umano che integra informazione geometrica con informazione cromatica di superficie. La nostra tecnica dapprima stima l'illuminazione nella scena, in modo da fattorizzare il colore della superficie osservata in effetti di luce e pura albedo; l'albedo estratta viene quindi usata per creare un modello 3D del corpo ad alta risoluzione. Tale modello viene allineato a una serie di immagini 2D, acquisite simultaneamente alle scansioni 3D, usando una funzione di matching robusta.

Secondo, sulla base delle registrazioni prodotte dalla nostra tecnica, proponiamo un nuovo dataset per algoritmi di registrazione 3D, FAUST (Fine Alignment Using Scan Texture). FAUST colleziona 300 scansioni 3D relative a 10 soggetti in differenti pose. È il primo dataset che fornisce sia scansioni reali, sia registrazioni accurate e affidabili ("ground truth") per esse.

Terzo, esploriamo possibili usi del nostro approccio in dermatologia. Combinando la nostra tecnica di registrazione con un algoritmo di segmentazione per lesioni melanocitiche, proponiamo un sistema di screening in grado di rilevare l'insorgenza di nuove lesioni o modifiche in lesioni preesistenti su quasi tutta la superficie cutanea;

tale sistema è di aiuto per i dermatologi nell'individuazione di potenziali melanomi.

Concludiamo questa tesi esaminando l'importanza di usare informazione cromatica per registrare scansioni 3D acquisite in sequenze dinamiche. In particolare, proponiamo un nuovo approccio per ottenere modelli 3D realistici e completi del corpo umano a partire da sequenze acquisite con un singolo Kinect.

CONTENTS

1 INTRODUCTION

1.1 PROBLEM AND THESIS STATEMENT

This thesis addresses the problem of registering 3D real scans of human bodies. *How can one reliably establish dense correspondences between articulated, deformable surfaces like the human body? Can one assess the accuracy of such correspondences in a quantitative, rigorous manner?* And, in this light, *are current state-of-the-art registration techniques accurate enough, or do they exhibit some limitations?* We would like to put virtually no constraints on the body surfaces we consider. Humans have different shapes, and move assuming a variety of different poses; furthermore, in generic setups, we cannot assume the presence of predefined markers on their bodies.

Studying humans has always been extremely important in computer vision and graphics. Detecting people in images or videos, estimating their shape and pose, modeling hair, clothing or muscle deformations are just a few examples of tasks that have received a lot of attention during recent years. With the diffusion of new, low-cost scanning technologies, the amount of 3D data on humans that is gathered is constantly increasing. Effectively exploiting this data has an impact on different fields, ranging from entertainment and garment industry to medicine and surveillance.

Data captured with scanner devices, however, cannot be used "as is". Most scanning systems provide partial 3D surfaces, with different resolutions and topologies. Prerequisite to extract useful information from these scans is *registering* them – i.e. defining a set of correspondences between their surfaces (see Fig. 1).

Registering human scans is a challenging task. Scanner devices suffer from noise and outliers, as a result of software and hardware limitations; they can acquire only a portion of the body surface at each capture, due to occlusions. Human bodies are extended and articulated, and they deform in a *non-rigid* manner. Even though similar geometric features are important indicators for matching surface regions, they can significantly differ when the body is deforming.

In particular, there is a delicate interplay between shape and pose in the geometry of the human body. For instance, the area around the elbow deforms in different ways, depending on whether or not the arm is flexed. And such deformations vary if the subject is more or less muscular.

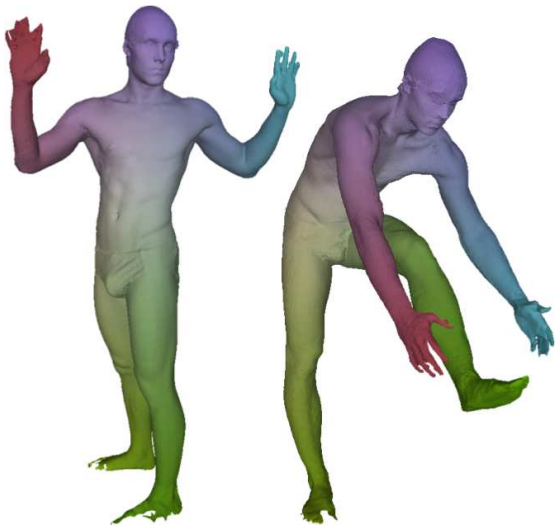


Figure 1: Registering two 3D scans means finding a set of correspondences between their surfaces: here, corresponding surface points are represented using the same color.

We find that geometric features alone are not enough to establish accurate correspondences between human bodies, particularly in smooth regions that are devoid of strong geometric features (e.g. the belly or the back). Due to this ambiguity, source and target surfaces may "slide" freely during registration.

The errors caused by sliding might be particularly significant when one considers bodies in motion: the wider the range of poses considered, the higher the difficulty in capturing non-rigid deformations based on geometry information alone.

We show that combining geometry with additional surface information that we call *appearance* (including texture and color), is a powerful tool for a) improving the quality of geometry-based registrations and b) quantitatively assessing their accuracy.

We develop a novel technique for registering 3D human scans that produces reliable correspondences by exploiting both geometric and dense texture features, and use it to create realistic body shape and appearance models.

To assess the accuracy of the correspondences produced by our algorithm, we initially rely on a high-frequency texture pattern applied to the skin. While assuming high-frequency texture can be reasonable when registering dressed people, it is somewhat restrictive when dealing with naked bodies. We later remove our initial assumption, and show that even the texture provided by the naked human skin (e.g. by the presence of moles or birthmarks) provides enough information to produce accurate registrations.

1.2 MOTIVATING APPLICATIONS

Obtaining accurate and reliable registration of 3D surfaces, and in particular of human bodies, has an impact on a variety of fields – ranging from graphics and computer vision to robotics and medicine.

GRAPHICS When we play a video game, watch an animated film, or interact with a 3D virtual environment, we often look at digital 3D representations of humans. A large number of these representations is inspired or directly adopted from the real world. For example, in the Digital Ira project [13], high-resolution scans and video recordings of an actor are combined to create a real-time, photoreal digital human character (see Fig. 2).

In general, data-driven approaches collect huge amounts of 3D data, and then aggregate this information to build representative 3D models. Prerequisite to this task is the definition of meaningful correspondences for the captured data.

Corpora of registered human scans have been used to learn statistical models of body shape and pose variation across people [15, 44, 65, 68]. These models find applications in a variety of tasks, like shape completion [14, 20, 68, 100], animation of motion capture sequences [89], morphing [14, 65], transfer of texture [14] and animation controls [14, 15, 127].

Surface registration is also used by data-driven approaches to model clothing [61] and specific body dynamics, like skin and muscle deformations [94, 99] and facial expressions [128]. Recent work on breath modeling [127] uses the registration technique proposed in this thesis.

Besides shape, texture plays an important role in digital rendering too. Recovering high-quality textures makes digital 3D reconstructions more realistic and appealing [13, 29, 32, 145]. Furthermore, accurate models of scene lighting and surface reflectance can help in capturing and reproducing fine geometric details [122, 137, 138].

COMPUTER VISION As mentioned above, defining dense correspondences between human scans in a given corpus allows the creation of low-dimensional statistical shape models. These models have been used to estimate human shape and pose from images [24, 62] and body shape under clothing [22, 64], and to predict anthropometric measurements [45, 126]. Tracking is another important task that can benefit from the introduction of a generative body model [67, 73, 93].

Although this thesis focuses on human body registration, we think that the insights we provide might be useful also in applications targeted to different object classes. Reg-

Figure 2: In the Digital Ira project [13], high-resolution scans and video recordings of an actor (top row) are combined to render a real-time digital character (bottom row).

istration of generic 3D surfaces is important for object identification and retrieval [36], and for building complete 3D models from partial scans [19, 29].

ROBOTICS Computer vision systems for autonomous robots can benefit from more accurate and robust 3D registration techniques, for instance in tasks like scene reconstruction [55], object identification or object retrieval [63]. Additionally, human body registration and modeling may have importance in the development of realistic humanoid robots.

MEDICINE One straightforward application of 3D human body registration is the monitoring of body shape evolution over time. This can be useful in nutrition and sport science: for example, in the analysis of the relations between physical exercise, diet and body shape, and in the prevention and treatment of eating disorders. It can also aid doctors in planning and monitoring rehabilitative activities for patients with motor disabilities. Furthermore, collecting corpora of registered body scans can help in understanding correlations between body shape and particular diseases (e.g. diabetes).

An example in this sense is given by the MPI BMI Visualizer [1] (Fig. 3). It is a web application that uses a statistical body shape model, learned from a corpus of

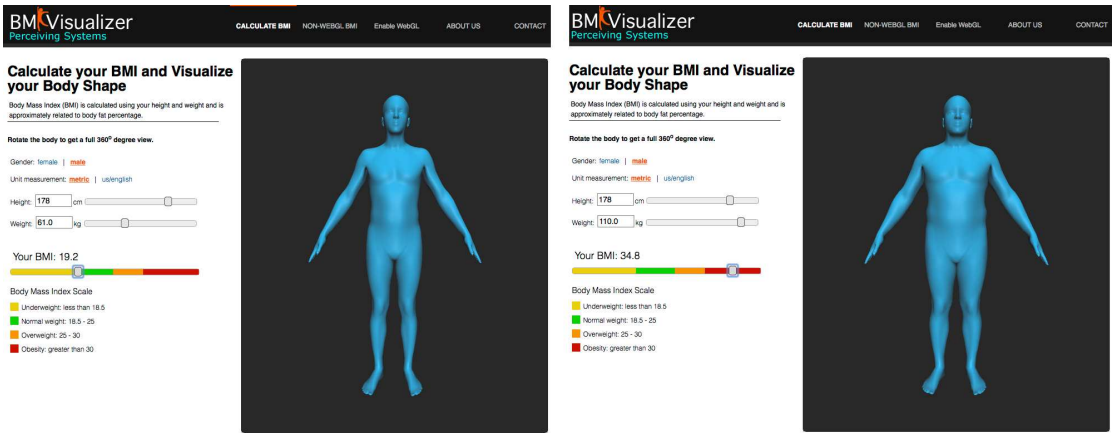


Figure 3: The MPI BMI visualizer [1] is a web application that uses a statistical human body model to associate a Body Mass Index (BMI) value provided by the user to a specific body shape.

registered scans, to visualize body shape variations as a function of the Body Mass Index (BMI).

Recovering appearance, besides shape, opens even further scenarios. Recently, our technique has been employed in a psychological study aimed at analyzing the influence of shape and appearance in the perception of self [104]. In Chapter 5, we show how our approach can be used to accurately track skin lesion changes over almost the entire body surface. This has potential applications in dermatology, for early diagnosis of melanoma.

1.3 CONTRIBUTIONS

In this thesis, we tackle the problem of defining accurate, dense correspondences between 3D surfaces, represented as polygonal meshes. We start by analyzing the current state of the art, and find that: a) most registration techniques rely only on geometric information, which is ambiguous on flat surface areas; b) there is a lack of adequate datasets and benchmarks in the field. The goal of this thesis is to overcome these limitations. While we focus on registration of human bodies, we believe that many of our insights can be of value also for more general approaches.

Our main contributions can be summarized as follows:

1. We identify an important shortcoming of current state-of-the-art registration techniques: accurate correspondences cannot be established in geometrically smooth

areas if we consider only geometry information. To quantitatively evaluate these inaccuracies, we combine the information given by geometry with another, complementary source of information: surface appearance.

2. We develop a novel registration technique for human bodies that provides dense and reliable correspondences. Our approach estimates scene lighting and surface albedo, and uses the albedo to construct a high-resolution textured 3D model that is brought into registration with multi-camera image data using a robust matching term.

3. We propose a quantitative evaluation metric for 3D mesh registration, that considers inaccuracies not only in terms of geometry, but also in terms of appearance.

4. Based on our technique and evaluation metric we build FAUST, a novel dataset collecting 300 high-resolution scans of different people in different poses. This is the first dataset providing both real meshes, and automatically computed "ground-truth" correspondences between them.

5. We evaluate several state-of-the-art registration techniques on our dataset and show that they have difficulties dealing with this real-world data.

6. We explore possible uses of our approach in the dermatological field. By combining our registration technique with a lesion segmentation algorithm, we propose a system that automatically detects new or evolving melanocytic lesions over almost the entire body surface.

7. We investigate the benefits of using appearance information to establish frame-to-frame correspondences in dynamic sequences; in particular, we propose a novel technique to obtain realistic shape and appearance models from challenging human performances captured with a consumer RGB-D camera.

1.4 THESIS OUTLINE

The remainder of this thesis is organized in six chapters.

Chapter 2 briefly reviews some basic notions. This preliminary background helps in contextualizing our work and makes subsequent chapter easier to understand. We address three major areas: scanning technologies, presenting the most widely used high- and consumer-quality acquisition systems; registration techniques, identifying

challenges and achievements in model-free and model-based non-rigid registration; appearance modeling, introducing the notions of texture map, albedo and shading, and common approaches to recover them.

The core of our work is presented in Chapter 3. After reviewing the most related approaches in the literature, we present our model-based registration technique for human scans. Our discussion proceeds in five steps. First, we describe our body model in detail. We then introduce our approach to estimate scene lighting and decompose scan surface color into albedo and shading. Based on this, we define our robust appearance-based error term and introduce the global objective function we minimize during registration. After providing some details about optimization, we qualitatively evaluate the impact of considering appearance, besides geometry, when defining dense scan-to-scan correspondences.

Our technique produces highly reliable registrations. We propose to use them to build a novel dataset collecting real human scans, FAUST. The dataset is described in Chapter 4. We discuss the main novelties introduced by FAUST in comparison with previous work, and provide details about its structure. We also present a novel evaluation methodology, that allows us to assess dense "ground-truth" correspondences between real scans. Finally, we define on FAUST a new 3D registration benchmark and evaluate, according to it, various state-of-the-art techniques, both model-based and model-free.

Chapter 5 explores possible uses of our approach for medical applications. We propose a novel system to monitor the evolution of melanocytic lesions over almost the entire body surface, and evaluate its accuracy through a preliminary pilot study.

Chapter 6 leverages our appearance-based technique to deal with data acquired with a consumer RGB-D camera. We identify the main challenges posed by this kind of data, outline a solution to recover body shape and appearance from challenging human motions, and show a set of preliminary results.

Finally, Chapter 7 summarizes our work and looks at directions of future research.

2 PRELIMINARIES

This chapter introduces some basic notions, that help contextualize our work and facilitate the comprehension of the subsequent chapters. We address three main areas: first, we briefly present the most important 3D scanning technologies currently used, with a focus on full-body scanners (Section 2.1); second, we provide an overview of 3D surface registration techniques, considering both model-free and model-based approaches (Section 2.2); third, we discuss how the problem of recovering high-quality appearance models for 3D data has been addressed so far in the literature (Section 2.3). We conclude the chapter providing some details about the mathematical notation used throughout this thesis (Section 2.4).

2.1 SCANNING SYSTEMS

Different applications may require, in general, to deal with different types of data. Some applications work with *static* 3D scans: a single "snapshot" of the object in time. This might be enough, for example, in shape retrieval or statistical shape modeling. For other applications, the temporal dimension is more important: for instance, to track people in a scene, or to learn a model of clothing or human soft tissue deformations, one should capture multiple 3D scans per second. In the following, we introduce for convenience a distinction between 3D and 4D scanning systems: the latter have a frame rate (i.e. number of frames/scans generated per second) suitable for capturing dynamic scenes, the former do not.

For both 3D and 4D systems, there is a tradeoff between quality of captured data and equipment costs. Quality can refer to the accuracy and completeness of the reconstructed surface geometry and texture, and to the absence of noise and outliers.

During recent years there has been an enormous diffusion of consumer depth cameras (like the well-known Kinect [2]); these cameras are commonly referred to as 2.5D devices, as they produce a depth image where each pixel contains the distance from the camera plane to the nearest object in the scene.

In this section, we present a brief overview of the most widely used scanning technologies available nowadays, ranging from expensive systems to home-scanning,

Figure 4: A full-body laser scanner from Cyberware [3] and a laser scan from the CAESAR dataset [107] (scan resolution: 124825 vertices). 3D reconstruction may be imperfect in areas that are either unreachable by the laser or invisible by the camera.

lightweight devices. Our presentation is not meant to be exhaustive, and focuses mainly on the technologies we will refer to in the rest of this thesis; we point the reader to [31, 109] for two surveys.

2.1.1 3D systems

For 3D static capture, *active* systems provide the most accurate results. In these systems, the light sources are specially controlled, as part of the strategy to arrive at the 3D information. In the following, we briefly review three important active scanning technologies: laser, structured light and active multi-stereo.

Triangulation-based laser scanners have been the state of the art for decades in 3D reconstruction of static scenes. In these scanners the triangulation takes place between a laser emitter and a camera. The emitter projects onto the target surface a single light stripe, that is easily detectable in the image taken by the camera; since the position and orientation of both emitter and camera are known, the 3D coordinates of all the points along the stripe can be determined. From a large sequence of images taken as the stripe sweeps across the target, a complete 3D surface can be reconstructed.

Laser triangulation produces highly accurate reconstructions [31]. One of the most comprehensive datasets of body shapes, the CAESAR 3D anthropometric dataset [107],

(a) (b)

Figure 5: Common structured light techniques project multiple parallel lines of light simultaneously on the target surface. To facilitate reconstruction, these patterns are usually encoded temporally (a) or spatially (b). (Image courtesy of C. Piccolo).

was built using two laser scanners from Cyberware [3] and Vitronic [4] (see Section 4.1.2 for details about the dataset).

To perform the "scanning" process (i.e. to sweep the laser stripe), these systems require a precise mechanical apparatus (Fig. 4). This affects costs, acquisition time (usually on the order of a few seconds or even minutes), and scanning volume; laser scans often exhibit holes in areas that are either unreachable by the laser or invisible by the camera (see again Fig. 4).

Structured-light scanners avoid the time-consuming "scanning" phase by directly projecting a two-dimensional pattern on the target surface. Common approaches project multiple parallel lines of light simultaneously, and rely on a precise temporal or spatial code to discriminate between the different lines in the pattern and the corresponding projection planes (Fig. 5). Projecting multiple patterns in sequence increases accuracy and robustness of the system [133], at the cost of a longer capture time.

In this work, we will use a full-body 3D active multi-stereo system from 3dMD [5]. Instead of using structured light, the system projects speckle patterns on the target surface; such patterns are captured and triangulated by multiple stereo cameras to reconstruct geometry (Fig. 6). The flexible positioning of the cameras allows scanning volumes which are larger than those available using laser scanners, reducing the number of holes in the scans; on the other hand, inaccuracies in stereo matching may produce noisier results in some areas (Fig. 7). The "one-shot" projection approach ensures acquisition times on the order of a few milliseconds; the recovery time between

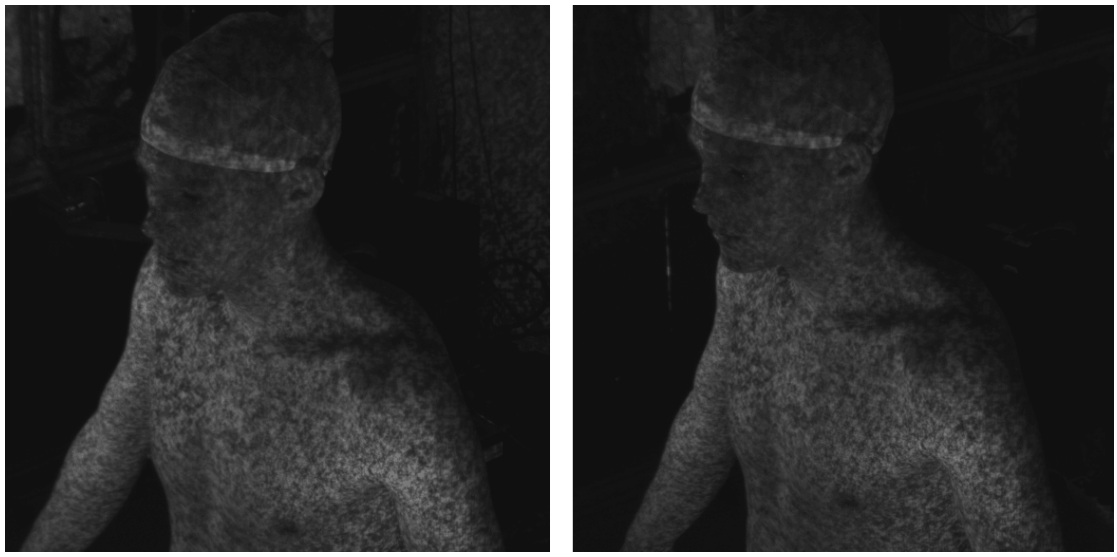


Figure 6: Example of projected speckle pattern captured by a pair of stereo cameras.

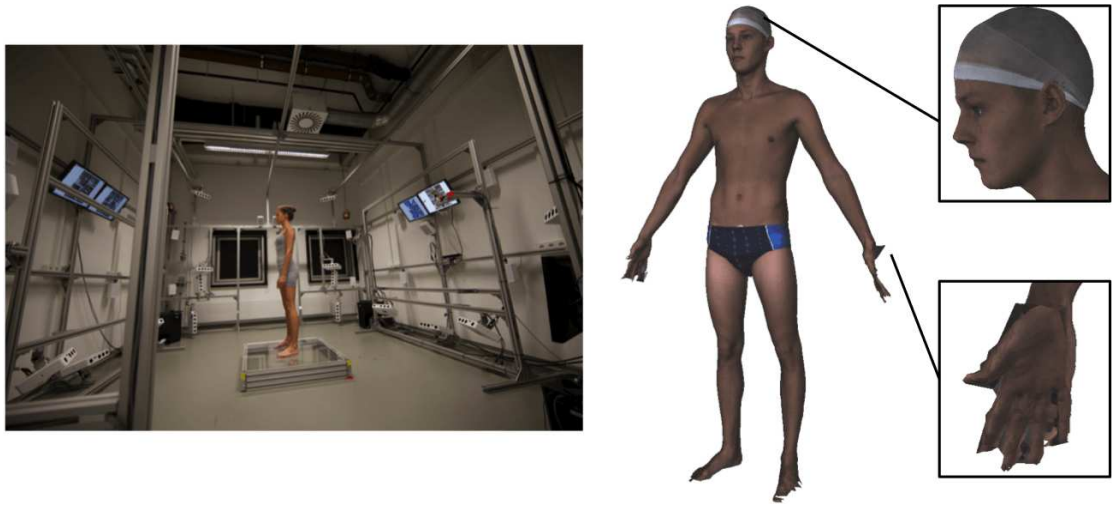


Figure 7: The full-body 3D active multi-stereo system from 3dMD [5] used to capture most of the data in this thesis, and an example scan (scan resolution: 174499 vertices). Reconstructions are comparable, in accuracy, to those produced by laser scanners. Active stereo may help in reducing holes in some parts of the scans (e.g. head), at the cost of increasing noise in others (e.g. hands).

subsequent scan captures requires a few seconds. Further details about the system are provided in Section 3.2.

2.1.2 4D systems

Laser sweeping or multiple structured pattern projection may be too time-consuming to capture full bodies in motion at a high frame rate [129]. As for active multi-stereo, 3dMD recently proposed a 4D scanner that extends the previous technology based on speckle projection to capture dynamic scenes too. The system captures and reconstructs dynamic scenes at a rate of 60 scans per second; it can be considered the first 4D scanner providing an accuracy comparable to that of static scanners.

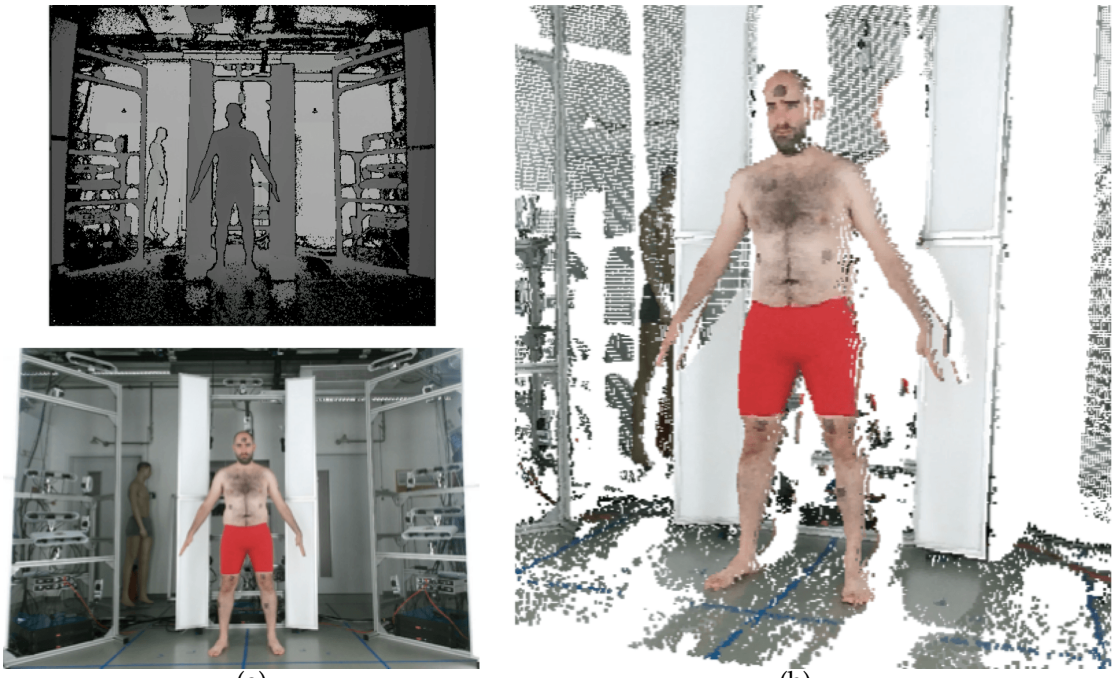
Other 4D scanners produce, in general, less accurate 3D reconstructions. In most cases, these scanners are multi-camera systems that adopt *passive* approaches for 3D surface reconstruction. Significant examples are multi-view passive stereo and shape from silhouette.

Multi-view passive stereo approaches simultaneously capture multiple images of the scene from different viewpoints, and then match them based on color information; the 3D coordinates of a point visible from multiple views can be reconstructed via triangulation. Matching accuracy may strongly affect the quality of the reconstruction; hence, the performance of these systems is highest with dense texture and many camera viewpoints. Recently, Infinite Realities [6] showed compelling results in recovering shape and high-resolution texture using more than one hundred synchronized DSLR (Digital Single-Lens Reflex) cameras; the system, however, still requires manual intervention in the reconstruction pipeline.

Shape-from-silhouette approaches extract silhouette images of an object captured from multiple cameras. A common way to facilitate silhouette extraction is to provide a simple background, like a homogeneous blue or green cloth. By backprojecting in 3D the silhouettes according to the camera parameters and intersecting these volumes, the visual hull of the object can be reconstructed. Customization of number, positioning and resolution of the cameras allows recovery of shape and texture at different frame rates and levels of detail. For instance, systems from 4D View Solutions [7] can ensure a frame rate up to hundreds of frames per second.

Since visual hulls provide only an approximate estimation of the object's geometry, shape-from-silhouette approaches are often combined with other techniques. For instance, in [115] the authors merge shape from silhouette and multi-view stereo to improve the quality of 3D reconstruction.

Shading cues can be exploited to infer high-frequency geometric details. Approaches proposed in [122, 129] combine multi-camera systems with sophisticated lighting in controlled studio setups to obtain highly detailed 3D reconstructions.



(a) (b)

Figure 8: Example data captured with a Kinect 2: depth and RGB images (a), reconstructed point cloud (b). The point cloud and the RGB image have been cropped for visualization purposes. (Scan resolution: 217088 points, including background).

2.1.3 Consumer depth cameras

During recent years the diffusion of consumer depth cameras has promoted the development of more lightweight, low-cost scanning systems.

These devices usually combine a depth sensor with a color (RGB) camera to produce a depth and an RGB stream; typical frame rates are on the order of 30 frames per second. Each frame of the depth data stream (a depth image) is made up of pixels that contain the distance from the camera plane to the nearest object; a depth image can be converted into a 3D point cloud, which contains a point for each pixel of the depth image. Note that this represents only a partial 3D reconstruction of the scene (areas not visible from the camera are not reconstructed). Throughout this thesis, we will refer to these devices also as RGB-D cameras.

One of the most widely used RGB-D cameras is probably the Microsoft Kinect [2], whose first version was released in 2010. The device uses a near-infrared camera and a near-infrared laser source to recover depth via structured light. Other low-cost RGB-D cameras rely on a similar technology [8, 9].

The new Kinect 2, released recently, improves 3D reconstruction by replacing structured light with a time-of-flight sensor. The sensor indirectly measures the time it takes for pulses of laser light to travel from a laser projector to the target surface, and then back to the image sensor. Figure 8 shows example RGB and depth data captured with a Kinect 2, together with the reconstructed point cloud.

Prices of consumer depth cameras are on the order of a hundred euros; professional systems can cost ten or a hundred times more. Lower prices often mean lower quality: cheaper devices produce more incomplete, noisier, lower-resolution data. Which acquisition system to use is application-dependent, and registration techniques should be carefully targeted to the data they must deal with.

2.2 MODEL-FREE VERSUS MODEL-BASED NON-RIGID REGISTRATION

Human bodies are *articulated*, *non-rigid* objects. Consider two scans of two different subjects (or of the same subject in two different poses): the relation between their surfaces cannot be expressed by a rigid transformation. We need to model more complex deformations.

Lifting the rigidity assumption makes the registration problem significantly more challenging: registration of non-rigid shapes requires estimating both a set of correspondences, and a suitable warping function that matches the deformation of the target shape; false correspondences are debilitating, since they can lead to strong distortions that are not consistent with the target [83]. And indeed, while many effective rigid registration techniques have been proposed in the last years, non-rigid registration remains an open problem [119].

This section briefly reviews the main approaches in the literature, dividing them into model-free and model-based. Model-free techniques do not make any assumption about the object to be registered; model-based techniques specifically address a class of objects. Our discussion will mainly focus on techniques we will refer to in the remainder of this thesis; we point the reader to [76, 119] for two exhaustive surveys.

2.2.1 Model-free approaches

Recent model-free techniques emphasize *intrinsic* surface representations, that are invariant to bending being based on properties like surface distances and angles. These representations can be used to embed the surfaces to be matched into a new space,

where their intrinsic geometry is preserved; in the embedded space, the matching problem reduces to rigid registration.

While many intrinsic representations have been proposed in the literature [76], here we focus on three of the most significant ones: geodesic distances, diffusion distances and Möbius transformations.

The geodesic distance between two surface points is defined as the length of the shortest path, traveling on the surface, that connects the points. Bronstein et al. [37] represent surfaces as metric spaces, whose metric structure is defined by the geodesic distance between pairs of points on the surface; a metric space is embedded into another via Generalized Multidimensional Scaling (GMDS). In this framework, registering two surfaces corresponds to finding the embedding with minimum distortion (i.e. the embedding that minimizes the discrepancy between the two metric spaces). Geodesic distances can be computed efficiently, but are sensitive to topological transformations (modifications in mesh connectivity alter the paths between points, and can result in significant changes of the geodesic distances).

Alternative representations based on diffusion distances provide greater robustness against topological changes [39]. Diffusion distances [46] are related to the probability of traveling on the surface from one point to another in a fixed number of random steps; they can be efficiently computed from the eigenvalues of a discrete approximation of the Laplace-Beltrami operator [105]. Mateus et al. [90] match articulated shapes combining Laplacian embedding with probabilistic point matching. Powerful descriptors used for shape analysis and matching, like the Heat Kernel Signature [36, 98, 117] and the Wave Kernel Signature [21], are based on diffusion distances.

Lipman et al. [87] pioneer the use of the Möbius transformations, that are conformal (preserve angles) and contain as a subset the group of isometric (distance-preserving) transformations. In [87], the authors propose to compute multiple conformal mappings between the surfaces to be matched, and then combine them with an algorithm that "votes" for the most reliable correspondences. Kim et al. [78] extend the approach, blending different maps instead of adopting a voting scheme. Zeng et al. [143] take a different direction, combining ideas from [87] with a Markov Random Field (MRF) framework to perform surface tracking.

Despite the amount of work in the field, techniques based on intrinsic representations still present a number of limitations. Many approaches can deal only with nearly isometric deformations. In some cases, they provide only sparse correspondences; and, depending on the adopted intrinsic representation, they may fail for human shapes due to reflective symmetries [52, 78] (e.g. the front of the body is mapped to the back). The main limitation, however, seems their scarce robustness to noise. Most methods require watertight meshes as input; topological changes and presence

of holes may severely affect the accuracy of these approaches, making them unsuitable for real-world applications.

2.2.2 Model-based approaches

Many practical applications require to align ("fit") a common template mesh to noisy scans [14, 18, 20, 65, 68, 82, 100, 139, 144] (see Fig. 9). The template represents a strong geometric and topological prior, that provides automated hole filling and noise removal; since it has fixed topology, fitting it to different scans is equivalent to defining dense correspondences between them.

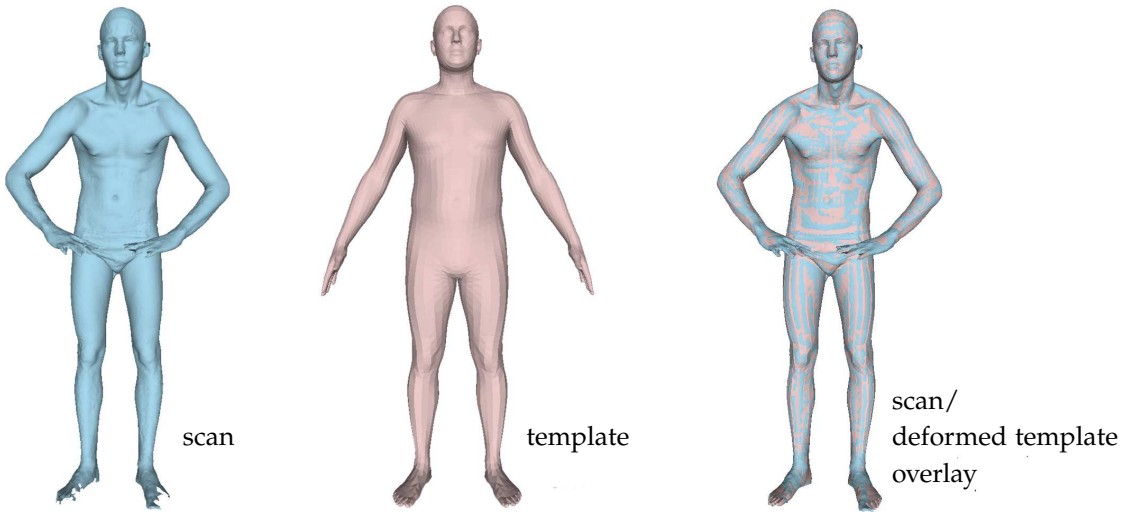


Figure 9: In many practical applications, a common template mesh is deformed to fit noisy scans.

The alignment process usually corresponds to the minimization of an error function, combining one or more data terms with some form of regularization.

Classic data terms exploit geometric information. Most approaches use extensions of the Iterative Closest Point (ICP) algorithm [30] to non-rigid surfaces [41, 71]. The iteration takes place between two subproblems: estimating a set of correspondences between scan and template surfaces, and, based on the current correspondences, estimating a non-rigid transformation deforming the template to fit the scan. Non-rigid ICP or variants have been used to align body parts like faces [18, 82], hands [82], and full bodies [14, 65, 68, 93, 139].

Since non-rigid ICP is sensitive to local optima [71], the registration is often initialized by identifying manually or automatically a set of corresponding landmarks on both surfaces. Several approaches [14, 15, 65] use markers placed by a human op-

erator, typically in correspondence with key anatomical locations (e.g. where bones are palpable through the skin). Wuhrer et al. [139] use automatically-detected landmarks to initialize a skeleton-based template; Anguelov et al. [20] detect initial sparse markers using the Correlated Correspondence algorithm [19].

More rarely, data terms consider color cues too. Dense texture information has been used to register face scans [32], while sparse constraints based on SURF descriptors [28] have been used for template-based tracking of human performances [85].

Common regularization terms used during alignment act on the deformations of the template surface. They fall into two main classes: "as-rigid-as-possible" terms and "smoothness" terms. As-rigid-as-possible terms penalize local deformation estimates as they deviate from rigidity [114]; they assume near-isometric deformations, such as those that occur when aligning scans of the same person [82, 83], so are commonly not used when registering different body shapes. Smoothness terms penalize deformations changing rapidly over the surface of the template [14, 18, 65, 68].

More informative regularization terms can be obtained using class-specific shape models. For example, Amberg [17] aligns a template to face scans using a deformable head model. The model is coupled to the aligned template by a prior that measures smoothness of the deformation between the alignment and an optimized fit of the deformed model. Compared to regularization acting only on geometric properties of the template surface, this approach ensures more consistent registration across scans.

Finding expressive and manageable models for complex objects like the human body, however, is challenging. The next section presents an overview of the most significant human shape models proposed in the literature.

2.2.3 Human shape models

Once a corpus of scans has been registered to a common template, standard multivariate statistical methods can be used to model the distribution of shapes [68]. In the case of faces [32] and bodies in a single pose [14, 110], low-dimensional models have been obtained by performing Principal Component Analysis (PCA) [75] on the aligned template vertices. For scans of multiple poses, approaches in the literature propose articulated, parametric body models that represent both shape and pose of each registered scan [15, 20, 44, 64, 65, 68]. These approaches may in general differ on 1) how they address registration and model learning stages and 2) how they model the correlation between body pose and shape.

In most cases, registration and model learning stages are addressed separately: first, a corpus of scans is brought into registration; then, model parameters are learned from

the registered data. Consequently, the quality of the model heavily depends on the quality of the registration.

Other approaches adopt different schemes. In [32], the authors bring a dataset of 3D face scans into registration in an iterative fashion: they alternate between learning a model based on the current registrations, and then using the model to initialize a new round of registration; iterations help in improving correspondence consistency across scans. The *coregistration* framework proposed in [68] registers a corpus of scans of different people in multiple poses and learns a set of body model parameters by minimizing a unique objective function. This "concurrent" approach brings a number of advantages: good data provided by some scans can be used to explain poor or missing data in others; consistency of a subject's shape across poses is modeled more explicitly; similarities in the deformations of different bodies as they change pose are captured more easily. Further details about the coregistration approach are provided in Section 3.3.

The relationship between body pose and shape has been modeled in different ways. The well-known SCAPE model [20] *decouples* pose and shape: it uses PCA to learn a pose deformation model from scans of an individual in multiple poses, and a shape model from many subjects in a neutral pose. SCAPE has been widely used for estimating human shape and pose from images [23, 24, 62, 112], body shape under clothing [22] as well as for reshaping human body in images and videos [73, 146]. The major limitation of SCAPE is that it does not capture the correlation between body shape and pose: similar changes in pose result in similar surface deformations, independently of the identity of the subject (e.g. male or female, more or less athletic).

Other approaches try to explicitly encapsulate the correlation between shape and pose. Allen et al. [15] propose Maximum a Posteriori (MAP) framework to learn a correlated model of shape- and pose-dependent variation. The optimization procedure is expensive, since it requires to solve a set of nonlinear functions with a high number of degrees of freedom. The Tensor-based Body (TenBo) model introduced in [44] represents surface deformations as a joint function over both shape and pose parameters using the tensor decomposition technique. Hasler et al. [65] introduce a translation- and rotation-invariant surface representation, and use it to learn a single PCA space of shape and pose variations. Since in this approach the shape and pose components cannot be decoupled and analyzed separately, the authors propose to train a set of regression functions to correlate PCA coefficients with significant values like weight, pose or body fat content. Subsequent work [64] extends this approach by learning two low-dimensional spaces of shape and pose, and combining them in a bilinear model.

In this thesis, we use the *BlendSCAPE* model, a modified version of SCAPE introduced in [68]. A detailed description of BlendSCAPE is provided in Section 3.3.

2.3 RECOVERING APPEARANCE

In addition to surface geometry, most scanning systems also capture surface texture and color, usually in the form of one or more RGB images. Combining such images into a unique texture map, representing the scanned object's appearance, helps in obtaining more realistic 3D reconstructions. As we will see in the following chapters, appearance may also represent a source of information complementary to geometry when defining correspondences across scans.

This section first formalizes the notion of texture mapping (Section 2.3.1) and introduces the fundamental distinction between surface albedo and shading (Section 2.3.2); then, it presents the most significant techniques proposed in the literature to recover high-quality appearance models from RGB and geometry data (Section 2.3.3).

2.3.1 Texture mapping

Texture mapping was introduced in computer graphics as early as 1974 [43] as a very effective means to increase visual rendering complexity of 3D polygonal meshes without the need to increase geometry details [54].

The most straightforward way to represent the appearance of a 3D mesh is to assign a color (e.g. a RGB triplet) to each vertex. Reproducing appearance with a high level of detail requires a high number of vertices, increasing computational complexity.

The introduction of texture mapping allows to decouple geometry and color. Surface color is stored as a texture map, i.e. an image that corresponds to a parameterization in 2D (*texture space*) of the original 3D surface (see Fig. 10). Each pixel (or *texel*) of the texture map is identified by a unique pair of coordinates.

A mapping function uv defined from 3D to texture space assigns to each surface vertex a pair of texture coordinates; the color assigned to vertex $\boldsymbol{v}$ is stored by texel $uv(\boldsymbol{v})$. Given this mapping, texture coordinates for an arbitrary surface point are calculated via linear interpolation (see Fig. 10).

In this way, the level of detail that can be reproduced is determined by the resolution of the texture map (i.e. by the number of texels), not by the number of mesh vertices. This allows the generation of detailed textured 3D models, without increasing their geometrical complexity.

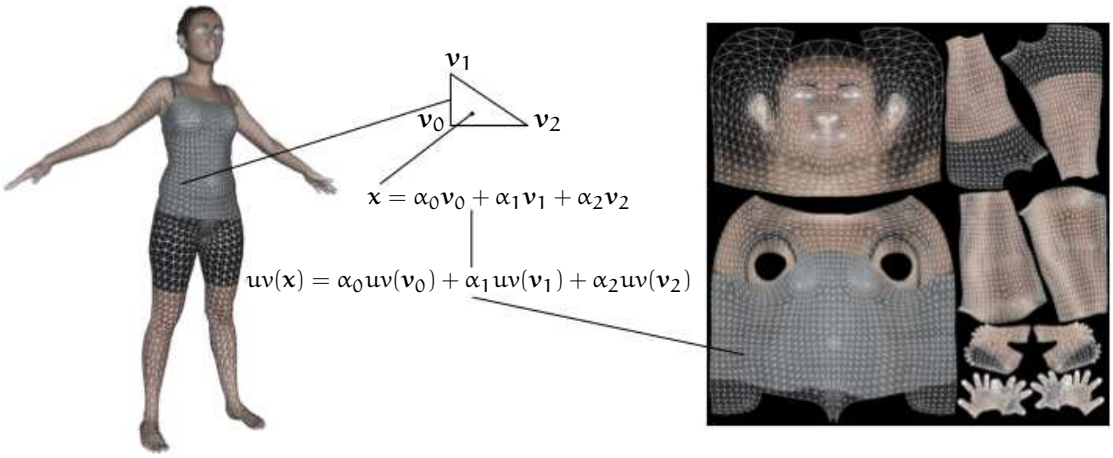


Figure 10: The basic mechanism used in texture mapping. A mapping function uv assigns to each mesh vertex a pair of texture coordinates; given this mapping, texture coordinates for an arbitrary surface point x are obtained by linearly interpolating the coordinates of the vertices of the face x belongs to.

2.3.2 Albedo and shading

Imagine observing a subject that is walking in the sunlight. Given a point x on his body surface, the color observed at x varies over time, due to shading effects depending on body pose. When recovering appearance models, it might be desirable to decompose the observed surface color into its "constant" and "transient" components – that is, to discriminate between *albedo* and *shading*. Albedo depends only on physical properties of the surface itself (like its material), so it is invariant to changes in scene lighting. Estimating it is necessary, for example, for rendering object appearance under novel, different lighting conditions; in our work, we will exploit albedo constancy when registering different scans of the same subject.

Albedo and shading can be estimated by *inverting* the pipeline used by common graphic renderers. That is, one starts with a plausible image formation model, and tries to estimate its components. The components combined to generate an image are essentially three: scene lighting, object geometry and object reflectance. Different models of lighting and reflectance have been proposed in the literature. In the following, we initially consider the simplest lighting model (a single point light source), and introduce two widely used reflectance models: Lambertian and Blinn-Phong. Then, we discuss how to capture more complex lighting environments.

Consider an object illuminated by a single light source. The color i_x observed at point x on the object's surface depends on the incoming light at x, on the surface normal n_x and on the surface reflectance at that point.

Different objects (or even different surface points of the same object) may have different reflectance. In the simplest case the surface is *Lambertian*, i.e. reflects the light equally in all directions. For Lambertian surfaces, the color i_x observed at surface point $\mathbf{x}$ is explained as

$$i_x = (\mathbf{n}_x \cdot \mathbf{l}_x) a_x l \tag{2.1}$$

where a_x is the surface *albedo* at $\mathbf{x}$, $\mathbf{l}_x$ is the direction from $\mathbf{x}$ to the light source, and l is the light intensity.

For glossy objects, reflectance may also have a specular component that depends on the viewing direction. The Blinn-Phong reflectance model [33] adds a specular term to the diffuse Lambertian one:

$$i_x = (\mathbf{n}_x \cdot \mathbf{l}_x) a_x l + (\boldsymbol{\xi}_x \cdot \mathbf{n}_x)^{\alpha_s} s_x l \tag{2.2}$$

where s_x is the specular coefficient, $\boldsymbol{\xi}_x$ is the halfway vector between $\mathbf{l}_x$ and the direction from $\mathbf{x}$ towards the observer, and α_s is the specular exponent.

Although human skin has a specular component [62, 135], the Lambertian model provides a good approximation for body reflectance and is used for its simplicity [23, 60, 136, 138]. In Chapter 3, we will estimate body albedo combining a Lambertian reflectance model with a visibility term that takes into account cast shadows.

In addition to reflectance, one can modify the lighting model to capture the scene more realistically. While for outdoor scenes light intensity can be assumed constant for each surface point (as in Eq. (2.1)), for a point light source in an indoor scene it is common to use an attenuation term that grows quadratically with distance [23, 62]. Multiple light sources in the scene can be modeled by simply summing multiple terms, one per light source.

With a high number of light sources – and, in general, for diffuse, low-frequency lighting environments – efficient models proposed in the literature use a low-order Spherical Harmonic basis [25, 113, 136, 138].

Spherical Harmonics (SH) define an orthonormal basis over the sphere, analogous to the Fourier transform over the 1D circle [113]. In the SH model, reflectance and lighting are represented as functions over the sphere, that are projected onto a low-order SH basis. Usually, 9 to 25 basis vectors are enough to obtain a realistic model [26, 113]: in this way, the computation of a rendering equation like the one in Eq. (2.1) boils down to the computation of a dot product between vectors of no more than 25 elements.

In our approach, we will model scene lighting using a 9-dimensional SH basis; further details are provided in Section 3.4.1.

2.3.3 Image-based texture reconstruction

In image-based texture reconstruction, one faces the problem of integrating multiple views of a 3D surface into a single texture map. These views may refer to a unique time instant (e.g. a static scene captured by multiple cameras), or to different ones (e.g. a dynamic scene captured by one or more cameras).

Integration across views is mainly addressed in two ways: blending information from all the views on a per-texel basis [27, 29, 132], or building the texture as a mosaic of unique-view contributions, whose seam locations are optimized to minimize appearance change between fragments [16, 74, 80, 81]. In both cases, one has to introduce some heuristics to assess the "quality" of each view, for a given surface patch. Common heuristics rely on viewing distance [80], angle between surface normal and viewing direction [29, 80, 81, 132], color variation for corresponding surface patches in multiple images [29], area of the patch projected on image space [16, 27, 74].

In general, the quality of the reconstructed texture is dependent on a good alignment between 2D and 3D data: inaccuracies in geometry reconstruction and camera calibration, and wrong frame-to-frame surface correspondences in temporal sequences, may result in blurring and ghosting (double imaging) artifacts. To address this, a number of methods employ some form of additional registration before estimating texel color [29, 54, 81, 122, 132]. In particular, the approaches in [54, 122, 132] use optical flow to warp and align texture data in dynamic sequences.

Achieving exact alignment between 2D and 3D data, however, is impractically difficult, particularly when computing high-resolution texture maps. Consequently, in methods blending contributions on a per-texel basis, the fewer the cameras influencing the result for a single texel, the sharper the resulting texture is [59]. On the other hand, if only the contributions of few cameras are blended for a given texture patch, well-visible seams and discontinuities might arise at patch boundaries. Both kinds of approaches do not exhibit scalability. To better exploit redundant information from multiple views, Goldluecke et al. [59] propose a super-resolution framework, that optimizes for a super-resolved texture map, a set of camera calibration parameters and a displacement field correcting local surface geometry. The approach works only for static scenes. Tsiminaki et al. [125] adapt and extend the super-resolution framework to deal with dynamic scenes, but only over very limited time intervals (no more than 7 frames).

In our work, we will explicitly address the problem of reconstructing a single texture map from multiple images, and leverage this map, in combination with geometry information, to define more accurate scan-to-scan correspondences. In contrast to most previous work, we consider bodies sampled at "arbitrary" (not necessarily close) time

instants: this means that our technique should be able to handle significant non-rigid deformations. Furthermore, we are mainly interested in solving the correspondence problem in an accurate manner; obtaining high-quality texture maps is a related, secondary effect. This is why, instead of simply warping texture data to improve visual quality, our approach optimizes for both an appearance and a shape body model.

2.4 MATHEMATICAL NOTATION

This section provides some details about the mathematical notation adopted throughout this thesis. In particular, we would like to draw the attention of the reader to four rules we adhere to:

- We denote objective functions using an upper-case E with a subscript (e.g. $E_x(\cdot)$). In general, we try to choose subscripts that provide an intuitive description of the function. In some cases, we use the same subscript for functions that differ in their formal definition, but are "operatively" analogous (e.g. they both compute the distance in 3D between the surface of a scan and that of a template, as in Eqs. (3.2) and (6.2)).
- Given a d-dimensional vector $\mathbf{x}$, we denote by $\|\mathbf{x}\|$ its Euclidean norm $\sqrt{\sum_{i=1}^{d} x_i^2}$.
- When defining the parameters of our objective functions, we use a semicolon to distinguish parameters that are optimized from those that are kept fixed. For example, objective $E(\alpha, \beta; \gamma)$ depends on parameters α, β and γ; we minimize it with respect to α and β.
- For simplicity, in some cases we refer to sets of elements using an abbreviated notation: given a set of N_X elements $X = \{X^i : i = 1, \ldots, N_X\}$, we denote it by $\{X^i\}$. We adopt this notation only if there is no ambiguity about the membership of the set.

3 REGISTERING HUMAN SCANS USING APPEARANCE

This chapter introduces our registration technique for 3D human meshes.

We register a corpus of scans of different people in multiple poses by aligning a common template mesh to each scan. Our approach adapts and extends the *coregistration* framework introduced in [68], that simultaneously builds a model of the body and its deformations while registering the scans using the model. The main novelty of our approach is the use of appearance information, in addition to geometry, during registration. We estimate scene lighting and surface albedo, and use the albedo to construct a high-resolution textured 3D model; the model is then brought into registration with multi-camera image data using a robust matching term. The use of appearance information helps in solving correspondence ambiguities in geometrically smooth areas, producing highly reliable registrations.

The chapter is organized as follows. Section 3.1 reviews the most related techniques in the literature. Section 3.2 describes the 3D scanning system we use to capture our scans. Details about the coregistration framework are provided in Section 3.3. Sections 3.4, 3.5 and 3.6 present the key components of our approach, formalize the objective functions we minimize during registration and provide details about optimization. Finally, Section 3.7 discusses – mainly from a qualitative point of view – the impact of considering appearance information during registration. Quantitative results will be provided in Chapter 4.

3.1 PREVIOUS WORK

We briefly reviewed the rich literature on 3D surface matching in Section 2.2. Here we summarize the key themes, with a focus on use of appearance information and human body registration. Human body shape modeling has received a great deal of attention recently [14, 15, 20, 44, 64, 65, 68, 110], but there is a paucity of high-quality registered scan data for building and evaluating such models.

We can roughly split registration techniques in the literature into model-free and model-based. The former do not make any assumption about the objects to be regis-

tered; the latter specifically address a class of objects, relying on stronger shape priors. We briefly review the most relevant techniques for both classes, and then analyze in greater detail approaches that make use of appearance information.

Recent model-free approaches start by defining an intrinsic surface representation that is invariant to bending. This representation is used to embed the surfaces to be matched in a new space, where their intrinsic geometry is preserved. In the embedded space the matching problem reduces to rigid alignment. Representative model-free techniques are Generalized Multi-Dimensional Scaling (GMDS) [37], Möbius voting [87], Blended Intrinsic Maps [78], heat kernel matching [98]. These approaches often provide only sparse correspondences, suffer from reflective symmetries (e.g. the front of the body is mapped to the back), and typically require watertight meshes as input; in some cases, they do not handle topological changes.

Model-based approaches for human body registration commonly fit a template mesh to noisy scans. Often the template is of lower resolution. Classic approaches employ non-rigid ICP in conjunction with simple regularization terms favoring surface smoothness [14, 18, 65, 68] or deformations that are as rigid as possible [82]. Since non-rigid ICP is sensitive to local optima, the registration is often initialized by identifying (manually or automatically) a set of corresponding landmarks on both surfaces [14, 15, 20, 65, 139]. The introduction of shape priors, by coupling the template to a learned model during alignment [68], can increase accuracy and robustness.

These approaches rely only on geometry information. Geometry alone may not prevent template vertices from being positioned inconsistently (i.e. sliding) across smooth scan surface areas. While many regularization methods have been proposed, without ground truth it is unclear how well they work at preventing this sliding.

A limited number of approaches exploits appearance information to solve geometry ambiguities. In most cases, they rely on sparse photometric features. Thorstensen and Keriven [123] extend GMDS adding a photometric-based error term to the formulation provided in [37]; the technique works only with sparse correspondences, and, like other model-free approaches, is not robust to mesh noise and topological changes. Zaharescu et al. [141] propose local feature descriptors for 3D surfaces based on geometric and photometric information, exploring possible applications to mesh matching. In [85], the authors use sparse texture-based constraints (SURF feature descriptors [28]) to improve their shape completion algorithm for dynamic scenes; correspondences are defined only over short time windows.

Dense texture has been used for 3D model-based alignment of body parts like faces [32]. Full bodies, however, are substantially different. Their articulated structure is too complex to represent with the cylindrical 2D parameterization in [32]; they self occlude and self shadow; they are too extended to assume a simple lighting model;

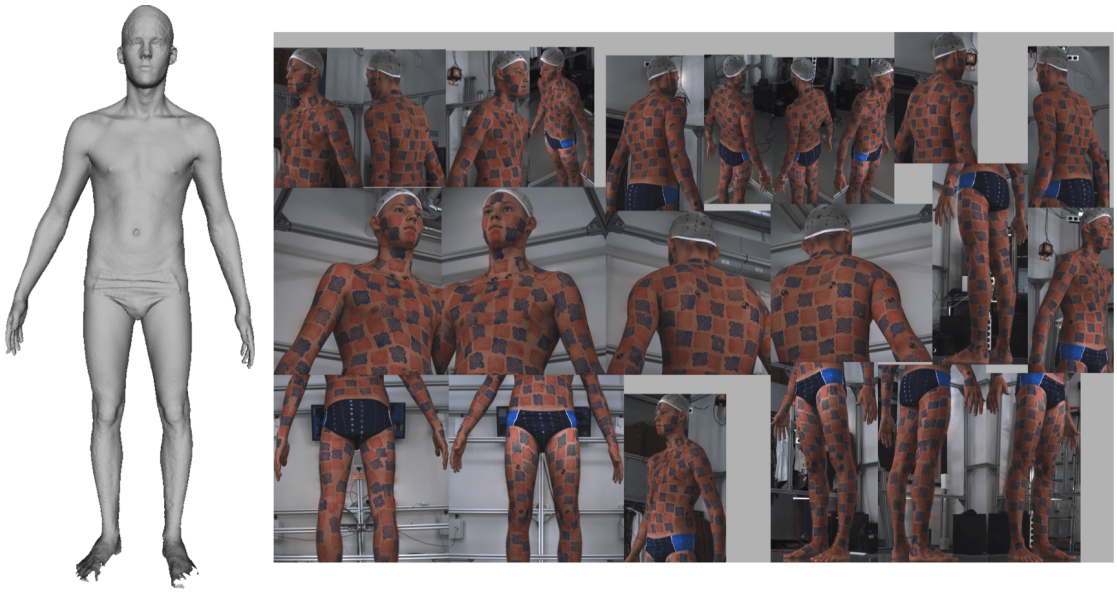


Figure 11: A scan reconstructed with our full-body 3D active multi-stereo system. Synchronized with each scan, we have 22 RGB cameras capturing surface texture.

the size of the body typically means lower-resolution texture as compared with face scans. We are aware of no full-body 3D mesh registration method that uses dense texture.

3.2 SCAN ACQUISITION

Our acquisition system is a full-body 3D active multi-stereo system, built by 3dMD [5] (see also Section 2.1 for details about multi-stereo capture). The system is composed by 22 scanning units; each unit contains a pair of stereo cameras for 3D shape reconstruction, one or two speckle projectors, and a single 5MP RGB camera. For efficiency purposes, we downsampled the RGB images to 612×512 pixels.

A set of 20 flash units illuminate the subject during capture, rendering a fairly diffuse light environment. The delay between speckle pattern projection and texture acquisition is around 2ms.

Each reconstructed scan is a triangulated, non-watertight mesh with $100000 - 200000$ vertices. Figure 11 shows a scan reconstructed with our system, together with the corresponding camera images.

To obtain high-frequency appearance information over the entire body surface, we painted the subjects with a dense texture pattern prior to scanning (see again Fig. 11).

This pattern implicitly defines full-body ground-truth correspondences between scans of the same subject; we will leverage it to quantitatively evaluate the accuracy of our registrations (see Section 4.2). In Chapter 5, we will remove this assumption and rely, during registration, only on texture information provided by the naked human skin (e.g. by the presence of small artifacts like moles and birthmarks).

3.3 THE COREGISTRATION FRAMEWORK

We adapt and extend the coregistration framework introduced in [68], that simultaneously brings a corpus of scans into registration and learns a set of body model parameters. In its original formulation, coregistration does not leverage appearance information; we add this and introduce a number of improvements.

Our approach proceeds in two stages: first, we *coregister* a corpus of scans based on geometry information alone; then, we refine our registrations by introducing a novel appearance-based error term. This section focuses on the first stage: it provides details about the body model we use, *BlendSCAPE*, and introduces the objective function we minimize to obtain a first round of registration and train our model. Sections 3.4 and 3.5 will focus on the second stage.

3.3.1 The BlendSCAPE body model

We assume we have a corpus of scans $\{S^k : k = 1, \ldots, N_{scans}\}$, collecting scans of N_{sbj} different subjects in multiple poses. We index subjects by p; for simplicity, we identify by p_k the subject of scan S^k.

We register the corpus by aligning a triangulated template mesh T^* to each scan. In our model-based approach, the deformations that fit T^* to a scan are regularized towards a deformable, statistical human body model. We use the *BlendSCAPE* body model, a modified version of SCAPE [20] introduced in [68]. For completeness we include here a brief description of the model, trying to adopt a notation as similar as possible to that used in the original work [68].

Let T^* be pre-segmented in 31 parts, connected in a kinematic tree structure. BlendSCAPE parameterizes the deformations that fit T^* to a scan S^k into a set of pose parameters θ^k and a set of shape (or "identity") parameters D^{p_k}: θ^k collects the relative rotations between neighboring kinematic tree parts, represented as Rodrigues vectors; D^{p_k} defines subject-specific deformations corresponding to the person's body shape. These deformations are applied to each triangle f of T^*.

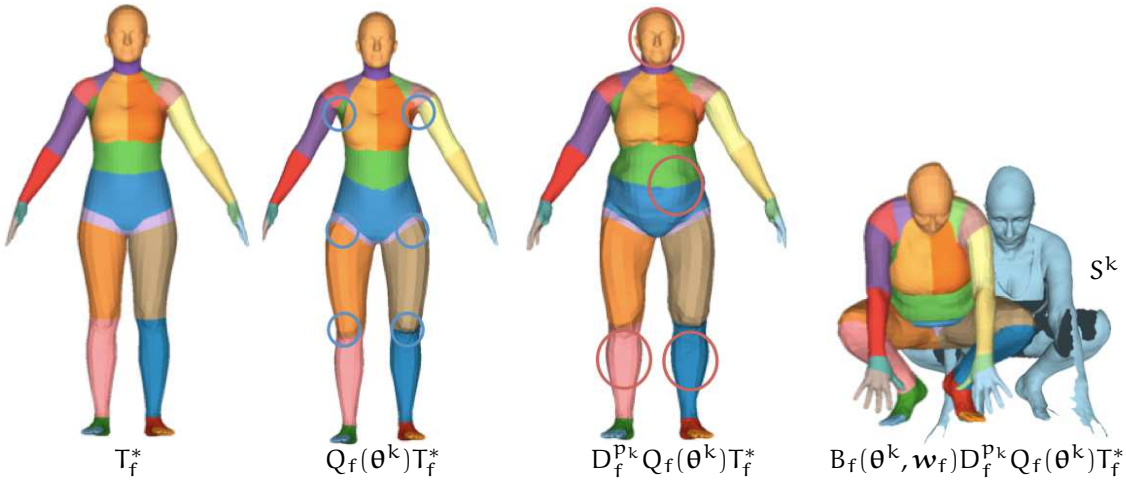


Figure 12: BlendSCAPE models the deformations that fit a template T^* to a scan S^k as a sequence of pose- and shape-dependent transformations, applied to each triangle T_f^* of T^*. Colors encode different segments of the underlying kinematic tree structure.

During alignment, T^* is first unstitched into disconnected triangles, T_f^*; each unstitched triangle is represented by a pair of its edge vectors, "forgetting" its location but retaining its shape and orientation. Then, each triangle is individually fit according to a sequence of pose- and shape-dependent deformations, represented with 3×3 linear transformation matrices. Namely, a transformed triangle $\mathcal{M}_f$ is obtained as

$$\mathcal{M}_f = B_f(\theta^k, \boldsymbol{w}_f) D_f^{p_k} Q_f(\theta^k) T_f^* \tag{3.1}$$

where $B_f(\theta^k, \boldsymbol{w}_f) = \sum_i w_{f,i} R^i(\theta^k)$ is a linear blend of rigid rotations $R^i(\theta^k)$ of body parts i, and $D_f^{p_k}$ and $Q_f(\theta^k)$ account for non-rigid deformations dependent on the subject's identity and on the pose, respectively.

We denote by W the set of blending weights for all the N_{trg} triangles in T^*: $W = \{\boldsymbol{w}_f : f = 1, \ldots, N_{trg}\}$. Analogously, Q and D^{p_k} collect the set of pose-dependent and shape-dependent per-triangle deformations: $Q = \{Q_f : f = 1, \ldots, N_{trg}\}$ and $D^{p_k} = \{D_f^{p_k} : f = 1, \ldots, N_{trg}\}$.

As in [20], Q is a linear function of the pose vector: $Q(\theta^k) = Q^0 + \sum_m \theta_m^k Q^m$, where θ_m^k is the m^{th} element of $\boldsymbol{\theta}^k$ and $Q^0, \{Q^m\}$ contain the linear coefficients, that are learned during coregistration. While in [68] the blending weights W are fixed, we optimize them together with D^{p_k} and Q.

After deformation, the disconnected triangles are stitched into a watertight mesh, $\mathcal{M}(\theta^k, D^{p_k}, Q, W)$, by solving for vertex positions via least squares (as in [20]). Figure 12 provides an intuition of how each transformation contributes to the final deformed mesh.

3.3.2 Geometry-based coregistration

The goal of the first stage is to perform a first round of registration and simultaneously train the BlendSCAPE model.

Given a corpus of scans $\{S^k : k = 1, \ldots, N_{scans}\}$ of N_{sbj} subjects in multiple poses, we optimize for:

- a set of alignments (deformed templates) $\{T^k : k = 1, \ldots, N_{scans}\}$
- a set of body model parameters:
  - a set of pose vectors $\{\theta^k : k = 1, \ldots, N_{scans}\}$;
  - a set of shape-dependent deformations $\{D^p : p = 1, \ldots, N_{sbj}\}$;
  - a set of pose-dependent deformations Q;
  - a set of blending weights W.

Note that, as mentioned above, we use the model as a regularizer during optimization: consequently, T^k is encouraged to be similar (but not necessarily equal) to $\mathcal{M}(\theta^k, D^{p_k}, Q, W)$.

Coregistration defines a unique objective function, that combines a geometry-based data term with a set of regularization terms; the data term penalizes distance in 3D between scan and template surfaces, while the regularizers "couple" the alignment to the learned model and encourage smoothness and rigidity of the deformations [68]. We now analyze each term in more detail.

The data term E_S evaluates the distance in 3D between the surface of the deformed template T^k and that of the scan S^k:

$$E_S(T^k; S^k) = \int_{x_s \in S^k} \rho\left(\min_{x_t \in T^k} \|x_s - x_t\|\right) \tag{3.2}$$

where ρ is the Geman-McClure robustifier [58]. The integral in Eq. (3.2) is approximated using a set of fixed locations x_s uniformly sampled over the surface of the scan.

In addition to E_S, we define four regularization terms: a "coupling" term E_{cpl}, a smoothness prior E_D, a rigidity prior E_Q and a pose prior E_θ.

The coupling term E_{cpl} penalizes discrepancy between the aligned template and the current model:

$$E_{cpl}(T^k, \theta^k, D^{p_k}, Q, W) = \sum_{\text{triangle } f} \|T_f^k - B_f(\theta^k, w_f) D_f^{p_k} Q_f(\theta^k) T_f^*\|_F^2. \tag{3.3}$$

The term evaluates the Frobenius distance between the pair of edge vectors T_f^k of the unstitched triangle of T^k, and the corresponding pair in the current posed model $\mathcal{M}(\theta^k, D^{p_k}, Q, W)$.

E_D promotes spatial smoothness of the shape deformations:

$$E_D(D^p) = \sum_{\text{adjacent triangles } f,f'} \|D_f^p - D_{f'}^p\|_F^2. \tag{3.4}$$

The rigidity term E_Q damps the pose-dependent deformations:

$$E_Q(Q) = \sum_{\text{triangle } f} \left(\|Q_f^0 - I_{3\times 3}\|_F^2 + \sum_m \|Q_f^m\|_F^2 \right) \tag{3.5}$$

where $I_{3\times 3}$ is the identity matrix.

Finally, the pose prior E_θ penalizes the squared Mahalanobis distance from a mean pose μ_θ:

$$E_\theta(\theta^k) = (\theta^k - \mu_\theta)^T \Sigma_\theta^{-1} (\theta^k - \mu_\theta). \tag{3.6}$$

We compute μ_θ and Σ_θ from a corpus of about 2000 scans of different people in a wide range of poses, pre-registered with the technique in [68].

Summarizing, given a corpus of scans of different people we obtain a set of preliminary alignments and learn a model of shape-dependent and pose-dependent deformations by minimizing the following objective function E_{coreg}:

$$\begin{aligned} &E_{coreg}(\{T^k\}, \{\theta^k\}, \{D^p\}, Q, W; \{S^k\}) = \\ &\sum_{\text{scan } k} \lambda_S E_S(T^k; S^k) + \\ &\sum_{\text{scan } k} (\lambda_{cpl} E_{cpl}(T^k, \theta^k, D^{p_k}, Q, W) + \lambda_\theta E_\theta(\theta^k)) + \\ &\lambda_Q E_Q(Q) + \sum_{\text{subject } p} \lambda_D E_D(D^p) \end{aligned} \tag{3.7}$$

where $\lambda_S, \lambda_{cpl}, \lambda_\theta, \lambda_Q$ and λ_D are weights for the different addends, and p_k identifies the subject in each scan.

In Eq. (3.7), the optimization is guided by the geometric data term. As a result, we obtain alignments that fit very closely the surface of the scans: the Euclidean distance between any scan vertex and its closest point on the surface of the corresponding alignment in no more than a few millimeters (see Fig. 13).

However, this is not sufficient to ensure consistent correspondence in areas where the scan surface provides no high-frequency geometric information. Figure 13 exemplifies the problem. We select the same vertex on two alignments relative to the same

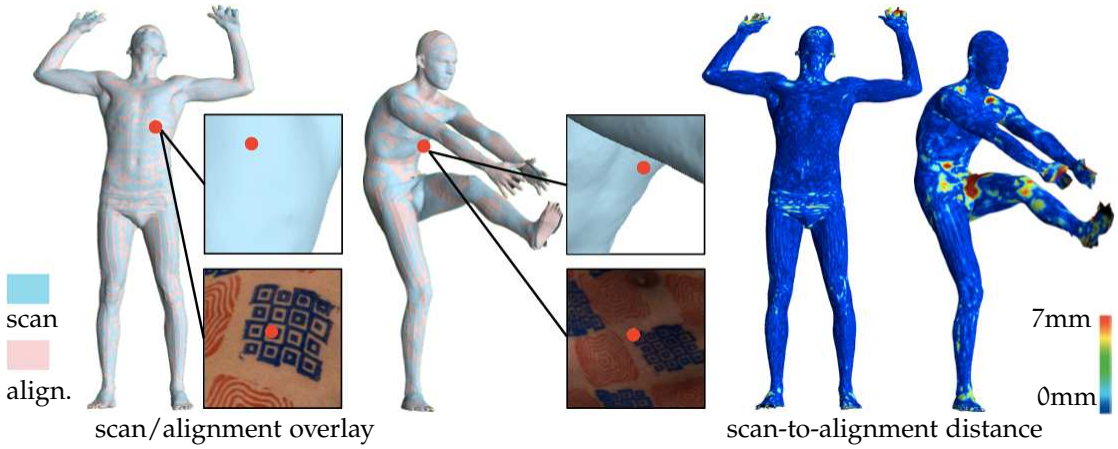


Figure 13: Alignments obtained minimizing the geometry-based objective function (3.7). On the left, we show two alignments and the corresponding scans; heat maps on the right encode the distance in 3D between scan and alignment surfaces (we compute the Euclidean distance between each scan vertex and its closest point on the alignment surface). Even though the deformed template fits very closely the scan surface, this does not ensure accurate correspondences in flat surface areas. We solve geometry ambiguities using appearance information: projecting the same template vertex (shown in red) on image space, we see that it maps to two different points on the subject's body.

subject, and project it onto the surface of the corresponding scan; since the alignments share the same topology, the projected points should map to the same point on the torso of the subject. In that area, however, the surface of the scans is smooth, making geometry information ambiguous; projecting the vertex on image space, we can verify that the correspondence is inconsistent.

In [68], the authors address this problem with a landmark-based error term. However, it is not clear how to precisely landmark smooth areas – exactly the places where landmarks are needed.

Our solution uses dense texture information: we estimate an appearance model for each subject and use this to define a novel appearance-based error term.

3.4 APPEARANCE MODEL

Optimizing objective (3.7) provides us with a set of initial alignments of all the scans in the corpus. These alignments are sufficient to build an initial subject-specific appearance model. To this end, we assume that the albedo of a subject is *consistent* across

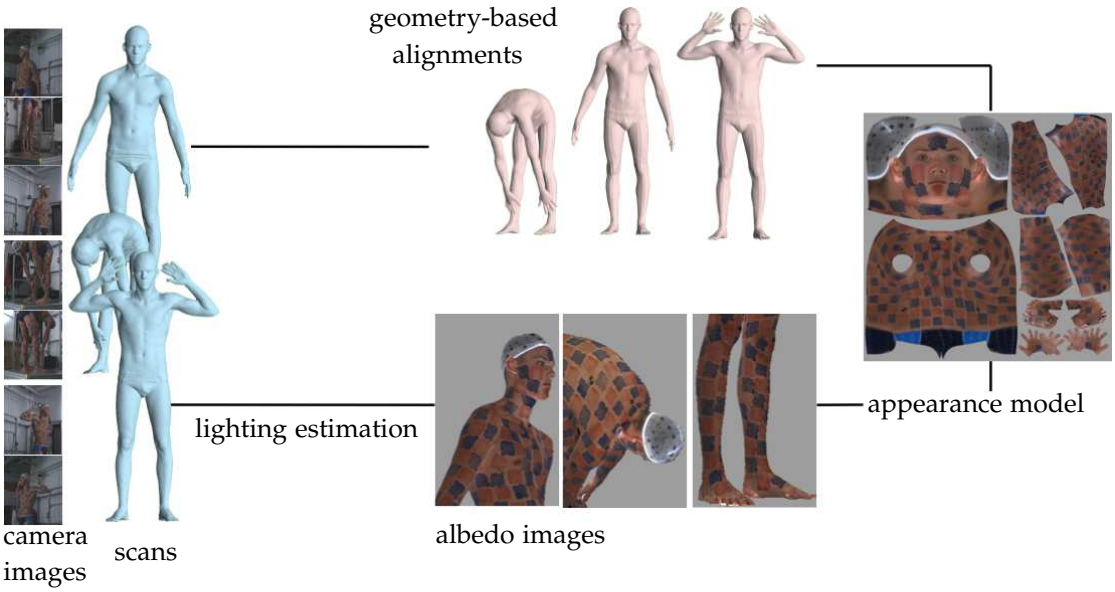


Figure 14: Creation of a per-subject appearance model. Synchronized with each 3D scan, we have 22 color cameras capturing images of the body from different views. We estimate scene lighting and preprocess the original images to remove shadows, obtaining a set of albedo images. We use these images and the initial geometry-based alignments to create a per-subject appearance model, represented as a texture map.

scans [24] – as is his shape D^p. Our key idea is to create a per-subject albedo model U^p, refining each alignment so that the estimated appearance model matches the observed scan appearance.

Figure 14 provides an overview of our approach. Synchronized with each 3D scan we have 22 color cameras capturing images of the body from different views. We denote by $I^k = \{I^k_j : j = 1, \ldots, 22\}$ the set of 22 images associated to scan S^k; furthermore, we denote by C^k the set of camera calibration parameters associated to I^k: $C^k = \{C^k_j : j = 1, \ldots, 22\}$. Given the calibration parameters C^k_j of camera j, we can project any 3D surface point $\mathbf{x}$ onto a 2D point $\pi^k_j(\mathbf{x})$ in the image plane of camera j; $I^k_j[\pi^k_j(\mathbf{x})]$ returns $\mathbf{x}$'s color if $\mathbf{x}$ is visible in I^k_j.

As a preliminary step, we preprocess the original images to discriminate between albedo and shading; based on the preprocessing results and on the initial alignments, we create a per-subject appearance model, represented as a texture map. The following sections describe each step in detail.

3.4.1 Albedo estimation

To discriminate between albedo and shading in the original camera images, we need to define an adequate model for human skin reflectance and scene lighting. Similarly to previous work [23, 60, 136, 138], we assume Lambertian reflectance for human skin. Given our diffuse lighting environment, we model illumination as a function defined over the sphere, and project it onto a low-dimensional Spherical Harmonic (SH) basis [113] (see Section 2.3.2 for a brief review of these concepts).

Since human bodies are extended and articulated, it is critical to model self-casting shadows. To account for this, we adopt a technique commonly used in real-time rendering: given a scan surface point $\mathbf{x}$, we compute a corresponding *shadowed diffuse transfer* [113], that maps incoming to ongoing radiance taking into account surface reflectance at $\mathbf{x}$ *and* occlusion effects.

We first formalize the definition of transfer function, and show how to combine it with a SH lighting model; then, we introduce the objective function we minimize to obtain a plausible lighting estimate; finally, based on the estimated lighting, we explain how to extract a set of albedo images from the original ones.

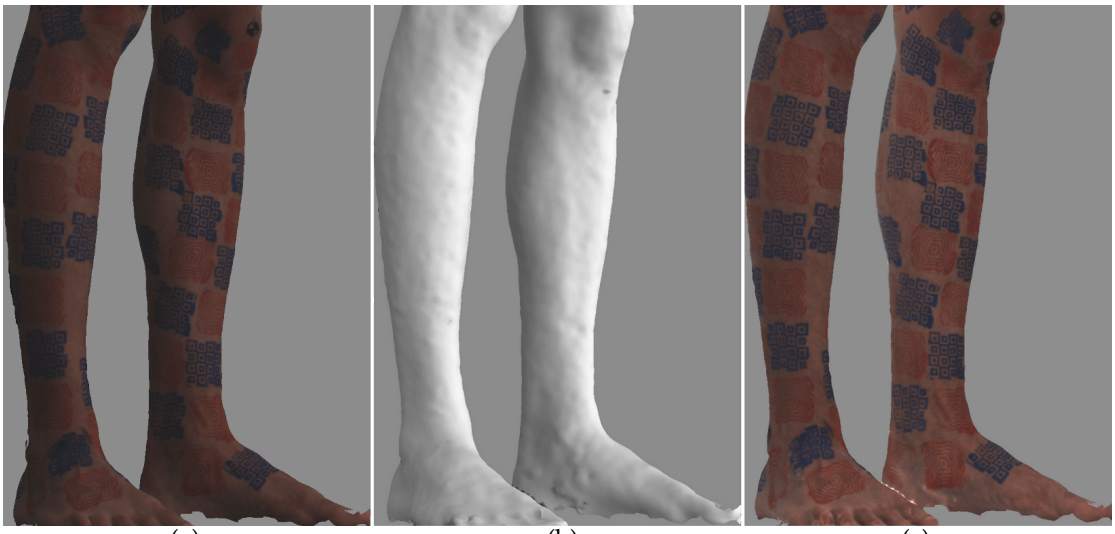
We denote by $L_x(\boldsymbol{\omega})$ the incoming light at $\mathbf{x}$ from direction $\boldsymbol{\omega}$, and by a_x and $\mathbf{n}_x$ the albedo and the surface normal at $\mathbf{x}$, respectively. The shadowed diffuse transfer $\mathcal{T}_x(L_x)$ at $\mathbf{x}$ is the function

$$\mathcal{T}_x(L_x) = \frac{a_x}{\pi} \int_{\omega \in \Omega} L_x(\boldsymbol{\omega}) \max(\mathbf{n}_x \cdot \boldsymbol{\omega}, 0) V_{\mathbf{x}}(\boldsymbol{\omega}) \tag{3.8}$$

where the integral is defined over the unit sphere Ω, and $V_{\mathbf{x}}(\boldsymbol{\omega})$ is a visibility function returning 1 if a ray from $\mathbf{x}$ in direction $\boldsymbol{\omega}$ fails to intersect the scan again (i.e. is unshadowed), 0 otherwise.

We assume that lighting variation over the body surface, not due to the presence of the body itself, is negligible (i.e. we assume that $L_{\mathbf{x}}(\boldsymbol{\omega}) \approx L_{\mathbf{x}'}(\boldsymbol{\omega})$, for any pair of surface points $\mathbf{x}$, $\mathbf{x}'$). On the right side of Eq. (3.8), we have a product between two functions, one depending on lighting and one depending on scan geometry: $L(\boldsymbol{\omega})$ and $\max(\mathbf{n}_x \cdot \boldsymbol{\omega}, 0) V_{\mathbf{x}}(\boldsymbol{\omega})$. By projecting both functions on a SH basis separately, their product reduces to a dot product of their coefficient vectors. In our approach we work in a 9-dimensional SH space, obtaining two 9-element vectors: a light vector $\mathbf{l}_{SH}$ and a *transfer vector* $\boldsymbol{\tau}_{\mathbf{x}}$, depending only on scan geometry.

We can now express the observed scan surface color as a function of SH coefficient vectors – namely, as a function of $\boldsymbol{\tau}_{\mathbf{x}}$ and $\mathbf{l}_{SH}$. Note that we work on each RGB



(a) (b) (c)

Figure 15: Results of light estimation in one camera image (original background not shown): (a) observed color, (b) shading, (c) albedo.

channel separately; for simplicity, we omit channel indices in our notation. Given a generic scan surface point $\mathbf{x}$, its color i_x and its albedo a_x are related as:

$$i_x = (\boldsymbol{\tau}_x \cdot \mathbf{l}_{SH}) a_x. \tag{3.9}$$

In Eq. (3.9), both albedo and lighting are unknown. We initialize the albedo computing the average color i_{avg} over the vertices of all the scans in the corpus, and estimate $\mathbf{l}_{SH}$ by minimizing

$$E_{light}(\mathbf{l}_{SH}; \{S^k\}, \{I^k\}, \{C^k\}) = \sum_{\text{scan } k} \sum_{\text{camera } j} \sum_{\text{vertex } h} V_{\mathbf{v}_h^k}(C_j^k) \| I_j^k[\pi_j^k(\mathbf{v}_h^k)] - (\boldsymbol{\tau}_{\mathbf{v}_h^k} \cdot \mathbf{l}_{SH}) i_{avg} \|^2 \tag{3.10}$$

where, like in Eq. (3.8), $V_{\mathbf{v}_h^k}(C_j^k)$ is a visibility function returning 1 if $\mathbf{v}_h^k$ is visible from a camera with parameters C_j^k, 0 otherwise. Recall that $I_j^k[\pi_j^k(\mathbf{v}_h^k)]$ denotes $\mathbf{v}_h^k$'s color as observed in image I_j^k.

We use $\mathbf{l}_{SH}$ to obtain a set of albedo images $A^k = \{A_j^k : j = 1, \ldots, 22\}$ from the original camera images I^k. Given $\mathbf{l}_{SH}$, we calculate the shading at vertex $\mathbf{v}_h^k$ as $(\boldsymbol{\tau}_{\mathbf{v}_h^k} \cdot \mathbf{l}_{SH})$; at a generic scan surface point $\mathbf{x}$ this is given by interpolating between vertices belonging to the same triangle. An albedo image A_j^k is then computed, for any pixel $\mathbf{y}$ with corresponding surface point $\mathbf{x}$ such that $\mathbf{y} = \pi_j^k(\mathbf{x})$, as

$$A_j^k[\mathbf{y}] = I_j^k[\mathbf{y}] / (\boldsymbol{\tau}_x \cdot \mathbf{l}_{SH}) \tag{3.11}$$

Figure 15 shows the results of our lighting estimation in one camera image.

3.4.2 From camera images to texture maps

Given the albedo images for each scan, we seek a per-subject appearance model U^p (represented as a texture map) that is consistent with all the scans of that particular subject (see Fig. 14). To this end, we first compute a texture map U^k for each alignment T^k; then, we combine the information given by all the maps corresponding to the same subject into a single appearance model.

We follow an approach that is quite standard in image-based texture reconstruction (see Section 2.3.3). We assume we have a 2D parameterization (a texture space) of the 3D surface of our template T^*. For any template surface point $\mathbf{x}$, we denote by $uv(\mathbf{x})$ its mapping from 3D to texture space.

Given an alignment T^k and the corresponding set of albedo images A^k, we obtain a texture map U^k by blending the contribution of each view on a per-texel basis; specifically, for each texel $\mathbf{y} = uv(\mathbf{x}_y)$ we weight the contribution of each view by computing the dot product between surface normal and viewing direction:

$$U^k[\mathbf{y}] = \frac{\sum_{\text{camera } j} V_{\mathbf{x}_y}(C_j^k) A_j^k[\pi_j^k(\mathbf{x}_y)] \max(\zeta_{\mathbf{x}_y}(C_j^k) \cdot \mathbf{n}_{\mathbf{x}_y}, 0)}{\sum_{\text{camera } j} V_{\mathbf{x}_y}(C_j^k) \max(\zeta_{\mathbf{x}_y}(C_j^k) \cdot \mathbf{n}_{\mathbf{x}_y}, 0)} \tag{3.12}$$

where $V_{\mathbf{x}_y}(C_j^k)$ is the visibility function returning 1 if $\mathbf{x}_y$ is visible from camera j, 0 otherwise, $\mathbf{n}_{\mathbf{x}_y}$ is the surface normal at $\mathbf{x}_y$ and $\zeta_{\mathbf{x}_y}(C_j^k)$ is the direction from $\mathbf{x}_y$ to the center of camera j.

Texture maps computed from single alignments are usually noisy and incomplete, since no single pose can provide full-body coverage. To compute a per-subject appearance model, we therefore integrate information over multiple per-subject poses: for each subject p, we initialize an appearance model U^p by averaging the set of per-alignment maps relative to p.

3.5 APPEARANCE-BASED ERROR TERM

The previous sections described how, given a corpus of scans of different people, we obtain a set of preliminary alignments and an appearance model for each subject in the corpus. We combine these components to define a novel appearance-based error term.

Figure 16 outlines our approach. Given an alignment T^k, an appearance model U^{p_k} and a set of camera calibration parameters C^k, we can render a set of synthetic images $\overline{A}^k = \{\overline{A}_j^k : j = 1, \ldots, 22\}$. Informally, a synthetic image $\overline{A}_j^k$ is an "estimate", based on

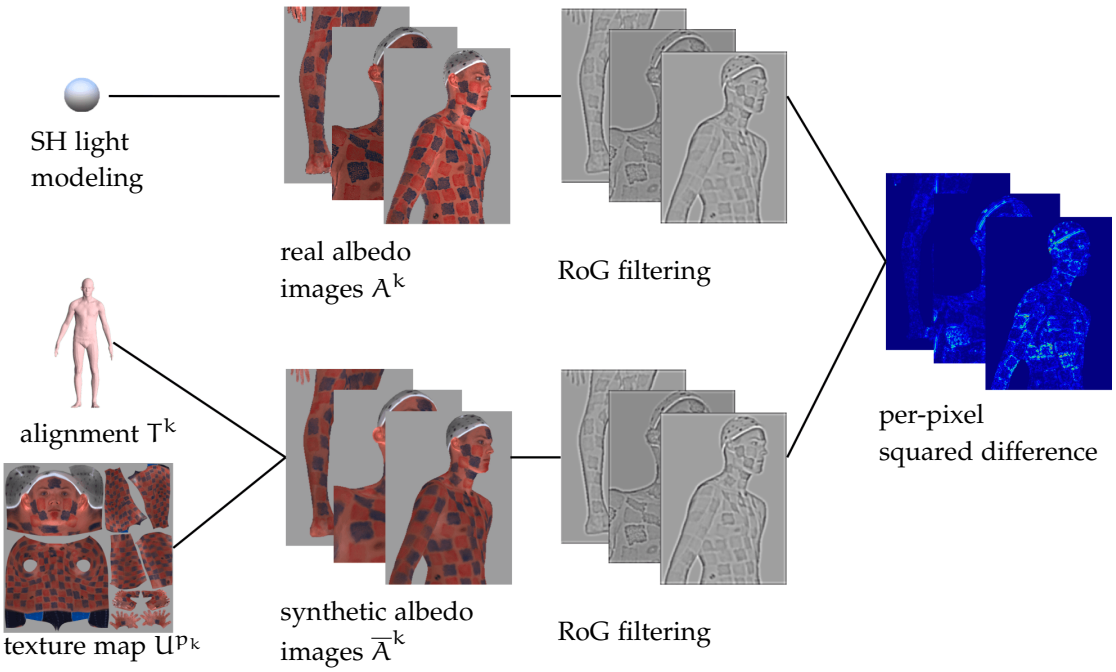


Figure 16: Our appearance-based error term penalizes differences between real albedo images, obtained after light preprocessing, and synthetic images, rendered from the model. We introduce a form of contrast normalization (RoG filtering) to account for inaccuracies in lighting estimation.

our alignment and model parameters, of the real albedo image A_j^k. The higher the similarity between synthetic and real images, the better our estimate is.

To define our appearance-based error term, we simply translate this intuition into a mathematical formulation. The term compares real albedo images against a set of synthetic ones, rendered from the model. We work on each RGB channel separately; for simplicity, we omit channel indices in our notation. Note also that we do not represent image background. Let $F_j^k(T^k)$ be the intersection between the foreground masks of A_j^k and $\overline{A}_j^k$; we can evaluate the discrepancy between A_j^k and $\overline{A}_j^k$ by computing a sum of squared differences between pixels in $F_j^k(T^k)$:

$$\sum_{\substack{\text{pixel } \mathbf{y}: \\ \mathbf{y} \in F_j^k(T^k)}} \|A_j^k[\mathbf{y}] - \overline{A}_j^k[\mathbf{y}]\|^2. \tag{3.13}$$

We observed that a direct computation of Eq. (3.13) may be problematic for some images: in some cases, shading artifacts may be still present due to inaccuracies in lighting estimation. To address this problem, we modify our formulation introducing a form of contrast normalization.

Before computing per-pixel squared differences, we apply a Ratio-of-Gaussians (RoG) filter to both synthetic and real images. The definition of a RoG filter of parameters σ_1, σ_2 is straightforward, and we include it for completeness. For a generic grayscale image I_{bw}, the corresponding RoG image $G_{\sigma_1,\sigma_2}(I_{bw})$ is computed as:

$$G_{\sigma_1,\sigma_2}(I_{bw})[\mathbf{y}] = \frac{I_{bw} * \frac{e^{-\|\mathbf{y}\|^2/2\sigma_1^2}}{2\pi\sigma_1^2}}{I_{bw} * \frac{e^{-\|\mathbf{y}\|^2/2\sigma_2^2}}{2\pi\sigma_2^2}} \tag{3.14}$$

where $*$ denotes the convolution operator and $\sigma_2 > \sigma_1$.

Including RoG filtering in Eq. (3.13), and summing over multiple images, we obtain the appearance-based error term E_U:

$$E_U(T^k; A^k, C^k, U^{p_k}) = \sum_{\text{camera } j} \sum_{\substack{\text{pixel } \mathbf{y}: \\ \mathbf{y} \in F_j^k(T^k)}} \|G(A_j^k)[\mathbf{y}] - G(\overline{A}_j^k)[\mathbf{y}]\|^2. \tag{3.15}$$

Note that E_U is computed over each RGB channel separately.

E_U can be incorporated into the original coregistration formulation as an additional data term, that complements the geometry-based data term introduced in Eq. (3.2). We define our objective function based on both geometry and appearance information as:

$$\begin{aligned} &E_{app-coreg}(\{T^k\},\{\theta^k\};\{S^k\},\{D^p\},Q,W,\{A^k\},\{C^k\},\{U^p\}) = \\ &\sum_{\text{scan } k} \lambda_S E_S(T^k; S^k) + \\ &\sum_{\text{scan } k} \lambda_U E_U(T^k; A^k, C^k, U^{p_k}) + \\ &\sum_{\text{scan } k} (\lambda_{cpl} E_{cpl}(T^k, \theta^k; D^{p_k}, Q, W) + \lambda_\theta E_\theta(\theta^k)) \end{aligned} \tag{3.16}$$

where, as in Eq. (3.7), λ_S, λ_U, λ_{cpl} and λ_θ are weighting coefficients; E_S, E_{cpl} and E_θ correspond to the error terms introduced in Eqs. (3.2), (3.3) and (3.6).

Objectives (3.7) and (3.16) are non-linear and exhibit a high-dimensional space of solutions. Details about their optimization are provided in the next section.

3.6 OPTIMIZATION

For the first phase (i.e. for minimizing objective (3.7)), our approach is similar to that proposed in [68]; we present it here for completeness.

We consider two separate subproblems, optimizing $\{T^k\}$ and $\{\theta^k\}$ first, and then W, $\{D^p\}$ and Q. Fixing W, $\{D^p\}$ and Q decouples the scans: we consider N_{scans} registration subproblems and minimize $\lambda_S E_S(T^k; S^k) + \lambda_{cpl} E_{cpl}(T^k, \theta^k, D^{p_k}, Q, W) + \lambda_\theta E_\theta(\theta^k)$ for each scan S^k separately. We use Powell's dogleg method [96] with Gauss-Newton Hessian approximation.

With $\{T^k\}$, $\{\theta^k\}$, $\{D^p\}$ and Q fixed, we solve for W via non-negative least squares, encouraging the sum of per-triangle blending weights $\sum_i w_{f,i}$ to be 1. Keeping fixed all parameters but $\{D^p\}$, the minimization with respect to each subject's D^p is an independent linear least squares problem for each subject p. Similarly, with all parameters but Q fixed, minimization with respect to Q_f is a linear regression problem (from pose parameters to pose-dependent deformations) independent for each triangle f.

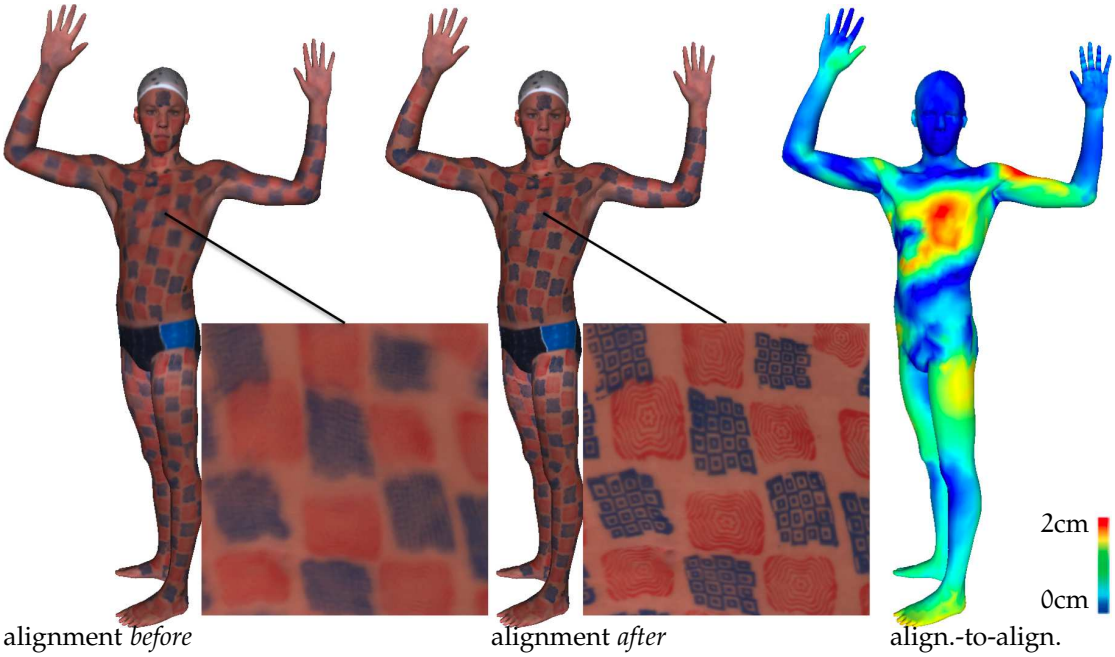
We initialize coregistration by fitting to each scan the BlendSCAPE model trained in [68], using E_S and a strong pose prior E_θ. During the model fit, shape is allowed to vary only within a low-dimensional shape space; we learn this space from almost 4000 registered scans from the CAESAR dataset [107] (more details about this shape space are provided in Section 6.3.1). For each subject p, D^p is initialized by averaging the shape of the alignments relative to p.

In the second stage, we use the set of initial geometry-based alignments to compute an appearance model U^p for each subject p. We compute texture maps at a resolution of 2048×2048 texels. We refine each alignment T^k by minimizing objective (3.16) for each scan separately. A single alignment, optimizing simultaneously over 22 images (of size 612×512 each), took less than 5 minutes on a common desktop machine.

Note that, after the refinement stage, one could compute again the model parameters $\{D^p\}$, Q, W and $\{U^p\}$. We observed that an iterative coarse-to-fine approach, in which the variance of both Gaussians in Eq. (3.13) becomes progressively narrower, leads to more accurate results. In our experiments, we ran two iterations; σ_1 and σ_2 decreased from 4 to 2 and from 8 to 4, respectively.

The ratio between λ_{cpl} and λ_U in Eq. (3.16) turned out to be a crucial parameter; we set it equal to 25 in the first iteration, and to 15 in the second one.

We observed good intra-subject consistency without the use of any "geometric" landmarks, by relying on a strong pose prior E_θ. However, this did not provide fully satisfactory inter-subject correspondence. In the absence of any constraint, D can induce different deformations in different subjects. We therefore introduced a weak landmark error term in the first phase, decreasing its weight progressively over several iterations. No landmarks were used in the second phase.



alignment *before* appearance-based refinement | alignment *after* appearance-based refinement | align.-to-align. distance

Figure 17: Example comparison between alignments obtained before and after appearance-based refinement. For each alignment, we show details of the corresponding appearance models. The heat map shows the Euclidean distance per vertex between the two alignments. Appearance information prevents sliding in geometrically smooth areas, producing more accurate correspondences and therefore sharper estimated models.

3.7 RESULTS AND DISCUSSION

In this chapter we have presented a novel registration technique for 3D human meshes. Our technique combines a classic geometry-based error term with a novel appearance-based error term, that uses dense surface texture information to solve correspondence ambiguities in geometrically smooth areas.

Figure 17 illustrates the benefits of our approach. Texture information adjusts vertex placement mostly in smooth 3D areas (like the stomach and back), complementing the partial or ambiguous information provided by geometry. Computing the Euclidean distance per vertex between the alignments obtained before and after appearance-based refinement, we noticed that such distance can be on the order of 1 or even 2cm.

Using a learned appearance model improves intra-subject correspondences between scans, resulting in much sharper appearance models (see again Fig. 17).

In Chapter 4, we will introduce a precise quantitative metric to assess the accuracy of our alignments; based on that, we will perform an extensive evaluation, compare geometry-based and appearance-based registration from a quantitative point of view, and discuss current limitations of our approach.

In this chapter we have relied on two assumptions. First, we painted the bodies of the subjects with a high-frequency texture pattern. Without painting, human skin provides much less texture information. It is therefore unclear how well our technique would perform. Second, we captured scans with a high-quality multi-camera system that provides, at each capture, almost full-body coverage. Ideally, the robustness of our approach should be tested using more "lightweight" systems, equipped with fewer cameras and producing lower-quality 3D data.

In Chapter 5 we will remove the first assumption, adapting our technique to work on naked bodies without any artificial texture pattern applied on them. In Chapter 6, we will extend our technique to deal with dynamic monocular sequences captured with a single Kinect camera. This will allow us to evaluate accuracy and robustness of our approach when registering lower-quality, noisier 3D and color data.

4 THE FAUST DATASET

Despite the rich literature on 3D surface registration, datasets and benchmarks are scarce. This lack is mainly due to the difficulty of dealing with real data.

The popular TOSCA dataset [38] contains synthetic meshes of fixed topology with artist-defined deformations. The SHREC benchmark [40] adds a variety of artificial noise to TOSCA meshes, but both artificial noise and meshes and deformation models created by artists lack sufficient realism.

To advance the field, datasets and benchmarks should contain noisy, realistically deforming meshes that vary in topology: this is the data real-world applications deal with. The problem is to establish dense ground-truth correspondences, and therefore a reliable evaluation metric, on such meshes. Common approaches like manual landmarking are time consuming, challenging and error-prone for humans – and provide only sparse correspondences.

In this chapter we introduce FAUST (Fine Alignment Using Scan Texture), a novel dataset and benchmark for 3D mesh registration. FAUST collects real scans of different people in multiple poses, with automatically computed dense ground-truth correspondences.

To build the dataset, we leverage our appearance-based registration technique. We remove ambiguities in the definition of dense correspondences between scans by painting the subjects with a high-frequency texture, and placing textured markers on key anatomical locations; then, we automatically compute dense scan-to-scan correspondences by aligning a common template to each scan. To ensure accurate ground truth, we verify the quality of our alignments through extensive validation.

This chapter describes our dataset in detail. Section 4.1 gives a brief overview on existing datasets for 3D mesh registration. Section 4.2 describes how we captured the data, presents our ground-truth creation process and introduces the FAUST benchmark. Benchmark results for different state-of-the-art registration techniques are provided in Section 4.3. Finally, Section 4.4 discusses the role played by texture information in building our dataset, and identifies current limitations of our approach.

4.1 PREVIOUS DATASETS

We review the most widely used datasets for 3D mesh registration dividing them into two classes, synthetic and real, based on the type of data they provide.

4.1.1 Synthetic data

Synthetic datasets collect meshes generated in an artificial fashion. Since establishing ground-truth correspondences between these meshes is straightforward, datasets in this class are widely used to evaluate 3D registration methods. In the following, we focus on three of them: TOSCA [38], the dataset used in the SHREC '10 robust correspondence benchmark [40] and the "kids" dataset recently proposed in [108].

TOSCA [38] collects 80 meshes of animals and people (with 3 subjects in a dozen different poses each); each mesh contains about 50000 vertices, and meshes in the same class share the same topology. There exists also a lower-resolution version of TOSCA (Non-rigid world 3D database [38]), that collects meshes with a resolution of approximately 3000 vertices.

The dataset is very popular, and has been used to evaluate several 3D registration techniques [39, 52, 78, 87, 98, 108]. Nonetheless, a major limitation of TOSCA is its lack of realism. Figure 18 illustrates some differences between TOSCA meshes and scans captured with our full-body 3D multi-stereo system: body deformations in synthetic meshes look artificial, and there is no noise or missing data.

The SHREC '10 robust correspondence benchmark [40] collects a subset of TOSCA meshes and adds to them a variety of noise (see Fig. 19): small and big holes, scaling, affine transformation, geometric Gaussian noise and shot noise, sampling, partial (missing parts), rasterization (simulating non-pointwise topological artifacts due to occlusions in 3D geometry acquisition) and view (simulating missing parts due to 3D acquisition artifacts). Yet, this artificial noise does not seem similar enough to the noise found in real-world applications.

The dataset proposed in [108] contains a set of 3D human meshes undergoing nearly-isometric and within-class deformations; namely, it provides two shape classes ("kid" and "fat kid") under different poses, where the same poses are applied to both classes. All the meshes have identical topology and compatible vertex ordering; the typical vertex count is about 60000. Unfortunately, the dataset exhibits the same limitations of TOSCA and SHREC '10: meshes are not representative of real-world data.

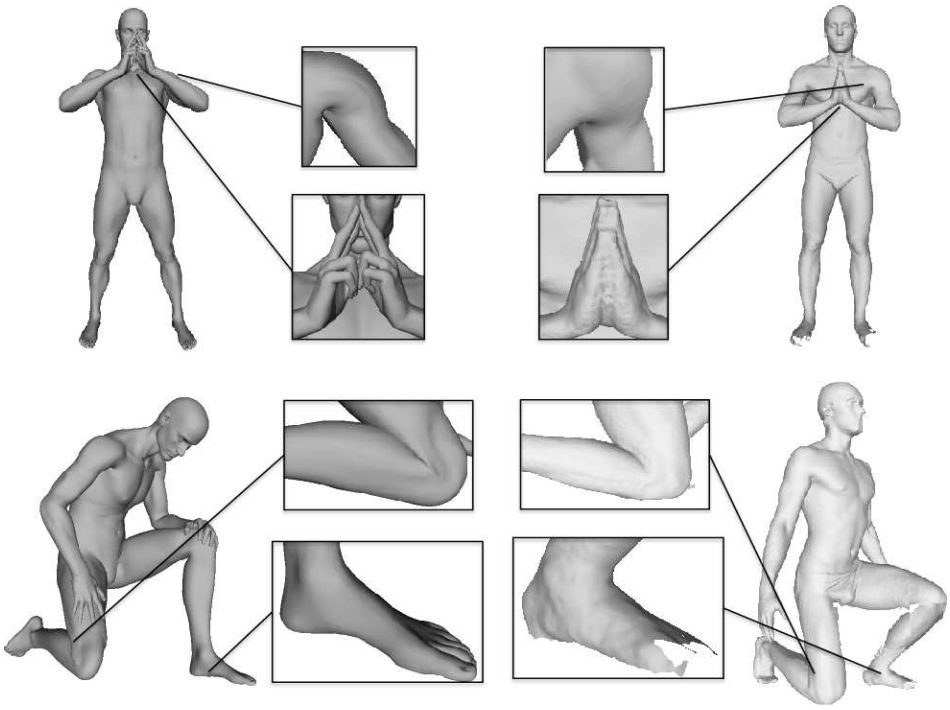


Figure 18: Comparison between two TOSCA [38] meshes (left) and two scans captured with our full-body 3D multi-stereo system. (right). Unrealistic deformations, plus the absence of noise and missing data, make TOSCA meshes not representative of real-world scans.

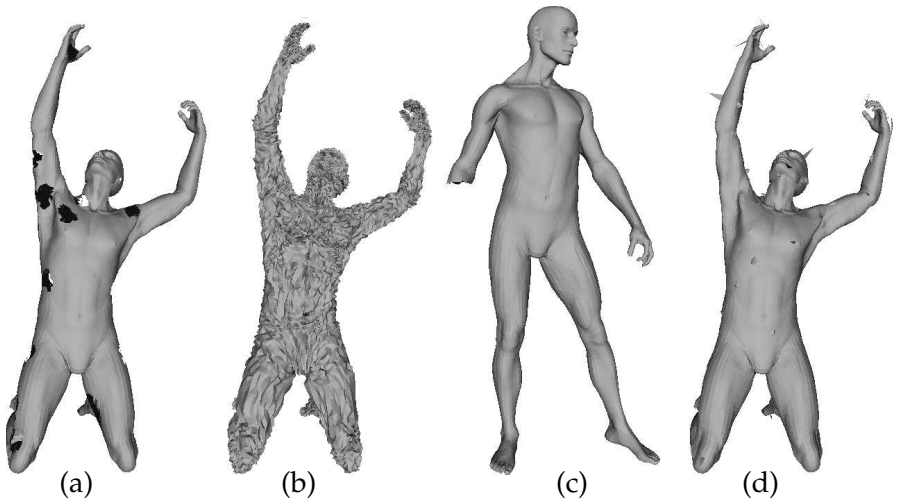


Figure 19: Examples of artificial noise added to TOSCA meshes in the SHREC '10 benchmark [40]: big holes (a), Gaussian noise (b), partial (c), shot noise (d).

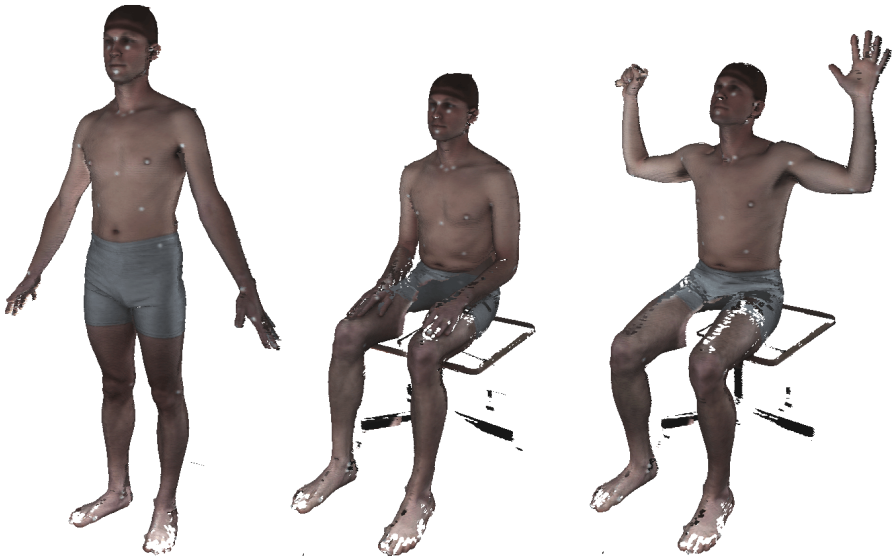


Figure 20: Three scans from CAESAR [107], exemplifying the three different poses included in the dataset. The white dots on the scans represent landmarks.

4.1.2 Real data

Establishing dense accurate correspondences between real scans is challenging; for this reason, real datasets usually do not provide dense ground truth. Three well-known datasets in this class are CAESAR [107], SCAPE [20] and Hasler's dataset [65].

CAESAR [107] contains several thousand laser scans of volunteers aged $18-65$ in the United States and Europe. Prior to scanning, 74 white markers were placed on the subjects at anthropometric locations, typically at points where bones can be palpated through the skin. Each subject was scanned in 3 different poses (see Fig. 20). CAESAR is widely used for alignment [14, 15, 20, 139] – though the only ground truth are sparse landmarks.

SCAPE [20] and Hasler's dataset [65] provide a set of registered scans.

The SCAPE dataset [20] contains 71 registered meshes of a single subject in different poses. The original scans were acquired with a laser scanner; a "template mesh" was selected among the scans, made watertight, and non-rigidly deformed to match the surface of the other scans. The process gives a set of meshes with the same topology, whose shape approximates well the surface of the original scans. These registrations have been used to evaluate surface matching algorithms [78, 87, 98]. Since the meshes are reconstructed from real scans, they are more realistic than synthetic data (e.g. they do not have exactly the same local shape features). However, they all share the same

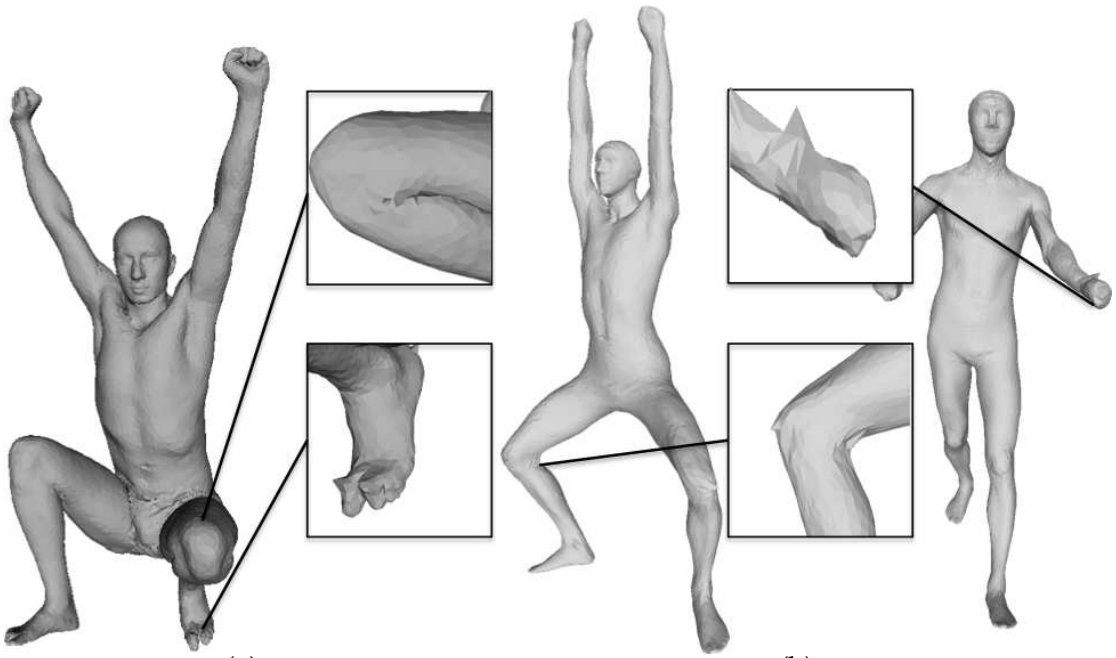


Figure 21: Meshes from the SCAPE dataset [20] (a) and from Hasler's dataset [65] (b). They exhibit artifacts in areas that are challenging to register (e.g. hands, feet, bent knees). In both datasets, the quality of the registrations is not evaluated in any manner; hence they cannot be trusted as ground truth.

topology and are watertight, simplifying the registration problem in comparison with raw scans.

Hasler et al. [65] provide a dataset of more than 500 laser scans of 114 different subjects (59 men and 55 women). All subjects were scanned in at least 9 poses selected randomly from a set of 34 poses. A common template has been fitted to each scan, and these alignments (together with the original scans) are publicly available.

In both datasets, registrations are obtained exploiting only geometry information; they exhibit artifacts in areas that are challenging to register (e.g. hands, feet, bent knees – see Fig. 21). Since the accuracy of these registrations is not quantified in any manner, they cannot be considered reliable ground truth.

4.2 BUILDING FAUST

FAUST addresses the limitations of previous datasets, providing *both* real scans of different subjects in multiple poses and reliable dense ground-truth correspondences between them. These correspondences are computed automatically, leveraging the registration technique presented in Chapter 3.

Sections 4.2.1, 4.2.2 and 4.2.3 describe in detail all the phases involved in the creation of the dataset: scan capture and registration, definition of reliable scan-to-scan correspondences based on a high-frequency texture pattern applied to the subjects' skin, validation of the quality of the registrations. Finally, Section 4.2.4 introduces the FAUST benchmark.

4.2.1 Scan capture and registration

The dataset collects 300 triangulated, non-watertight meshes of 10 subjects (5 male and 5 female), each scanned in 30 different poses. All the meshes were captured using the 3D multi-stereo system described in Section 3.2; the average mesh resolution is 172000 vertices.

The subjects are all professional models who have consented to have their data distributed for research purposes; their age ranges from a minimum of 18 to a maximum of 70. During the scan sessions they all wore identical, minimal clothing: tight fitting swimwear bottoms for men and women and a sports bra top for women. Figure 22 shows a set of example scans.

To compute scan-to-scan correspondences, we bring a common template into alignment with each scan using the technique described in Chapter 3. The alignments we obtain are watertight meshes with identical topology and resolution of 6890 vertices.

4.2.2 Painted bodies

It is essentially impossible, for both an algorithm and a human, to define *dense* ground-truth correspondences on a naked body's skin. Large uniformly-colored areas are uninformative, making the problem ill-posed. Note that unlike high-resolution face scans, we do not have sufficient resolution to see pores.

In order to provide high-frequency information across the whole body surface, we painted the skin of each subject. We applied body makeup of two different colors (red and blue) by using two woodcut stamps with different patterns (see Fig. 23). Each stamp has a surface of 45×45mm and pattern details up to 2mm in width.

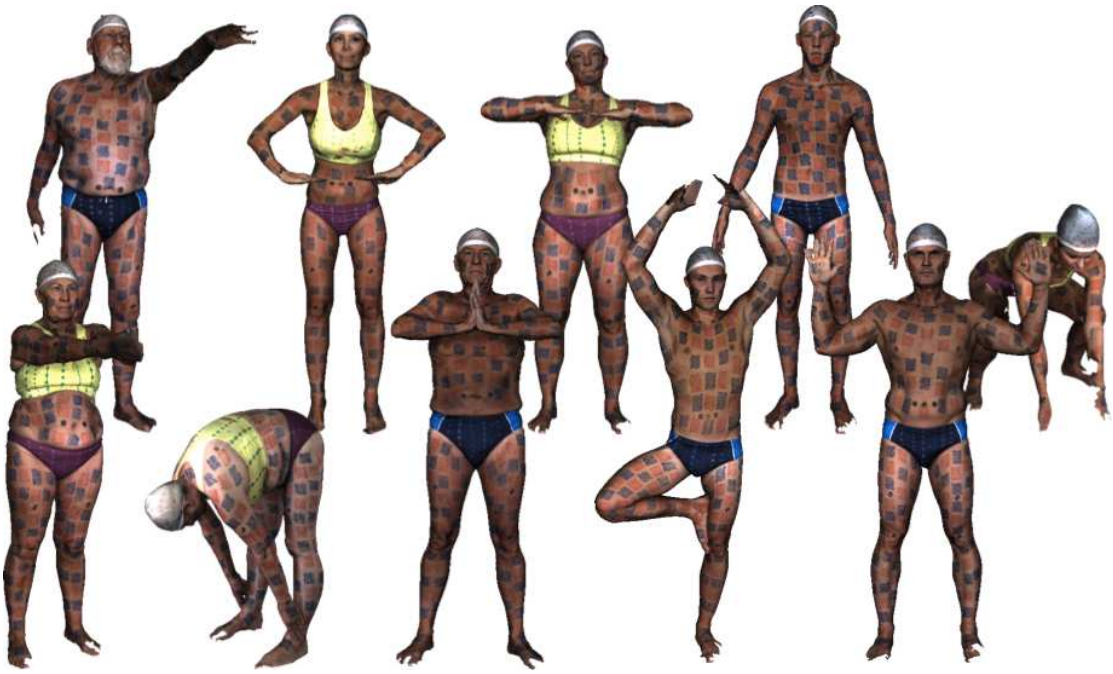


Figure 22: Example scans of all 10 subjects (all professional models) showing the range of ages and body shapes. A sampling of the poses shows the wide pose variation.

Figure 23: We define intra-subject correspondences by painting the skin of the subjects: we apply red and blue body makeup using two woodcut stamps with different patterns. To define inter-subject correspondences, we rely on a set of sparse landmarks drawn on key anatomical locations.

This painting provides reliable dense intra-subject correspondences. Between different subjects, we define only a set of *sparse* correspondences. Indeed, neither the natural texture of different people, nor our painted texture, can be matched across

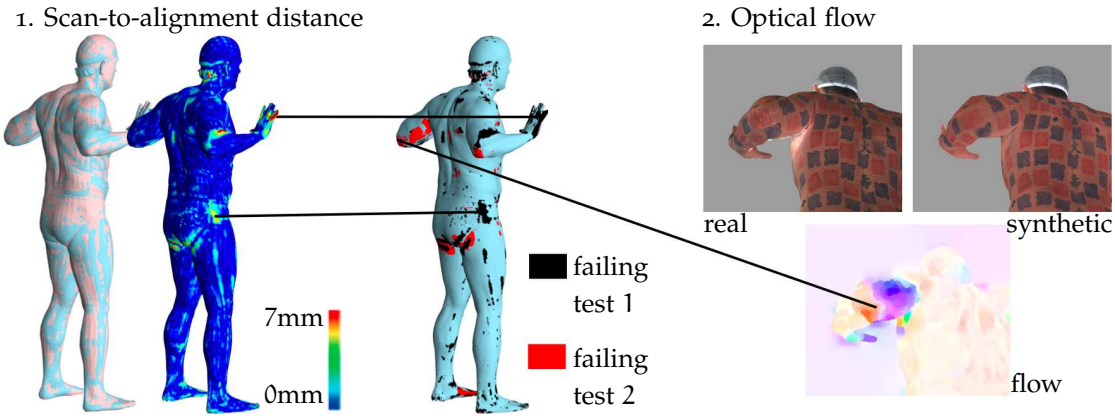


Figure 24: We evaluate the quality of our alignments in terms of both geometric error (measuring the distance in 3D between scan and alignment surfaces) and sliding error (measuring the optical flow between real and synthetic images). Scan vertices exhibiting an error higher than 2mm in either test are deemed misaligned.

subjects. And in general, correspondences across different body shapes may not be well defined – while key anatomical regions clearly can be matched, there are large non-rigid regions for which correspondences are less clear. To address this we took an approach that is common in the anthropometry and motion capture communities of identifying key landmarks on the body, and we used these to establish sparse correspondences. We drew a set of 17 easily identifiable landmarks on specific body points where bones are palpable; each landmark corresponds to a half-filled circle, with a diameter of approximately 2.5cm (see again Fig. 23).

4.2.3 Ground-truth correspondences

Our alignments implicitly define a set of scan-to-scan correspondences. Some correspondences are less reliable than others, since scans are noisy and incomplete and our alignments are the result of an optimization process. To ensure we have "ground truth", we identify vertices that are not aligned to an accuracy of 2mm using two tests: scan-to-alignment distance and optical flow between real and synthetic images (see Fig. 24).

1: Scan-to-alignment distance. Since all scans have been registered to a common template, we can compute the scan-to-scan correspondence between two scans, S^k and $S^{k'}$, as follows.

For any vertex v_h^k on S^k, find the closest point on the *surface* of the aligned template mesh T^k. Call this point $P_{T^k}(v_h^k)$. If the distance between v_h^k and $P_{T^k}(v_h^k)$ is greater

than a threshold, t_{dist}, we say that we are not able to provide any correspondence for v_h^k. Otherwise, we can uniquely identify $P_{T^k}(v_h^k)$ by a face index and a triplet of barycentric coordinates. Since T^k and $T^{k'}$ share the same topology, the same face and barycentric coordinates identify a point $P_{T^{k'}}(v_h^k)$ on $T^{k'}$. Given this point, we find the closest *point*, $P_{S^{k'}}(v_h^k)$, on the *surface* of scan $S^{k'}$.

Note our emphasis on the fact that this does not compute vertex-to-vertex correspondence but vertex-to-surface (mesh) correspondence.

If the distance between $P_{T^{k'}}(v_h^k)$ and $P_{S^{k'}}(v_h^k)$ is larger than t_{dist}, then we say that the vertex v_h^k on S^k does not have a corresponding point on $S^{k'}$. We take $t_{dist} = 2$mm.

2: Optical flow between real and synthetic images. Even scan vertices that are "near enough" to the alignment's surface can still suffer from sliding. This issue is ignored in most matching techniques, that simply rely on some surface distance metric for assessing correspondences. We quantitatively assess this sliding in image space by measuring the optical flow between the real albedo images A^k and the synthetic images $\overline{A}^k$ rendered by our final model.

We compute the optical flow between A_j^k and $\overline{A}_j^k$ using Classic+NL [116] with default settings. This performs well when differences in lighting between the images are homogeneous.

To better understand how we use the computed flow in our evaluation, it is useful to recall the notation introduced in Section 3.4. Given a scan point x, we denote by n_x the surface normal at x and by $\zeta_x(C_j^k)$ the direction from x to the center of camera j; furthermore, $\pi_j^k(x)$ represents the projection of x onto the image plane of camera j.

For any vertex v_h^k that is sufficiently visible (i.e. $n_{v_h^k} \cdot \zeta_{v_h^k}(C_j^k) > t_{vis}$, where $t_{vis} = 0.7$), we evaluate the flow magnitude at the image pixel $\pi_j^k(v_h^k)$. We set a threshold t_{flow} to 1 pixel; vertices mapped to pixels with flow magnitude higher than t_{flow} in at least one image are considered unmatched. In the 612×512 images we consider, this threshold corresponds to at most 2mm on the scan surface.

The two tests ensure that the accuracy of alignments is within 2mm. This excludes 20% of all scan vertices; note that test 1 alone excludes 10% (see Section 4.4 for a detailed discussion about these results).

Inter-subject, sparse ground-truth correspondences are obtained from landmarks manually drawn on the subjects' skin (see Section 4.2.2). We easily detect the position of each landmark in camera images, and backproject identified 2D points to *scan* surface points. For completeness, we also evaluated the accuracy of these landmark correspondences on our alignments. The average error for the inter-subject correspondences defined by our alignments, computed over all the landmarks, is 3mm.

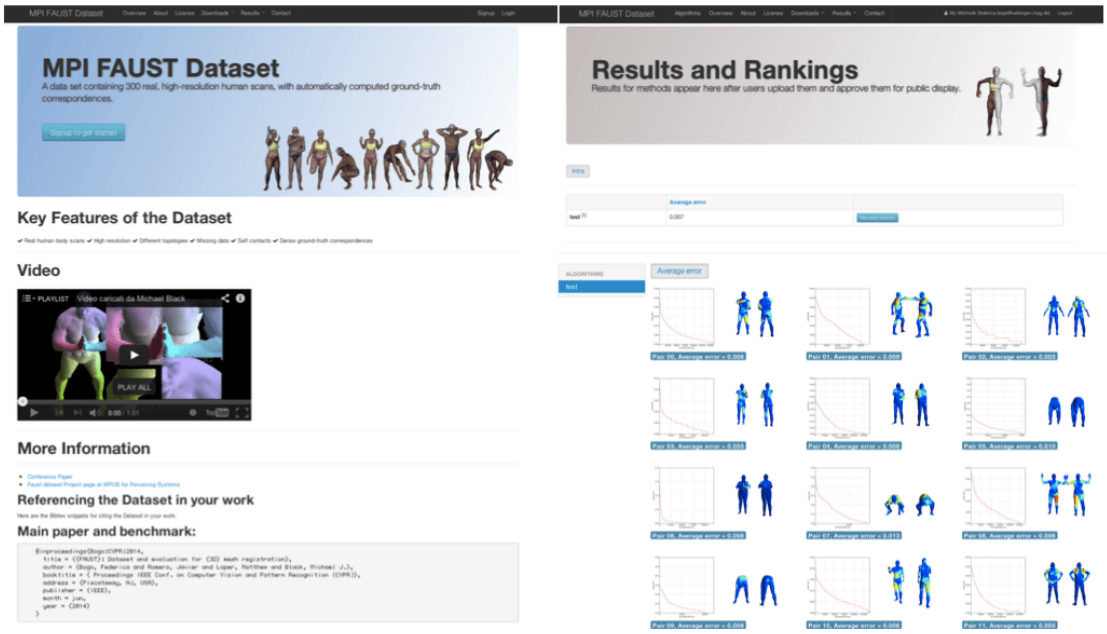


Figure 25: The FAUST website [10] allows users to download training and test sets, to submit their entries and visualize their results.

4.2.4 Benchmark

Based on the ground-truth correspondences defined by our alignments, we define a benchmark that evaluates 3D surface matching algorithms on real scans.

We split FAUST into a training and a test sets. The training set includes 100 scans (10 per subject) with their corresponding alignments; the test set includes 200 scans. The partition between training and test sets was chosen uniformly at random.

The FAUST benchmark defines 100 preselected scan pairs, partitioned into two classes – 60 requiring intra-subject matching, 40 requiring inter-subject matching. For each scan pair, $(S^k, S^{k'})$, we require a 3D point on the surface of $S^{k'}$ for every vertex on S^k. If the matching point is not a surface point of $S^{k'}$, we compute the closest point on the surface and use this.

To compute the error, we consider the Euclidean distance between the estimated point and the ground truth. Benchmarking is performed on each class (inter and intra) separately; for each class, we compute the average error over all the correspondences and the maximal error.

The dataset and the evaluation website are freely accessible for research purposes [10]. The website provides information about data and file formats, and lets users submit entries and visualize their results. Figure 25 shows two example webpages.

4.3 EXPERIMENTAL EVALUATION

We benchmark different state-of-the-art registration techniques on FAUST, partitioning them into model-free and model-based. Our main goal is to evaluate the impact of using real data when testing the performance of surface matching algorithms.

4.3.1 Model-free registration

We tested three embedding techniques, focusing on methods with publicly available code: Generalized Multi-Dimensional Scaling (GMDS) [37], Möbius voting [87] and Blended Intrinsic Maps (BIM) [78]. The first method achieves good results on TOSCA, while the last two perform well on both TOSCA and SCAPE.

The three algorithms require watertight meshes as input. Technically none of these methods can be evaluated on the FAUST benchmark, but to get a sense of how FAUST compares in complexity to TOSCA and SCAPE we converted our original scans to watertight meshes via Poisson reconstruction [77], keeping them at a fairly high resolution.

The algorithms return as output a set of sparse (GMDS and Möbius voting) or dense (BIM) correspondences. We computed the Euclidean distance between the returned correspondences and the ground truth; to compare our results with those reported in [78], we computed also a normalized sum of geodesic distances. We used only the intra-subject test set; the inter-subject test was not used because it requires correspondences of specific points on the scan, which are not provided by the sparse algorithms. Möbius voting and BIM did not return any result for 6 and 15 pairs of scans, respectively. While this does not comply with our benchmark, we report errors for successful pairs. We could not run GMDS at all because the method does not handle meshes with more than 4000 vertices; we chose not to downsample our meshes so drastically, since we are interested in evaluating techniques on higher-resolution data.

Benchmark results are summarized in Table 1. Möbius voting and BIM achieved an average error of 283mm and 120mm, respectively; the maximum errors were 1770mm and 1698mm. For geodesic error, Möbius voting and BIM had error lower than 0.05 units for 38% and of 64% of the correspondences, respectively. For a rough comparison, on 71 mesh pairs from SCAPE, [78] reports the same error threshold for 45% and 70% of the correspondences; on 80 mesh pairs from TOSCA, the same error is reported for 60% and 85% of the correspondences. We noticed that the algorithms tend to return many correspondences with high errors on challenging pairs.

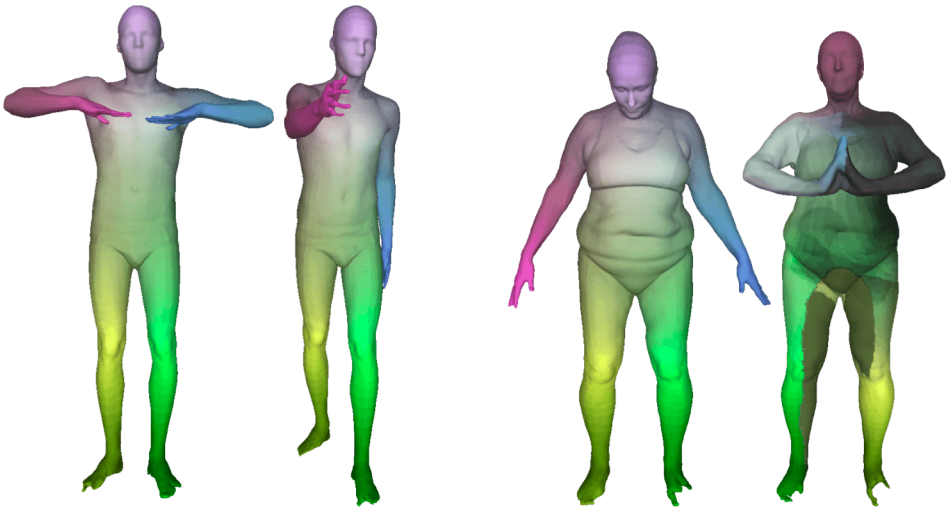


Figure 26: BIM algorithm [78] evaluated on two pairs of FAUST meshes made watertight. Correspondences are rendered with identical color. BIM handles pose variation (left pair), but fails to match meshes in the presence of self contacts (right pair).

In particular, we identify four principal challenges: missing data, differing mesh topologies between scans, high resolution, and self contacts. The algorithms returned correspondences with high error even for similar poses when meshes have missing parts (e.g. truncated hands or feet) or self contacts. Pose variance had in general minor (although not negligible) impact (see Fig. 26).

FAUST highlights the deficiencies of current algorithms when dealing with real scan data. This should drive the field in a useful direction.

4.3.2 Model-based registration

We are aware of no publicly available code for model-based registration, so we evaluated our method after removing its appearance-based component; this corresponds to the geometry-based coregistration approach described in Section 3.3. In this case, to obtain a fair evaluation, we used no landmarks during the alignment process.

On the full FAUST test set, the intra-subject error averaged 7mm; the maximal error was 926mm. We observed a very large error over a limited set of vertices; in most cases, this was due to self penetrations. When matching different subjects, the average error was 11mm, while the maximal error was 74mm.

These numbers can be seen as a rough estimate of the error one can expect due to sliding of surface points during mesh registration.

Table 1: Results obtained by GMDS [37], Möbius voting [87], BIM [78] and geometry-based coregistration [34] on the FAUST benchmark. All the errors are in millimeters.

Technique	Intra-subject error (avg/max)	Inter-subject error (avg/max)
GMDS[1] [37]	-	-
Möbius voting[2] [87]	283 / 1770	-
BIM[3] [78]	120 / 1698	-
Coregistration [34]	7 / 926	11 / 74

1. Not tested since it requires lower-resolution meshes.
2. Returns sparse correspondences. Tested over 54 scan pairs.
3. Returns dense correspondences. Tested over 45 scan pairs.

4.4 DISCUSSION

In this section, we discuss in greater detail the quantitative results presented in the chapter. First, we try to identify the reasons why 20% of all FAUST scan vertices are deemed misaligned by our validation process. Through this analysis, we identify the current limitations of our registration technique. Second, we apply the same validation process to judge the quality of the geometry-based registrations obtained after the first stage of our approach (see Section 3.3); in this way we can assess, from a quantitative point of view, the impact of using appearance information during registration. Finally, we conclude with some observations about the role and the significance of FAUST with respect to previous datasets for 3D mesh registration.

4.4.1 Unreliable correspondences

Our evaluation (see Section 4.2.3) defines as "reliably" registered (i.e. with an accuracy within 2mm) 80% of the vertices of all the scans. In particular, the percentage of reliably registered vertices *per scan* ranges from a minimum of 45% to a maximum of 93%.

Analyzing these results, we identify two main limitations of our approach: inadequate modeling of skin stretching, and scarce robustness against noise and missing data in small, detailed areas like hands.

Capturing skin stretching is problematic particularly in old subjects. Figure 27 exemplifies the stretching problem occurring for an old woman's knee: our appearance model erroneously captures skin wrinkles, that are present in most poses, generating

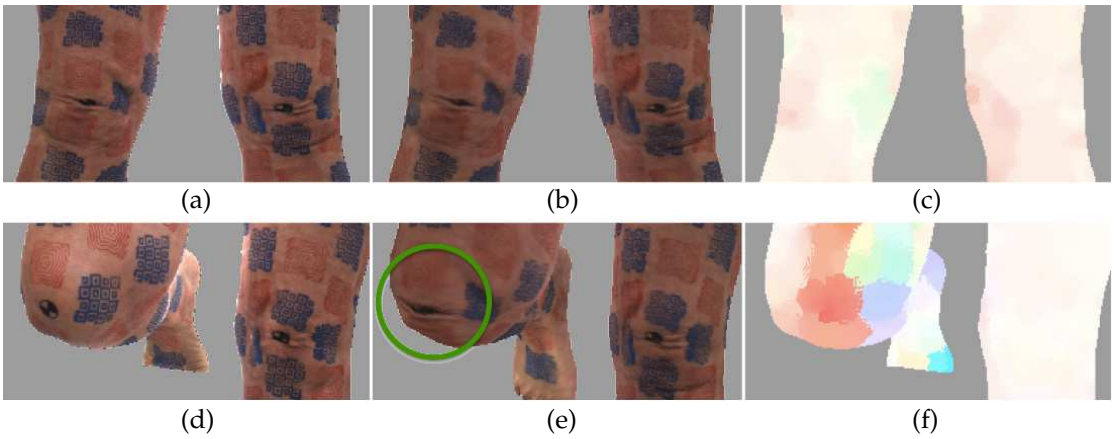


Figure 27: Problems due to skin stretching when registering the knee of an old woman in two different poses. For each pose, we show real image ((a),(d)), synthetic image ((b),(e)) and optical flow between the two ((c),(f)). Wrinkles are erroneously captured by our appearance model, resulting in wrong synthetic image patches when the knee is bent (green circle) – and therefore in a greater flow magnitude. The absence of toes in image (d) is due to missing data in the original scan.

inadequate synthetic images when the knee is bent. Our approach is currently not able to model these high-frequency details.

Besides creating inaccurate apperance models, this limits the applicability of our approach: wrinkles and small-scale deformations due to facial expressions, or clothing folds, would not be captured by our models as well. We plan to extend our approach to cope with these fine-scale details including a shape-from-shading component (similar to [122, 138]). Currently, we estimate scene lighting and surface albedo in a preprocessing stage. An alternative formulation of our appearance-based objective function may include lighting and albedo as parameters to optimize, leading to more accurate albedo estimates. Ideally, instead of using contrast normalization to obtain robustness against shading artifacts, our technique could exploit shading cues to reconstruct high-frequency geometric details.

Hands turn out to be very problematic areas, with nearly 100% of the scans having at least one hand vertex not reliably registered (see also Fig. 30 in the next section). Original scans are particularly noisy in these areas, due to the difficulty of capturing low-level details like fingers in scans covering the full body of a subject: in many cases fingers are missing or, when there is a contact between hands and other body parts, they cannot be clearly distinguished. Figure 28 shows some significant failure cases when registering hands.

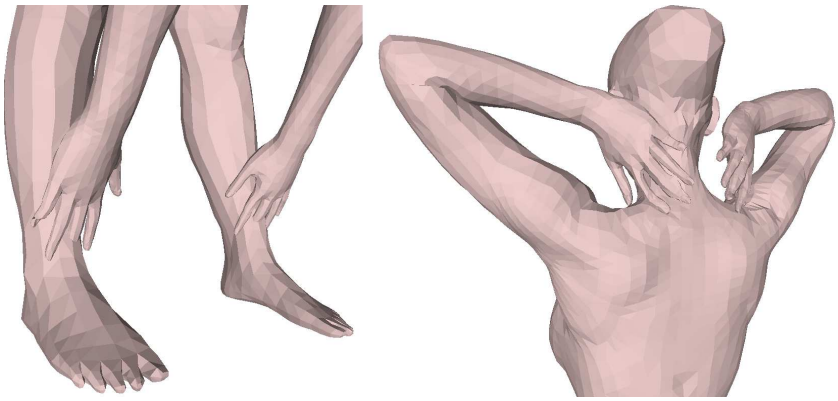


Figure 28: Failure cases when registering hands: self penetration (left) and unrealistic finger shape (right).

4.4.2 Appearance vs. geometry information

By applying the same validation process presented in Section 4.2.3, we estimated the number of "reliably" registered vertices obtained by our method after removing its appearance-based component.

Only 37% of the vertices of all the scans are registered within an accuracy of 2mm when scan textures ares not exploited. The percentage of reliably registered vertices *per scan* ranges from a minimum of 8% to a maximum of 67%.

We find that, in general, sliding plays an important role. We evaluated the percentage of vertices failing test 1 alone, and compared this with the percentage of vertices failing both tests. When the appearance-based component is considered during alignment, test 1 excludes 10% of all the scan vertices; test 1 and 2 together exclude 20% of the vertices. When evaluating geometry-based alignments, test 1 excludes only 4% of the vertices, but at the cost of a high sliding error: the final percentage of unreliably registered vertices is 63%. Figure 29 shows the sliding error detected by optical flow in one camera image when evaluating the geometry-based approach, and compares it with the results obtained after appearance-based optimization.

Figure 30 shows the distribution along the body surface of vertices failing test 1 alone (first row) and failing both tests (second row), for all the scans in FAUST; the left and right columns compare the results obtained when considering appearance during alignment or ignoring it, respectively. Red represents vertices deemed misaligned in all the scans; blue represents vertices accurately registered in all the scans. As expected, the number of vertices failing test 1 decreases when appearance is not considered. Sliding is still detectable even after appearance-based alignment, mostly in areas like armpits, forearms and shoulders; in some cases, we noticed errors due to movement

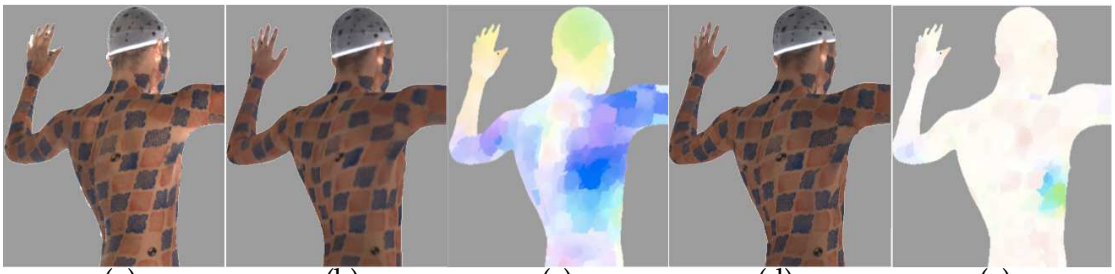


Figure 29: Sliding errors detected by optical flow significantly decrease after appearance-based optimization. We show an example real image (a), two synthetic images generated before (b) and after (d) appearance-based registration, and the corresponding computed flow, (c) and (e). Misalignments due to stretching, well visible in (c), are almost absent in (e).

of clothing between one scan capture and another. Without the appearance-based component, sliding is high in all the areas providing low-frequency shape information.

4.4.3 About the dataset

FAUST was officially released in June 2014. After half a year, we collected some statistics about its usage: more than a hundred people signed up to download the data; a dozen users submitted benchmark entries privately (i.e. without making results public); nobody submitted entries publicly. While FAUST alignments are used in statistical body shape modeling, the benchmark itself is less widely used.

We can draw on these numbers, and on the feedback we received by the users, to discuss the position of our dataset with respect to previous work. FAUST seems quite different from previous datasets for 3D mesh registration, for two reasons: first, the dataset is far more challenging; second, some choices we made, like the benchmark evaluation metric, are somewhat "unusual" in the field. In the following, we try to analyze both aspects in more detail.

As pointed out in Section 4.3.1, registration techniques have difficulties when dealing with FAUST scans for the following reasons:

- high resolution: FAUST scans consist on average of 172000 vertices – more than three times the average vertex count in TOSCA, the most popular dataset in the field. Some techniques we tested just failed in handling meshes with this resolution;

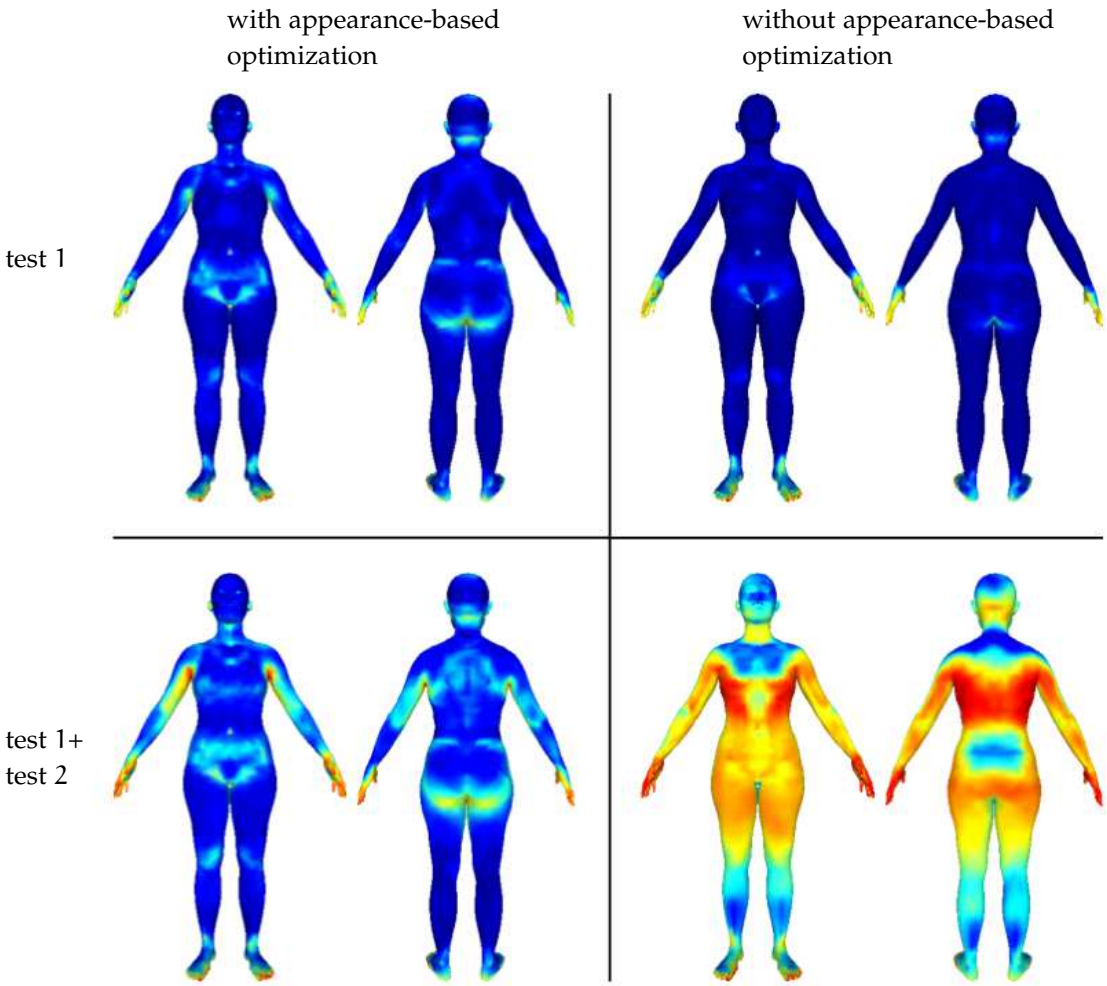


Figure 30: Distribution along the body surface of vertices failing test 1 alone (first row) and failing both tests (second row), for all the scans in FAUST; left and right columns compare the results obtained when considering appearance during alignment or not, respectively. Red represents vertices deemed misaligned in all the scans; blue represents vertices accurately registered in all the scans.

- differing mesh topologies between scans: too many techniques still rely on the (unrealistic) assumption that the meshes to be matched have identical number of vertices and topology;

- missing data and self contacts: this particularly affects model-free techniques, that seek intrinsic representations for the surfaces to be matched.

As for our "implementation" choices, we identify two main aspects differentiating FAUST from previous work:

- the computation of errors using Euclidean instead of geodesic distances: this choice is due to the fact that computing geodesic distances on non-watertight meshes is problematic;
- the computation of point-to-surface correspondences, instead of point-to-point (i.e. vertex-to-vertex) correspondences.

These characteristics may make more difficult the adoption of the FAUST benchmark in the community. We are currently considering the addition of more evaluation metrics, and the release of more training data. Our hope is that FAUST can encourage the development of more robust, real-world oriented registration techniques.

5 DETECTION OF NEW OR EVOLVING MELANOCYTIC LESIONS

In previous chapters, we assumed the presence of a high-frequency dense texture pattern painted on the skin of the subjects. This pattern was necessary to establish dense ground-truth correspondences between scans, and we leveraged it to assess the accuracy of our registrations.

A natural question that arises at this point is: can one exploit texture information given by the human skin alone – for example, by the presence of small artifacts like birthmarks or moles – to establish accurate scan-to-scan correspondences? To answer this, we explore possible uses of our approach in dermatology.

The problem we target is the automated screening of melanocytic lesions. Detecting lesion changes over time is crucial for a prompt diagnosis of melanoma, an aggressive form of skin cancer. By adapting our technique and integrating it with a lesion segmentation algorithm, we develop a fully automated pre-screening system for detecting new lesions or changes in existing ones, as small as 2–3mm, over almost the entire surface of the body. The system provides a first level of surveillance; putative changes can then be evaluated by a dermatologist.

The integration of lesion segmentation with a 3D body model is a key novelty that makes the approach robust to illumination variation, changes in subject pose and shape, and presence of sparse body hair.

We provide some background on melanocytic lesion screening in Section 5.1. After providing a brief overview of the most relevant literature (Section 5.2), we describe our system in Section 5.3. In Section 5.4, we present experimental results obtained in a pilot study using synthetic lesions. In Section 5.5, we draw some conclusions and look at directions of future work.

5.1 MELANOCYTIC LESION SCREENING

Malignant melanoma is an aggressive form of skin cancer whose incidence is rapidly increasing worldwide [53]. In its early phases, a melanoma is almost indistinguishable from a benign melanocytic lesion (a common "mole"). A sensitive sign of a malignant

lesion is its *evolution*: the appearance of a new lesion or changes in an existing one suggest an increased probability of a melanoma, while stability speaks against the presence of a disease. Early detection promptly followed by excision is the key to a favorable prognosis [53, 106]; this requires periodic monitoring of the skin and in particular of melanocytic lesions (sometimes hundreds on an individual).

Monitoring typically takes place at three levels of detail. First, the dermatologist conducts a naked eye, "full-body" inspection of the skin, detecting suspect lesions. Then, these lesions are examined through a dermatoscope, which provides 10× to 60× magnification and specific illumination revealing more details of the lesion's 3D structure. Finally, lesions deemed at risk are excised and, a posteriori, histopathologically analyzed.

Digital images are often acquired at both the first ("full-body") and at the second ("dermatoscopic") level of detail, allowing the dermatologist to track the evolution of the patient's melanocytic lesions. The periodic acquisition of these images at both levels of detail is an extremely time-consuming task; it is rarely performed more than once per year on a given patient. Furthermore, manual comparison – especially of "full-body" images – is challenging and error-prone: there might be changes in the pose (and in the shape) of the patient, and in the illumination of the scene.

The introduction of automated tools can help reduce the time spent by the dermatologist during a visit, costs and errors [51, 76]. In the next section, we review the literature on automated melanocytic lesion screening, with a focus on the first ("full-body") level.

5.2 PREVIOUS WORK

Most previous work on lesion change detection addresses the problem at the "dermatoscopic" level, analyzing high-magnification images of small regions surrounding a lesion (see [76] for a survey). Tracking multiple lesions, however, is a challenging problem that has received surprisingly little attention [76].

The task can be subdivided into two parts: segmentation/detection and registration/matching. Techniques proposed in the literature usually do not undertake both tasks simultaneously, but rather focus either on one or on the other.

Segmentation/detection approaches usually identify a set of lesion candidates using simple image processing methods [56, 101, 103, 118, 131] and then filter the results using unsupervised [101, 103] or supervised [118] classification.

Matching lesions in images taken at different times is challenging and approaches take many forms. The diameter of a lesion is generally small compared to its displacement between different images, making matching hard. One approach solves for a rigid 2D transformation between images given user-provided matches [102]. In [91], back torso images are mapped to a common 2D template. Pose variation and non-rigid changes in body shape cause non-linear, anisotropic deformations of the skin, further complicating matching. Other approaches focus on the topological relations between lesions and use graph-matching methods to find the relationship between images [70, 91, 92, 130]. While able to produce robust matchings, these approaches may have difficulty with large numbers of lesions.

Voigt and Classen [131] perform both segmentation and registration. Images of the patient's front and back torso are acquired with a single camera and a positioning framework for adjusting the patient's pose. Lesion borders are detected by thresholding the output of a Sobel operation; due to the large number of skin features easily mistaken for lesions, such as hair, this can lead to poor performance. The precise positioning of the patient attempts to remove the registration problem but, in practice, humans are deformable and there will always be non-rigid changes in pose and shape.

Virtually all previous segmentation and registration techniques are evaluated only on a small part of the body, commonly the back or front torso. These methods do not provide a solution to the full-body analysis/screening problem. In contrast, we consider the entire body surface (or most of it) at once, simplifying the acquisition process for both patient and dermatologist.

5.3 METHOD

To build our system, we adapt the registration technique presented in Chapter 3 and integrate it with a novel lesion segmentation algorithm.

The approach proceeds in five steps (see Fig. 31). First, we capture a scan of the subject, obtaining a triangulated non-watertight mesh and a set of camera images. Given scans of a subject taken on different days there will be changes in lighting, pose, shape, hair and skin texture. We preprocess the original images to remove shadows, and perform a preliminary lesion segmentation in image space. Then, we bring the scans into registration with our 3D body model. This defines dense correspondences between scan surfaces, normalizing for variations in subject pose and shape. Once scans have been registered, we refine our segmentation and detect lesion changes in the parameterized space defined by the model.

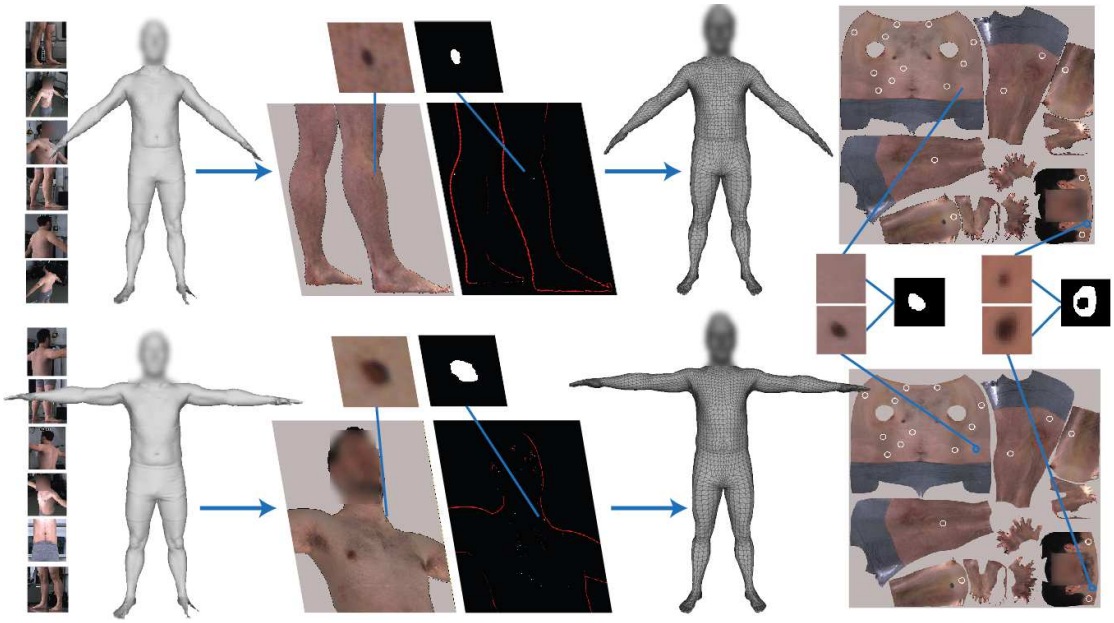


Figure 31: Overview of our approach. After scan capture, we preprocess camera images in order to eliminate shadows and obtain a preliminary lesion segmentation. We bring scans into registration with a 3D body model, that normalizes for pose and shape variations. Once scans are registered, we compare them and identify changes in the parameterized space defined by the model.

The following sections describe each step in detail.

5.3.1 Scan acquisition

We use the 3D multi-stereo system described in Section 3.2. Each scan is a triangulated, high-resolution, non-watertight mesh. Synchronized with each scan are 22 color cameras, capturing images of the body from different views. Differently from the approach described in Section 3.2, we now process images at a higher resolution (1224×1024 pixels).

5.3.2 Lesion segmentation

Automated lesion segmentation on the whole body may suffer from the presence of shadow artifacts. To reduce these effects, we preprocess the original camera images $I = \{I_j : j = 1, \ldots, 22\}$, in order to discriminate between albedo and shading.

We assume lighting is constant across scan sessions, and simply precompute it as described in Section 3.4.1. The estimated lighting model is used for computing shadows, which are then removed from each original image I_j to produce the corresponding albedo image A_j (see Fig. 31).

After obtaining a set of albedo images $A = \{A_j : j, \ldots, 22\}$, we segment them to identify putative lesion borders.

As a preliminary step, in each image we isolate skin from background and clothing by means of a simple thresholding of the hue. We choose a conservative threshold, since subsequent steps can deal with skin false positives.

We obtain an initial estimation of lesion borders using Laplacian-of-Gaussian (LoG) filtering [103, 118]. Since lesion radii can vary depending on the subject and on the camera viewpoint, the LoG filter is applied at five different scales. Linear Discriminant Analysis (LDA) [66] is used to classify each pixel in A_j into a lesion binary mask H_j based on the output of the multi-scale LoG filter. LDA classification produces, for each albedo image A_j, a binary mask H_j, marking each pixel as lesional or non-lesional. We collect these masks in a set $H = \{H_j : j = 1, \ldots, 22\}$.

Facial features and occlusion boundaries, due to their high second-derivative response, may be erroneously identified as lesional. However, these artifacts tend to be elongated, while lesions are spatially compact. We postprocess each mask H_j in order to keep only compact connected components. For each connected component in H_j, we consider its minimum bounding box; if the ratio between its major and minor side is too high, or fewer than half of the pixels inside it are lesional, the component is discarded (such components are colored in red in the preliminary masks in Fig. 31).

5.3.3 Scan registration

We register scans of the same subject captured in different sessions by aligning a common template T^* to each scan. As in Chapter 3, we minimize an objective function considering both geometry and appearance information, and couple the template towards a BlendSCAPE body model during the alignment. However, in contrast with the approach previously described, here we do not coregister a corpus of scans, but rather register scans relative to a single subject. We therefore have to define how the parameters of our body model are learned. After providing a few details about model learning, we formalize our objective function.

MODEL LEARNING As described in Section 3.3.1, our BlendSCAPE model factorizes human body deformations into a set of shape-dependent deformations, D, and a set

of pose-dependent deformations, $B(\theta, W)$ and $Q(\theta)$. Both $B(\theta, W)$ and $Q(\theta)$ are parameterized by the pose vector θ; W collects a set of blending weights. We also model subject appearance by means of a texture map U.

In our screening system, we are interested in obtaining a personalized model of shape deformations and an appearance model of the subject, but do not define an objective function including Q and W as parameters to optimize. Instead, we learn Q and W from a corpus of 1832 scans of 78 people (41 women and 37 men) in a wide range of poses, registered with the approach presented in Section 3.3.

To learn a subject-specific shape model D, we capture a set of initial scans $\{S^k : k = 1, \ldots, N_{init}\}$ of the subject and coregister them based on geometry information.

More precisely, we obtain a set of alignments $\{T^k : k = 1, \ldots, N_{init}\}$, a set of pose parameters $\{\theta^k : k = 1, \ldots, N_{init}\}$ and a set of shape-dependent deformations D by minimizing the following error function E_{init}:

$$E_{init}(\{T^k\}, \{\theta^k\}, D; \{S^k\}, Q, W) = \sum_{\text{scan } k} \lambda_S E_S(T^k; S^k) + \sum_{\text{scan } k} (\lambda_{cpl} E_{cpl}(T^k, \theta^k, D; Q, W) + \lambda_\theta E_\theta(\theta)) \tag{5.1}$$

where E_S, E_{cpl} and E_θ correspond to the geometric data term, the coupling term and the pose prior introduced in Section 3.3.2 (Eqs. (3.2), (3.6) and (3.3)), and λ_S, λ_{cpl} and λ_θ are weighting terms. Note that a very limited number of scans is sufficient to learn an adequate shape model; in our experiments, we set $N_{init} = 2$ (see also Section 5.4).

After obtaining a set of initial geometry-based alignments, we compute an appearance model U using the same methodology described in Section 3.4.2.

Note that D and U need to be computed only once per subject, during the first acquisition session. In the subsequent sessions, we use D and U during scan registration, but do not optimize them.

REGISTRATION Our formulation is similar to that provided in Section 3.5. The quality of the correspondence between the scan S captured during a generic session and the corresponding deformed template T is measured in terms of an error with 4 components: a geometry-based data term E_S, a "coupling" term E_{cpl}, a pose prior E_θ and a slightly modified appearance-based data term E_{U_H}. E_S, E_{cpl} and E_θ are analogous to the terms defined in Eqs. (3.2), (3.3) and (3.6).

Differently from the original formulation introduced in Section 3.5, the appearance-based error term now includes a weighting factor to give more importance to appearance consistency around lesions. Given U, T and the calibration parameters C_j of camera j, we render a synthetic image $\overline{A}_j$ (Fig. 32(b)). The appearance-based error

term E_{U_H} encourages consistency between each albedo image A_j and the corresponding synthetic image $\overline{A}_j$:

$$E_{U_H}(T; A, H, C, U) = \sum_{\text{camera } j} \sum_{\text{pixel } \mathbf{y}} w_{H_j}(\mathbf{y}) \| G_{\sigma_1,\sigma_2}(A_j)[\mathbf{y}] - G_{\sigma_1,\sigma_2}(\overline{A}_j)[\mathbf{y}] \|^2 \tag{5.2}$$

where C is the set of camera calibration parameters $C = \{C_j : j = 1, \ldots, 22\}$, and $w_{H_j}(\mathbf{y})$ is a weighting function assigning higher weight to pixel $\mathbf{y}$ if $\mathbf{y}$ is marked as lesional in H_j. As in the formulation provided in Section 3.5, G_{σ_1,σ_2} defines a Ratio of Gaussians (RoG) of parameters σ_1 and σ_2.

Summarizing, we deform the template T to fit the scan S minimizing the following objective function:

$$\begin{aligned} E_{algn}(T, \theta; S, A, H, C, U, D, Q, W) = & \lambda_S E_S(T; S) + \\ & \lambda_{U_H} E_{U_H}(T; A, H, C, U) + \\ & \lambda_{cpl} E_{cpl}(T, \theta; D, Q, W) + \lambda_\theta E_\theta(\theta) \end{aligned} \tag{5.3}$$

where E_S, E_{cpl} and E_θ are analogous to the terms defined in Eqs. (3.2), (3.3) and (3.6); λ_S, λ_{U_H}, λ_{cpl} and λ_θ are weights assigned to the different terms.

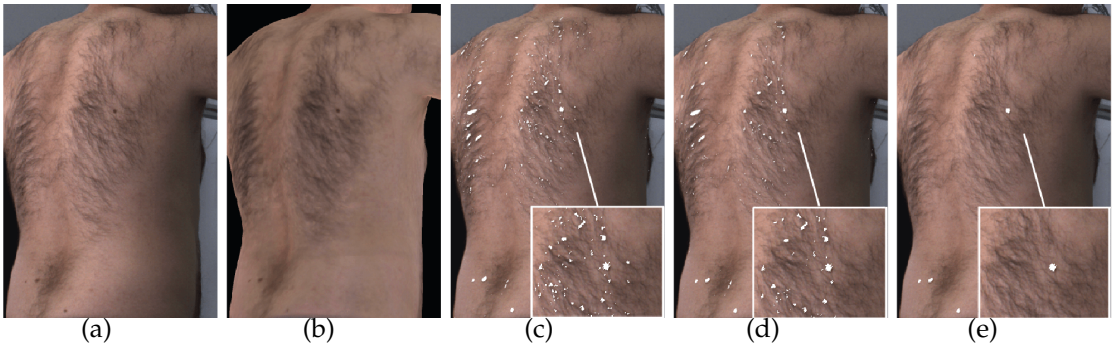
5.3.4 Lesion segmentation refinement and change detection

The presence of sparse hair, small skin artifacts or generic image noise may affect the performance of the pre-segmentation described above, producing a high number of false positives. Using more restrictive classification thresholds or artifact removal algorithms (as in [118]) may produce false negatives, i.e. discard actual lesions. Crucially, these artifacts tend to be mistaken as lesions only from specific viewpoints.

We exploit our multi-camera capture framework to filter out lesions that are not consistently detected by a number of relevant (i.e. with a good viewpoint) cameras. While previously we were working in image space, now we perform our segmentation refinement in texture space.

More formally, for any template surface point $\mathbf{x}$, we denote by $uv(\mathbf{x})$ its mapping from 3D to texture space, and by $\pi_j(\mathbf{x})$ its projection onto the image plane defined by camera j. $H_j[\pi_j(\mathbf{x})]$ equals 1 if $\mathbf{x}$ is classified as lesional according to camera j, 0 otherwise. Furthermore, we denote by $\mathbf{n}_x$ the surface normal at $\mathbf{x}$, and by $\zeta_x(C_j)$ the direction from $\mathbf{x}$ to the center of camera j. Texel $\mathbf{y} = uv(\mathbf{x}_y)$ is classified as lesional if and only if

$$\frac{\sum_{\text{camera } j} V_{x_y}(C_j) H_j[\pi_j(\mathbf{x}_y)] \max(\zeta_{x_y}(C_j) \cdot \mathbf{n}_{x_y}, 0)}{\sum_{\text{camera } j} V_{x_y}(C_j) \max(\zeta_{x_y}(C_j) \cdot \mathbf{n}_{x_y}, 0)} > \delta \tag{5.4}$$



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

Figure 32: A real (a) and a synthetic albedo (b) image of a subject's back. Figures (c)-(e) show the final segmentation obtained by setting δ in Eq. (5.4) to 0, 0.2 and 0.5, respectively.

where $V_{x_y}(C_j)$ is a visibility function returning 1 if $\mathbf{x}_y$ is visible from camera j, 0 otherwise, and δ is a system parameter. This corresponds to computing a weighted average of the classifications provided by different cameras – where the contribution of each camera is weighted according to the quality of its viewpoint. Figure 32 shows how the final segmentation varies depending on δ: artifacts like sparse hair tend not to be consistently detected across different cameras, and are therefore filtered out; lesions exhibit more consistency (see e.g. the bottom of the back and the right shoulder). We quantitatively evaluate the sensitivity of the system to the value of δ in Section 5.4.

Detected lesions are integrated into a full-body texture map (see Fig. 31). This greatly simplifies the tracking of changes in lesions compared to using multiple single images. Each texel is associated with the same template surface point, independently of subject pose and shape. Texture maps from different times are therefore directly comparable. A detection that does not overlap with one in a previous map reveals a new lesion; a detection that does overlap, but comprises a higher number of pixels, is likely to reveal a lesion that has grown.

5.4 EVALUATION

We evaluated our system on a set of 6 male and 6 female subjects of ages 23 to 44 years, height 160 to 186 cm, and weight 55 to 82 kg. There was considerable variation in terms of skin tone, number of melanocytic lesions, and presence of sparse body hair (Fig. 33(b)). We trained the LDA classifier (Section 5.3.2) on a set of 50 images of 10 different subjects, captured with our 3D multi-stereo system; there is no overlap between the subjects used for evaluation and those used for training.

(a) (b)

Figure 33: (a) Skin patch exhibiting a synthetic lesion (large lesion towards upper left). (b) Scans of subjects, showing varied skin phenotype and pose.

For this pilot study, we artificially created and altered lesions by drawing with a marker on the subjects' skin. Note that these synthetic lesions look realistic at the resolution of our images, as seen in Fig. 33(a).

Each subject was scanned in 2 poses, respectively with arms held horizontally ("T-pose"), and pointing downwards at an angle ("A-pose") (Fig. 33(b)).

For each subject, we captured two initial scans in order to learn D and U (as described in Section 5.3.3). After the initial scans, for each subject we created 10 synthetic lesions with a diameter of 3mm, and re-scanned them in the 2 poses. We then expanded each synthetic lesion first to a diameter of 5mm, and then to a diameter of 7mm, re-scanning the subjects in the 2 poses each time. This yielded 4 timepoints with increasing lesion diameter (0, 3, 5 and 7mm): 3 pairs of timepoints ($0 \rightarrow$ 3mm, $0 \rightarrow$ 5mm, $0 \rightarrow$ 7mm) correspond to the appearance of new lesions of different diameters, while the other 3 pairs ($3 \rightarrow$ 5mm, $3 \rightarrow$ 7mm, $5 \rightarrow$ 7mm) correspond to changes in existing lesions. For each pose, and pair of timepoints, our system identifies a set of "suspect" lesions – lesions deemed either new or modified.

For different values of the parameter δ (see Section 5.3.4), our system yields different values of precision (the fraction of suspect lesions that were actually new or modified lesions) and recall (the fraction of new or modified lesions that were reported as suspect lesions).

Figure 34 reports the results for the A-pose, since it was the most comfortable for all subjects. The results for the other pose are almost identical. Precision and recall values are computed by aggregating the values obtained for all the subjects. On average, a high recall ($> 90\%$) was achieved for all pairs of timepoints, with a precision $> 50\%$ in the case of small (3mm) new lesions, $> 80\%$ in the case of larger new lesions (5mm and 7mm), and $60 - 80\%$ in the case of changes in existing lesions. Note that, while high precision is desirable, high recall is more important since the consequences of missing a potential melanoma are much direr than those of a false alarm.

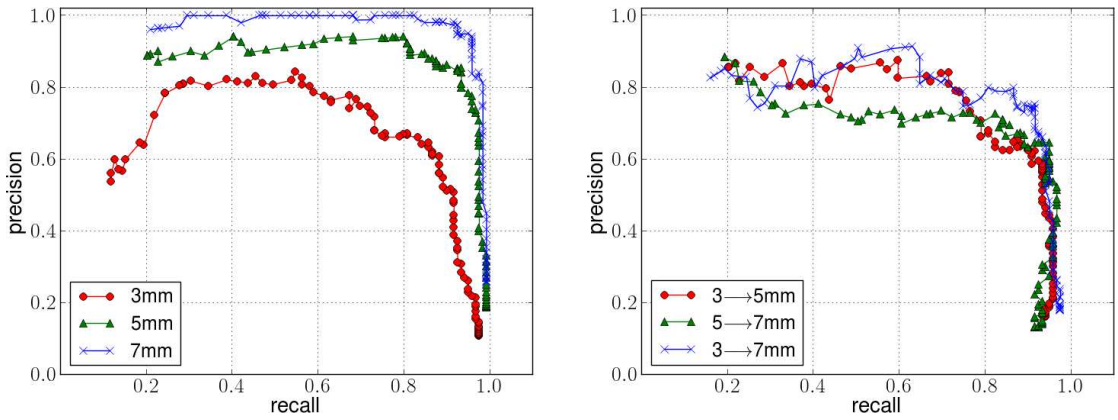


Figure 34: Precision/recall curves for detecting new lesions (left) and increased lesion sizes (right), for 100 values of the parameter δ evenly spaced between 0 and 1. Precision and recall values are computed by aggregating the values obtained for all the subjects.

The acquisition of each scan requires a few milliseconds. Further processing (scan generation, alignment, texture map analysis) can be performed off-line; in our experiments, it required a few minutes per scan on a common desktop machine.

5.5 CONCLUSIONS

In this chapter we have proposed a novel solution for "full-body" screening of melanocytic lesions. A multi-camera system captures the 3D shape and skin texture of a subject. Given two such scans of the same subject, taken at different times, we bring them into registration by aligning each scan with a learned, parametric 3D body model. Once scans are registered, we compare them across time and identify changes in skin lesions. In a pilot study, we show that our method automatically detects changes on the order of 2–3mm.

In comparison with previous work, our approach introduces a number of novelties. Most techniques in the literature try to detect lesion changes by comparing 2D images of a patient taken at different times. This comparison may be challenging due to illumination variation, and due to non-rigid changes in the patient's pose and shape. In contrast, we propose to compare 3D textured scans, after bringing them into registration with a 3D body model. Working in 3D makes it easier to estimate shadows, and remove them; the approach is robust to non-rigid deformations due to changes

in body pose and shape, since lesion changes are detected in the parameterized space defined by the model.

Based on our preliminary results, a longitudinal study of dermatological patients should be pursued. We would be interested also in evaluating the usefulness of our system for acquiring and analyzing data in epidemiological studies.

Future work should explore higher-resolution RGB imagery, and the effect of varying number/resolution of cameras on detection.

Currently, one major limitation of the system is its cost. A different, intriguing research line would explore less expensive scanning devices (e.g. the Kinect) for the acquisition of 3D data and texture. Here our 3D body model could be exploited to integrate information from multiple poses and, given accurate alignment, image super-resolution could be used to obtain high-quality texture. We tackle the problem of registering data acquired with a single Kinect camera in the next chapter.

6 TEXTURED 3D BODY MODELS FROM KINECT

In this chapter we tackle the problem of learning personalized body shape and appearance models from dynamic sequences captured with a single Kinect sensor.

Working with dynamic, monocular Kinect sequences poses novel challenges compared to the ones we faced in the previous chapters. It represents a testbed to evaluate the robustness of our approach, and highlight its shortcomings; and it is an opportunity to push forward our algorithms to deal with more complex scenarios. In contrast to previous chapters, here we present work that is not yet fully mature; we outline an approach and show some preliminary results – but, mostly, we look ahead to future research directions.

This chapter is organized in five sections. Section 6.1 introduces the problem we address, identifying the main challenges involved. Section 6.2 presents the most relevant literature. Section 6.3 describes our approach: we build on the algorithms presented in previous chapters, extending and adapting them to work with noisier, incomplete data. Section 6.4 presents preliminary results obtained on a set of 24 sequences, capturing 6 subjects in different motions. Section 6.5 discusses the significance of our results, analyzes current limitations of our approach and proposes further improvements.

6.1 PROBLEM AND CHALLENGES

During recent years there has been an enormous diffusion of consumer RGB-D cameras. Tens of millions of devices have already been shipped, and new versions for integration into laptops, tablets and smartphones are currently being developed [145].

A lot of research effort has been put into recovering high-fidelity geometric reconstructions from depth data captured with such devices. People can now scan static scenes or objects at home, and obtain shape models with an impressive level of detail.

But besides shape, appearance plays an important role in obtaining realistic models. Recent devices offer good-quality RGB streams, often at a higher resolution than depth ones. Crucially, RGB data at this resolution is useful not only to recover appearance: it is also a source of (dense or sparse) features to accurately track moving objects.

Leveraging this idea, we tackle the problem of building high-quality, personalized appearance and shape body models from monocular RGB-D sequences captured with a single Kinect. In contrast with most previous work, we try not to put any constraint on the subject's motion during capture. In fact, we consider large non-rigid deformations and viewpoint changes important to obtain complete, detailed models.

Working with Kinect monocular sequences poses a series of challenges. First, data provided by the device is noisier and at a lower resolution than scans obtained from professional scanning systems. Second, fast body motions may increase blurring and outliers. Third, the body is never fully visible at once, so one faces the problem of reconstructing a complete model from partial (per-frame) views.

Solutions proposed in the literature can be roughly divided into two classes. "Model-free" approaches consider multiple depth frames, corresponding to different views of the subject, and bring these frames into registration to obtain a complete model. To minimize the presence of non-rigid deformations, the subject is usually asked to keep a similar pose between frames (e.g. to turn around); this limits the amount of subject-specific shape and appearance details that can be captured. "Model-based" approaches fit a statistical body model to partial views, searching for optimal parameters into a low-dimensional shape space; often, the range of poses that can be handled is limited. The expressiveness of the model is bound to that of the low-dimensional space, so high-frequency shape details are smoothed out.

Our goal is to develop a different solution, that a) can handle significant non-rigid deformations and b) is not limited by a low-dimensional shape space. As in the previous chapters, we rely on the BlendSCAPE model described in Section 3.3.1. We bring our model into registration with depth and RGB data captured in each frame, obtaining a set of alignments (i.e. topologically coherent meshes deformed to fit data in each frame). These alignments are used to learn a personalized model of shape deformations, and an appearance model.

Note that, in our formulation, we are actually combining temporally coherent motion reconstruction and model learning: our approach produces a sequence of deformed meshes (one per frame), and use them to learn a set of body model parameters.

6.2 RELATED WORK

Our work tries to address two related, but different, problems: shape and appearance modeling from consumer RGB-D cameras, and motion and geometry reconstruction

from dynamic sequences. In the following, we briefly review the most significant approaches proposed in the literature for either problem.

SHAPE AND APPEARANCE FROM CONSUMER RGB-D CAMERAS Early approaches try to recover detailed geometry from rigid scenes. Systems such as KinectFusion [72, 95] enable a user holding and moving a Kinect camera to create highly detailed 3D reconstructions of indoor scenes in real time; data captured in different frames is registered using ICP to obtain a dense 3D model. Recent work [145] extends this approach, correcting for optical distortions introduced by the camera to recover high-quality appearance models.

Interestingly, a number of approaches to estimate human body shape draws inspiration from KinectFusion [47, 84, 111, 124, 142]. In these systems, the user is often asked to rotate in front of the device maintaining a roughly rigid pose; partial data captured from different viewpoints is merged to produce a single mesh, using non-rigid registration algorithms to correct for small deviations in the motion between viewpoints. These approaches treat non-rigid deformations as "noise", instead of explicitly capturing and modeling them.

Model-based techniques [44, 134] rely on low-dimensional statistical body models to handle a wider range of poses. These approaches fit a set of pose and shape parameters to one or multiple frames in order to recover complete models from partial data. Working in a low-dimensional shape space tends to limit the expressiveness of the reconstructed shape, smoothing out high-frequency geometry (e.g. subject-specific face details).

Note that the goal of all these systems is to produce a single mesh as output from an RGB-D sequence; they do not address non-rigid motion reconstruction.

To the best of our knowledge, no attempts have been made so far to learn an appearance body model, or even to recover high-quality texture maps from consumer RGB-D cameras. Common approaches reconstruct per-scan appearance by averaging RGB information from different views [47], and smoothing texture transitions between views via Poisson blending [84, 111, 124]; in [84], the authors correct for illumination changes between views using a general-purpose albedo estimation technique [25]. This does not avoid artifacts such as ghosting or blurring, leading to low-quality textures.

NON-RIGID GEOMETRY AND MOTION RECONSTRUCTION Many approaches address the problem in high-quality multi-camera systems. In such settings, impressive results have been shown in the reconstruction of complex human motion and dynamic geometry of non-rigid surfaces, including people with general clothing [42, 48, 49, 57, 85].

In some cases, fine geometric details are captured by estimating surface reflectance in controlled studio setups with sophisticated lighting [122, 129].

Recently, the availability of consumer depth cameras promoted the creation of more "lightweight" setups. Some approaches employ multiple devices: Ye et al. [140] capture multi-subject performances with three moving Kinects; Dou et al. [50] deform a human template to match scans acquired with an eight-Kinect rig.

Working with monocular sequences, however, is more challenging; at each frame, typically less than half of the considered object is visible [147], so the behavior of unobserved parts has to be inferred.

Most approaches focus on capturing and modeling small objects. Li et al. [82] reconstruct non-rigid surface deformations from high-resolution monocular depth scans, using a smooth template as geometric prior; the limited scanning volume makes the approach suitable only for body parts like hands or faces. Zollhöfer et al. [147] try to continuously reconstruct non-rigid motions of small objects or body parts, acquired with a custom RGB-D camera, fitting a rough template to each frame in real time. Successful model-based approaches have been proposed for real-time tracking of hands from Kinect [97, 121]. These approaches do not recover object appearance.

Less effort has been put in reconstructing the motion of full human bodies. Liao et al. [86] develop an algorithm to model complete 3D deformable objects (including human bodies) from monocular depth sequences, evaluating it on synthetic data. Helten et al. [67] estimate a personalized body shape model from two Kinect depth images, and then use it to track the subject's pose in real time from a stream of depth images. The system does not provide satisfactory tracking performance when the subject does not face the camera, or when parts of the body are occluded or outside the recording volume of the Kinect.

6.3 APPROACH

Our approach takes as input a sequence of frames, acquired with a single Kinect camera, capturing a subject in an arbitrary motion; it produces as output a subject-specific shape and appearance model, together with a set of alignments – i.e. a set of topologically coherent meshes deformed to fit depth and RGB data in each frame.

The task is challenging. At each frame, only part of the body is visible; data may suffer from noise; furthermore, prior information about the motion is scarce. We address this with a two-stage approach, summarized in Fig. 35. In the first stage, we rely on a low-dimensional shape space; namely, we represent shape-dependent deformations

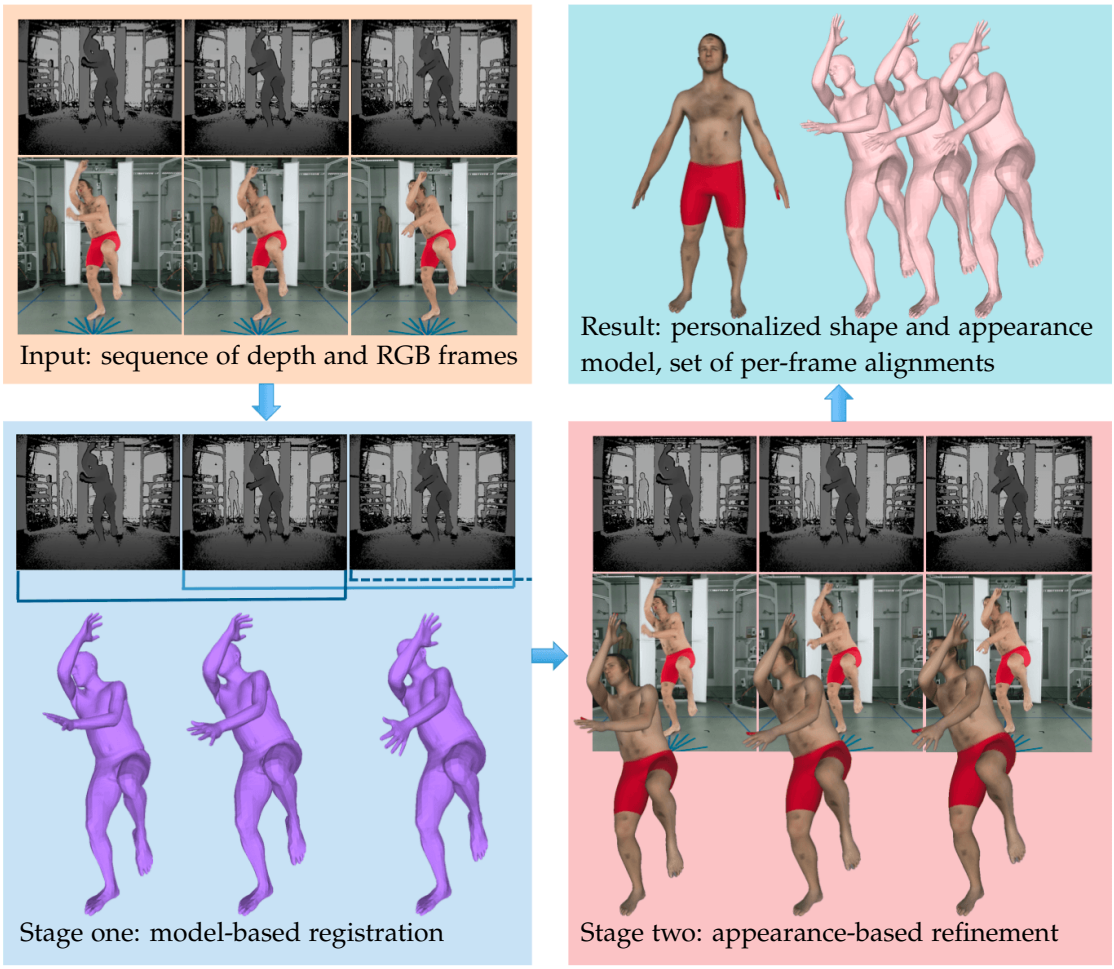


Figure 35: Our approach proceeds in two stages. First, we optimize a set of body model pose and shape parameters over short windows of consecutive depth frames. We use this estimate to initialize a second round of registration, in which we exploit both RGB and depth data. After optimization, we obtain a personalized shape model D and an appearance model U (here we show a model deformed according to D, and textured with U), plus a set of alignments, one per frame.

in our BlendSCAPE model as a linear combination of shape basis vectors, learned via Principal Component Analysis (PCA) from an aligned database of thousand bodies. We exploit the fact that body deformations are continuous and predictable over short time intervals, and jointly optimize pose parameters and shape coefficients over small windows of consecutive depth frames. We use these shape and pose estimates to initialize a second round of registration. In the second stage, we exploit both depth and

RGB data to obtain a set of refined alignments. Based on these alignments we learn a personalized model of shape deformations, which is not constrained by the initial low-dimensional space, and an appearance model.

Section 6.3.1 clarifies how we define a PCA shape space in BlendSCAPE; Sections 6.3.2 and 6.3.3 formalize the objective functions minimized during alignment and model learning, while Section 6.3.4 provides some details about their optimization.

6.3.1 Introducing a low-dimensional shape space

Before introducing the low-dimensional shape space adopted in our approach, it is useful to recall some details about BlendSCAPE. As described in Section 3.3.1, BlendSCAPE parameterizes the deformations that fit a template mesh T^* to a scan into a set of pose parameters θ and a set of shape parameters D. During alignment, T^* is first unstitched into disconnected triangles. Each triangle is fit according to a sequence of pose- and shape-dependent deformations: $B(\theta, W)$, D and $Q(\theta)$. $B(\theta, W)$ is a linear blend of rigid rotations of body parts: it is parameterized by a set of blending weights W and by a pose vector θ, collecting relative rotations between neighboring parts. D and $Q(\theta)$ account for deformations dependent on the subject's identity and on the pose, respectively. After deformation, the disconnected triangles are stitched into a watertight mesh, $\mathcal{M}(\theta, D, Q, W)$, by solving for vertex positions via least squares.

As already proposed in the original SCAPE formulation [20], given a corpus of registered scans we can learn a low-dimensional space of shape deformations. In Chapter 3, we explained how coregistration brings template T^* into alignment with a corpus of scans of N_{sbj} subjects, and simultaneously learns Q, W, and a set of shape-dependent deformations $\{D^p : p = 1, \ldots, N_{sbj}\}$. Once a corpus of scans has been coregistered, we can apply PCA to the set of deformations $\{D^p\}$ to compute a low-dimensional shape space. A shape in this space is fully described by a vector of coefficients, β; we denote the corresponding deformations by $D(\beta)$.

We compute a PCA shape space from 3803 coregistered scans (1700 men and 2103 women) from CAESAR [107]; note that we learn two separate spaces for men and women. From these spaces, we extract the first 300 principal components. Furthermore, we learn Q and W from a corpus of 1832 scans of 78 people (41 women and 37 men) in a wide range of poses (note that these values of Q and W coincide with those used in Section 5.3.3).

Since we use precomputed values for Q and W and do not optimize them when fitting Kinect data, for simplicity we omit these parameters in our notation. We introduce instead a novel parameter γ, representing the global translation center of model

$\mathcal{M}$. In Chapter 3, we did not need to introduce γ since our objective functions were invariant to model translation (error terms were defined on pairs of edge vectors, and not on vertices, of the model). In Section 6.3.2, we will introduce an error term defined on model vertices, that does depend on γ; consequently, we represent our model as a function $\mathcal{M}(\theta, D(\beta), \gamma)$.

6.3.2 First stage: shape and pose estimation

Our approach takes as input a sequence of Kinect frames; let ℓ_{seq} denote the number of frames in the sequence. In the following, we assume that the calibration parameters of depth and color cameras, C_Z and C_{RGB}, are known.

For each frame t, the device produces a depth image Z^t and a color image I^t. We process Z^t to obtain a triangulated mesh. Since we know C_Z, we extract a 3D point cloud, containing a 3D point for each pixel in Z^t. To triangulate the point cloud, we exploit neighboring relationships between pixels in Z^t. Given a pixel y in Z^t, denote by $\mathtt{left}(y)$, $\mathtt{right}(y)$, $\mathtt{up}(y)$ and $\mathtt{down}(y)$ the pixels adjacent to y horizontally and vertically: we connect in a triangle the triplets of 3D points corresponding to pixels $(y, \mathtt{right}(y), \mathtt{down}(y))$ and $(\mathtt{right}(y), \mathtt{right}(\mathtt{down}(y)), \mathtt{down}(y))$. This produces a triangulated scan S^t.

For each sequence, we assume we have a "background shot", i.e. one frame capturing the scene without the subject. We denote background scan and RGB image by S_{bg} and I_{bg}, respectively.

The goal of the first stage is to fit the model pose (θ), translation (γ) and shape (β) parameters to the geometric data captured in each frame, S^t. We cannot assume prior knowledge about the subject's motion over long frame sequences; we exploit the fact that body deformations are continuous and predictable over short time intervals.

We optimize model shape, pose (and translation) parameters over short frame intervals of fixed length, ℓ_{int}. Consider an interval starting at frame t, including frames $t, t+1, \ldots, t+\ell_{int}-1$ (for simplicity, we assume $t+\ell_{int}-1 \leqslant \ell_{seq}$ for now). We look for a set of pose vectors Θ^t (where $\Theta^t = \{\theta_i^t : i = 1, \ldots, \ell_{int}\}$), a unique shape vector β^t and a set of translation vectors Γ^t ($\Gamma^t = \{\gamma_i^t : i = 1, \ldots, \ell_{int}\}$). Adopting an analogous notation, we denote by $\mathcal{S}^t$ the set of scans $\{S^t, S^{t+1}, \ldots, S^{t+\ell_{int}-1}\}$ captured during the interval.

Our objective function is a sum of four error terms: a data term E_S and three regularization terms E_θ, E_Θ and E_β. Formally, over an interval starting at frame t, we define the following objective E_t:

$$E_t(\boldsymbol{\beta}^t, \Theta^t, \Gamma^t; S^t, S_{bg}, C_Z) = \sum_{i \in [1, \ell_{int}]} \lambda_S E_S(\boldsymbol{\beta}^t, \theta_i^t, \boldsymbol{\gamma}_i^t; S^{t+i-1}, S_{bg}, C_Z) + \sum_{i \in [1, \ell_{int}]} \lambda_\theta E_\theta(\theta_i^t) + \lambda_\Theta E_\Theta(\Theta^t) + \lambda_\beta E_\beta(\boldsymbol{\beta}^t) \tag{6.1}$$

where λ_S, λ_θ, λ_Θ and λ_β are weighting factors. We now analyze each term in Eq. (6.1) in detail.

Similarly to the geometry-based error term introduced in Section 3.3 (Eq. (3.2)), E_S penalizes distance between model and scan surfaces in 3D space. There are however two differences to take into account. First, unlike the scanning system we used in the previous chapters, Kinect produces scans including both the subject's body and the background: that is, S^t includes "foreground" vertices (reconstructing the subject), as well as "background" vertices (reconstructing the static scene where the subject moves). Background vertices should not be explained by our model. Hence, we define E_S between the surface of S^t and that of mesh $\mathcal{M}_{bg}(\theta_i^t, D(\boldsymbol{\beta}^t), \boldsymbol{\gamma}_i^t)$, which includes the vertices of $\mathcal{M}$ and those of S_{bg}. Second, since we deal with monocular sequences, we need to take into account which model vertices are visible from the depth camera: these are the vertices that should explain the data. For simplicity, we denote by $\mathcal{M}_{vis}$ this set of vertices. Formally, we define the following error term:

$$E_S(\boldsymbol{\beta}^t, \theta_i^t, \boldsymbol{\gamma}_i^t; S^{t+i-1}, S_{bg}, C_Z) = \sum_{\boldsymbol{v}_s \in S^{t+i-1}} \rho\left(\min_{\mathbf{x}_{\mathcal{M}_{vis}} \in \mathcal{M}_{vis}} \|\boldsymbol{v}_s - \mathbf{x}_{\mathcal{M}_{vis}}\|^2 \right) \tag{6.2}$$

where ρ is the Geman-McClure robustifier [58].

We define two pose priors, E_Θ and E_θ. E_Θ assumes that changes in pose occurring within the interval are smooth:

$$E_\Theta(\Theta^t) = \sum_{i \in [2, \ell_{int}-1]} \|2\theta_i^t - \theta_{i-1}^t - \theta_{i+1}^t\|^2 \tag{6.3}$$

E_θ penalizes the squared Mahalanobis distance from a mean pose vector $\boldsymbol{\mu}_\theta$:

$$E_\theta(\theta_i^t) = (\theta_i^t - \boldsymbol{\mu}_\theta)^T \Sigma_\theta^{-1} (\theta_i^t - \boldsymbol{\mu}_\theta). \tag{6.4}$$

To obtain $\boldsymbol{\mu}_\theta$ and Σ_θ, we compute the pose for 39 subjects across more than 700 mocap (motion capture) sequences in the CMU dataset [11] (for a total of 1.7 million frames).

Finally, E_β penalizes the squared Mahalanobis distance from the mean CAESAR shape μ_β:

$$E_\beta(\beta^t) = (\beta^t - \mu_\beta)^T \Sigma_\beta^{-1} (\beta^t - \mu_\beta). \tag{6.5}$$

As explained in Section 6.3.1, we learn the Gaussian parameters μ_β and Σ_β from a corpus of almost 4000 CAESAR scans.

We optimize E_t for each frame t in the sequence. We proceed sequentially, optimizing E_t, then E_{t+1}, then E_{t+2} and so on; we initialize parameters in E_t using the values obtained after minimizing E_{t-1}. Note that, when $t + \ell_{int} - 1 > \ell_{seq}$, we simply consider a shorter interval of frames (precisely, we consider frames within the interval $t, \ldots, \min(t + \ell_{int} - 1, \ell_{seq})$).

After optimization, for each frame we obtain one shape vector and ℓ_{int} pose and translation vectors. Ideally, we would like to have a unique shape estimate for the entire sequence, plus one pose vector and one translation vector per frame. We obtain a shape vector β_{avg} by averaging the set $\{\beta^t : t = 1, \ldots, \ell_{seq}\}$. Analogously, for each frame t we average the set of pose and translation parameters $\{\theta_j^{t-(j-1)} : j = 1, \ldots, \ell_{int}\}$ and $\{\gamma_j^{t-(j-1)} : j = 1, \ldots, \ell_{int}\}$, estimating a mean pose vector θ_{avg}^t and a mean translation vector γ_{avg}^t.

Based on the shape vector β_{avg}, on the pose parameters $\{\theta_{avg}^t : t = 1, \ldots, \ell_{seq}\}$ and on the translation parameters $\{\gamma_{avg}^t : t = 1, \ldots, \ell_{seq}\}$, we obtain a set of meshes $\{\mathcal{M}(\theta_{avg}^t, D(\beta_{avg}), \gamma_{avg}^t) : t = 1, \ldots, \ell_{seq}\}$: they represent complete, topologically coherent 3D reconstructions of the data in each frame.

We use these meshes to compute an initial appearance model U. Analogously to the technique described in Section 3.4 (Eq. (3.12)), we compute U by blending contributions from multiple views. While in the multi-camera setting considered in Section 3.4 multiple views correspond to a unique time instant and multiple viewpoints, here, since we have one RGB camera, the viewpoint is unique; what changes is the time instant considered. As in Eq. (3.12), weights assigned to different views are proportional to the dot product between surface normal and direction towards the camera.

6.3.3 Second stage: appearance-based registration and model learning

We use the per-frame pose and translation parameters and the shape and appearance models obtained during the first stage to initialize a second round of registration.

An important difference between first and second stage is the variables we optimize. In Eq. (6.1), we optimize a set of body model parameters – namely, pose and translation, plus a vector of shape coefficients. The meshes $\{\mathcal{M}(\theta_{avg}^t, D(\beta_{avg}), \gamma_{avg}^t)\}$ we obtain are fully determined by the optimized parameters of the model; the shape of

the subject is approximated by a linear combination of shape basis vectors. In the second stage, we additionally optimize template vertex positions. We obtain a set of deformed templates (alignments) $\{T^t : t = 1, \ldots, \ell_{seq}\}$ and a personalized model of shape deformations, D, which is not constrained in the low-dimensional shape space. During registration, we couple the deformed template T^t to the model, but allow T^t's surface to freely deform to better fit the data. To achieve more accurate registration, in this stage we exploit both depth and color data; note that we do not consider frame intervals, but work on each frame separately.

Our approach draws inspiration from the coregistration framework described in Section 3.3. We define a unique objective function, E_{seq}, as the sum of five error terms: a geometry-based data term E_S, an appearance-based error term E_U and three regularization terms E_{cpl}, E_θ and E_D. Formally, we minimize the following objective function:

$$\begin{aligned} E_{seq}(\{T^t\},\{\theta^t\},D;\{S^t\},S_{bg},C_Z,U,\{I^t\},I_{bg},C_{RGB}) = \\ \sum_{\text{frame } t} \lambda_S E_S(T^t;S^t,S_{bg},C_Z)+ \\ \sum_{\text{frame } t} \lambda_U E_U(T^t;U,I^t,I_{bg},C_{RGB})+ \\ \sum_{\text{frame } t} (\lambda_{cpl}E_{cpl}(T^t,\theta^t,D)+\lambda_\theta E_\theta(\theta^t))+\lambda_D E_D(D) \end{aligned} \tag{6.6}$$

where λ_S, λ_U, λ_{cpl}, λ_θ, and λ_D are weights for the different terms.

The error term E_S is formulated as in Eq. (6.2). Note, however, that here E_S is parameterized by T^t, not by the BlendSCAPE model parameters. The regularization terms E_{cpl} and E_D are analogous to those introduced in Section 3.3.2 (see Eqs. (3.3) and (3.4)). The coupling term E_{cpl} penalizes discrepancy between the aligned template and the model (recall that, as explained in Section 6.3.1, parameters W and Q are not included in the optimization pipeline in this case, and for simplicity are omitted from our notation); E_D promotes spatial smoothness of the shape deformations.

E_U simply computes the sum of per-pixel squared differences between the color image I^t and the synthetic image $\bar{I}^t$. The synthetic image corresponds to our "estimation" of the scene, according to the model: we obtain $\bar{I}^t$ by projecting T^t, textured with texture U, over the background image I_{bg}, according to the calibration parameters C_{RGB}. Mathematically, we define E_U as:

$$E_U(T^t;U,I^t,I_{bg},C_{RGB}) = \sum_{\text{pixel } \mathbf{y}} \|I^t[\mathbf{y}] - \bar{I}^t[\mathbf{y}]\|^2 \tag{6.7}$$

Note that Eq. (6.7) differs from the appearance-based error terms formulated in previous chapters in two ways: in this case, we do not estimate scene lighting and do not

preprocess images applying Ratio-of-Gaussians (RoG) filtering. This is motivated by specific implementation choices, which we discuss in Section 6.5.

6.3.4 Optimization

We minimize objective function (6.1) using Powell's dogleg method [96], with Gauss-Newton Hessian approximation. To compute the gradients of the objective functions, we use an auto-differentiation package called Chumpy [12]. We adopt a a coarse-to-fine approach, in which the resolution of template T^* and the dimensionality of the PCA shape space are progressively increased. We run two iterations: initially, we use a template T^* with 863 vertices and consider only the first 10 principal components; in the second iteration, we increase the resolution of T^* to 6890 vertices, and use 100 principal components. We observed that this helps in avoiding local minima, in particular when aligning geometrically detailed parts like faces. In our experiments, we set ℓ_{int} to 3.

Objective (6.6) is minimized in an iterative fashion. We consider two separate sub-problems, optimizing for $\{T^t\}$ and $\{\theta^t\}$ first, and then for D. As in the first stage, we optimize $\{T^t\}$ and $\{\theta^t\}$ using Powell's dogleg method; minimization with respect to D is a linear least squares problem. For efficiency purposes, the appearance-based error term E_U (Eq. (6.7)) is computed over gray-scale images. We run three iterations, progressively increasing the resolution of the RGB images we consider (we use the following scaling factors: 0.25, 0.5 and 1). We compute texture maps at a resolution of 2048×2048 texels. We keep fairly high the ratio between λ_{cpl} and λ_U, setting it to 15: this means that we adopt a strong regularization towards the model during alignment. We believe that a more accurate tuning of λ_{cpl} and λ_U could bring significant improvements in the quality of our shape models (see also Section 6.4).

Optimizing objective (6.1) over $\ell_{int} = 3$ frames took between 4 and 5 minutes on a desktop machine; optimizing a single alignment in objective (6.6) took about 3 minutes.

6.4 PRELIMINARY RESULTS

We conduct a preliminary study to evaluate the quality of the shape and appearance models provided by our approach. As already pointed out, our technique cannot be considered fully mature: while providing encouraging results, it still exhibits a number of limitations. We analyze them and propose further improvements.

6.4.1 Data acquisition

We used a Kinect 2 device to capture dynamic performances of 6 subjects (4 female and 2 male). For each subject, we captured 4 sequences, corresponding to 4 different motions: walking, dancing, turning around, and an "arbitrary" motion chosen by the subject. Most sequences included fast movements; in many cases, subjects changed significantly their orientation with respect to the camera (providing frontal, lateral and posterior views).

During the capture sessions all subjects wore tight clothing (bicycle shorts for men and women and a sports bra for women). They were allowed to freely move at various distances from the device, from a minimum of 40cm to a maximum of 2.5m. Note that the minimum distance supported by the Kinect depth sensor is 50cm [2]; when the subject is closer, the sensor tends to produce very noisy (and incomplete) data. Prior to scanning, we applied to the skin of each subject black body makeup using a woodcut stamp (the same described in Section 4.2.2), on a dozen locations across the body; this allows us to visually assess the quality of our alignments. Not all subjects gave consent to make their data public, so we show images only for a subset of them.

For each sequence, we captured 100 frames at 30fps. Resolutions of depth and color images are 512×424 and 1920×1080, respectively.

6.4.2 Recovery of shape and motion

Our approach aims at recovering a subject-specific body shape model D, plus a complete 3D reconstruction of the surface of the body (a deformed template T) for each sequence frame. We evaluate results achieved in each task separately.

To evaluate our estimation of D, we adopt an approach similar to [89]: we compute how well this estimated model fits a registration obtained from a higher-quality device, such a full-body 3D scanner. We scanned each subject with the high-resolution 3D body scanner described in Section 3.2, and aligned a template mesh to each scan using the technique presented in Chapter 3. This produces, for each scan, a registered mesh that faithfully represents the scan and conforms the topology of our model. We define a "registration error" in terms of model-to-registration distance: namely, we compute the Euclidean distance per vertex between the registered mesh and our model $\mathcal{M}(\theta, D, \gamma)$, where θ and γ are adjusted to minimize such distance. Figure 36 shows original scan, registration and posed model for one subject.

Figure 37 (right) shows registration errors (average and standard deviation) for all sequences. For comparison, we computed the same errors for the shape model $D(\beta)$ obtained after the first stage (Fig. 37 (left)). Note that for two sequences ("turning

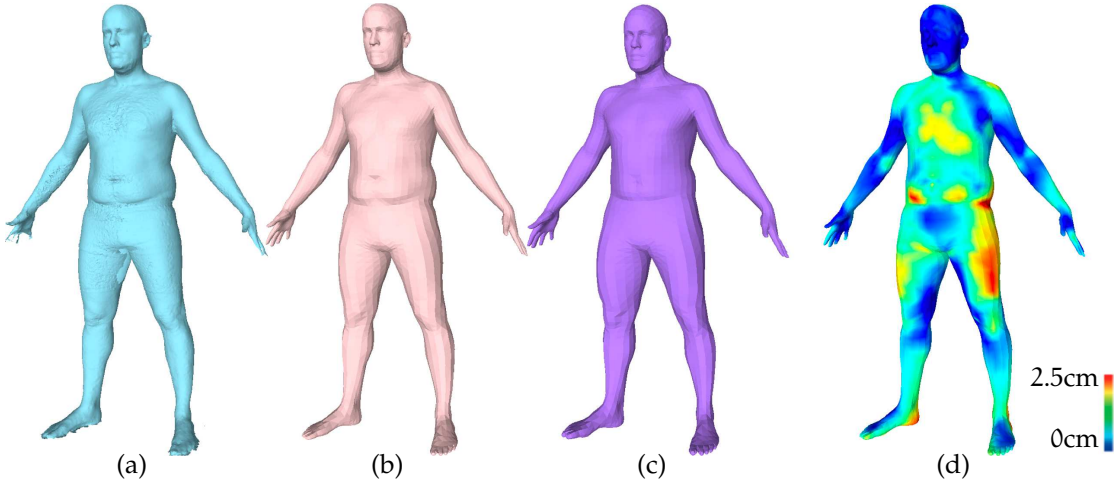


Figure 36: Evaluation of estimated body shape. We capture a high-resolution 3D scan of the subject (a), and align to it a template mesh with the same topology of our model; this produces a registered mesh (b), that we compare against the estimated model (c). We compute a "registration error" as the Euclidean distance per vertex between registration and model. The heat map in (d) shows the error computed between (b) and (c); blue means zero and red means 2.5cm.

around" for subject 4, and "walking" for subject 6), we were not able to run both stages of our technique: alignments obtained after the first stage were not accurate enough to compute an adequate appearance model. This was mostly due to the challenging poses assumed by the subjects: arms adhering to the body in one case, face pointing towards the floor instead of towards the camera in the other. Our approach is currently not robust enough to deal with this.

For all subjects and sequences the average error is below 1cm, except for the "arbitrary" sequence of subject 2 (which corresponds to a challenging "handstand", see Fig. 38(b)). The average error computed over all sequences is 8mm. The approach seems quite robust to variation in motions ("arbitrary" sequences do not exhibit significantly higher errors); this however should be verified on a wider range of motions.

The second stage produces only a minor improvement: errors decrease of 1–2mm on average. We think that a more finely tuned optimization during the second stage (in which, for instance, the weight λ_{cpl} assigned to the coupling term progressively decreases) should help in obtaining more accurate reconstructions.

We evaluate the alignments produced by our method from a qualitative point of view. Figure 38 shows the results obtained on three "arbitrary" sequences: shape and pose estimates are adequate even in the presence of fast motions or challenging poses.

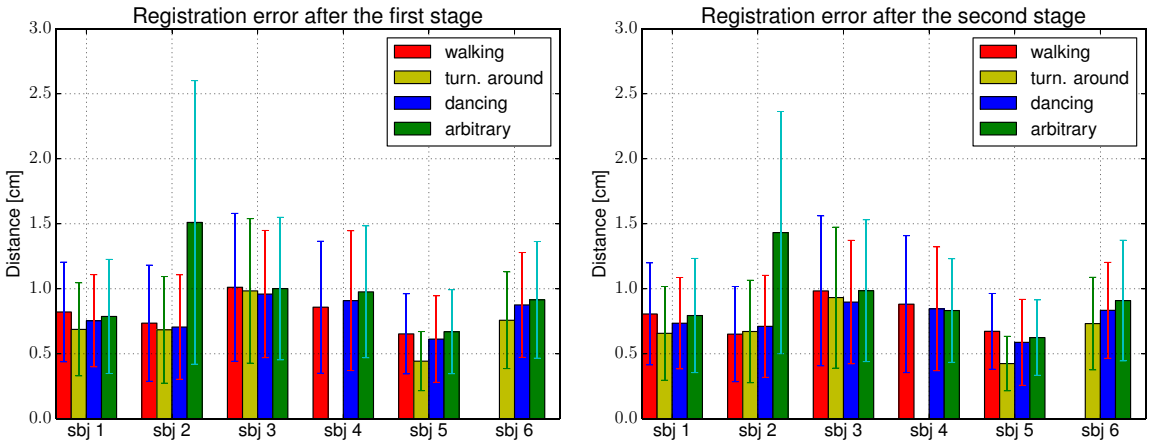


Figure 37: Registration error (average and standard deviation) for all subjects and sequences, showing the accuracy of the estimated body shape. We compute the error for the shape estimated after the first stage of our approach (left) and for the final estimated shape (right).

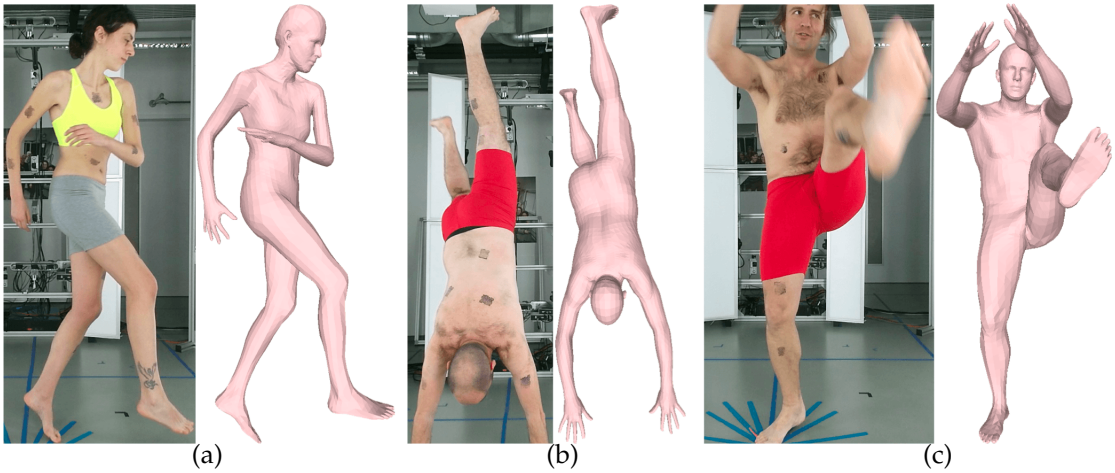


Figure 38: Example RGB frames from three sequences, with the corresponding alignments produced by our method. Alignments adequately represent the data even in challenging situations: arms closed to the body (a), unusual pose (b), fast motion (c).

However, motion reconstruction for body parts that are not visible from the camera over long time intervals tends to be inconsistent. Figure 39 exemplifies the problem. The top row shows 7 subsequent RGB frames; the bottom row shows the corresponding alignments, rendered from a point of view *opposite* to the camera. The estimated motion of the subject's left arm, not visible from the camera in most frames, is inconsistent: it unnaturally changes trajectory twice, in frames 3 and 4. The problem could

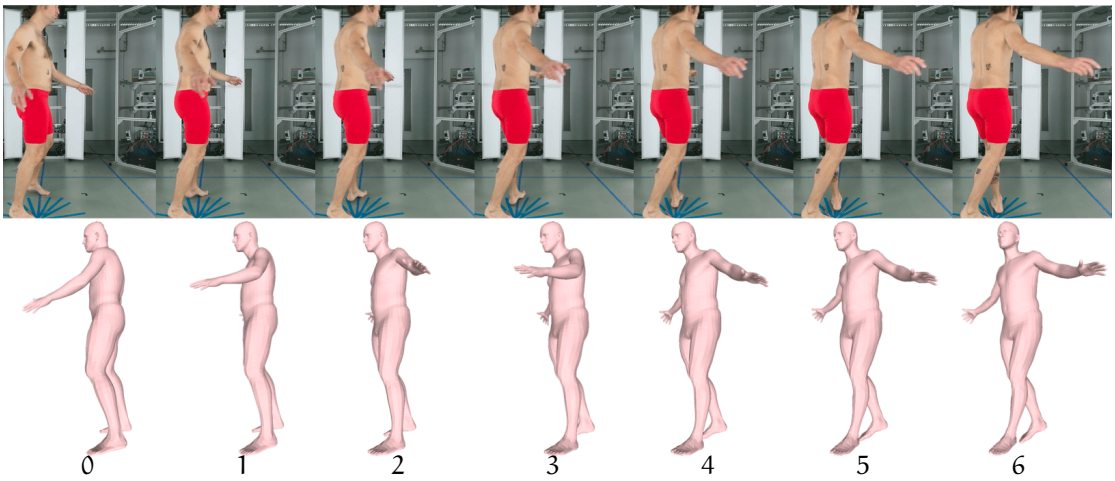


Figure 39: Problematic motion reconstruction for body parts not visible from the camera. Top row: 7 subsequent RGB frames. Bottom row: alignments obtained for those frames, rendered from a point of view opposite to the camera. The estimated motion of the left arm, not visible in most frames, is inconsistent.

be mitigated by including in objective (6.6) a more sophisticated temporal pose prior for human motion (e.g. [79]).

Another limitation of our alignments is that they do not capture transient, fine-scale geometric details (e.g. wrinkles and folds). Note that, in many cases, such details are not captured by the Kinect depth sensor either; however, as recently shown in [138], one could infer high-frequency surface geometry by exploiting shading information in color images (similarly to what is done in shape-from-shading approaches).

6.4.3 Recovery of appearance

We evaluate the quality of our appearance models by comparing them with the texture produced by our high-quality 3D scanner system.

Figure 40 shows a high-quality textured scan and our textured model (obtained from a "walking" sequence). The overall appearance of our model looks realistic; however, high-frequency details (e.g. those of the painted pattern) are not faithfully reproduced. This means that the accuracy of the correspondences established by our alignments should be improved: not surprisingly, "arbitrary" sequences gave worse (though comparable) results in terms of smoothness. Furthermore, textures exhibit seams and discontinuities in some areas (see detail in Fig. 40); we think this is due to alignment inaccuracies and problems of calibration/synchronization between cameras. In partic-

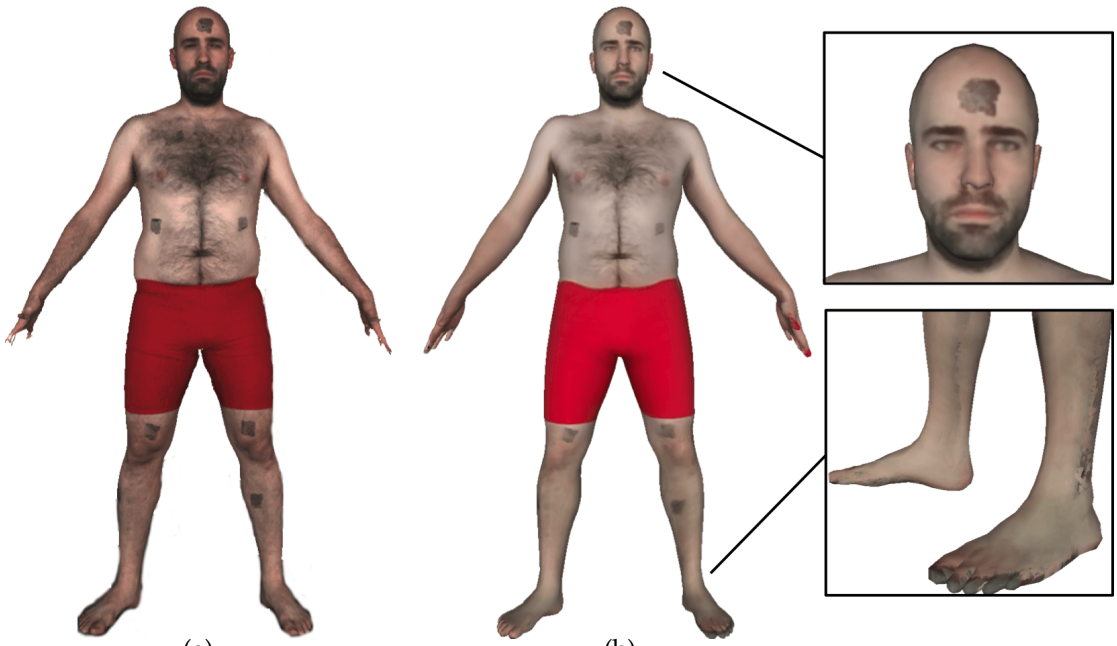
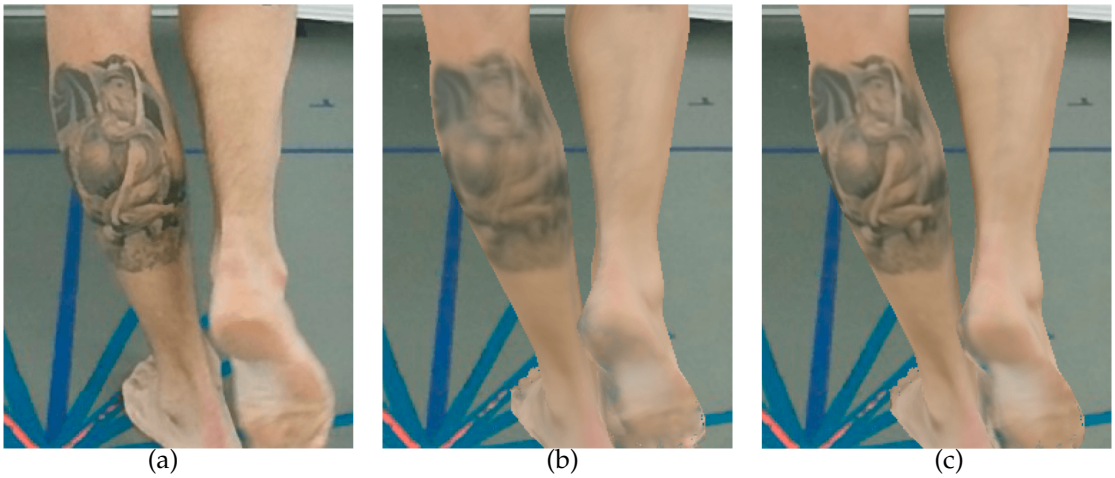


Figure 40: Comparison between a textured high-resolution scan (a) and our textured model (b). Whereas the model looks realistic, our texture does not capture high-frequency details (e.g. the pattern painted on the face is smoothed out) and exhibits some artifacts (seams are well visible on the legs).

ular, we noticed synchronization problems between the shutters of color and depth cameras, and optical distortions in color images.

Note also that the completeness of the appearance models we recover depends on the percentage of body surface actually visible during capture: our approach currently does not infer color information for body parts that are never observed.

To evaluate the impact of using color information during alignment, we compare the appearance models obtained before and after optimizing the second stage of our approach. Figure 41 shows a real image and two synthetic images, obtained after the first and the second stage, respectively (the appearance model is obtained from a "dancing" sequence). The use of color information improves correspondence consistency across frames; we think, however, we could obtain more accurate results if the aforementioned problems of synchronization and optical distortion were explicitly addressed by our method.



(a) (b) (c)

Figure 41: Comparison between a real camera image (detail) (a) and two synthetic images obtained before (b) and after (c) the second stage of our approach.

6.5 THE ROAD AHEAD

In this chapter we have described an approach to estimate personalized body shape and appearance models from monocular Kinect sequences, showing preliminary results on more than 20 sequences capturing 6 subjects in different motions. As discussed in Sections 6.4.2 and 6.4.3, our approach still suffers from a number of shortcomings; we believe that important extensions to our work would be the following:

- When estimating appearance, we do not exploit information about the distance, at each frame, between subject and device; we could obtain higher-quality models by assigning to each view a weight inversely proportional to this distance.
- Our appearance-based error term (Eq. (6.7)) does not include any form of contrast normalization; instead of estimating lighting in a preprocessing stage and then seeking robustness against residual light artifacts (as described in Section 3.5), we would like to improve our estimate including lighting and surface albedo as variables to optimize, following an approach that is quite standard in the literature [60, 88].
- Combining our technique with a shape-from-shading approach would allow us to infer high-frequency geometric details: as in [138], shading cues detected in color images could be exploited to enhance the depth data provided by the device.

- During the second stage we refine each alignment independently from the others; adopting a multi-frame approach (in which, for example, front and back views of the subject are optimized simultaneously) could add more constraints to our optimization, diminishing the need for a strong regularization towards the model.

- Lack of synchronization between depth and RGB cameras and optical distortions may affect the quality of our appearance model, and reduce the effectiveness of our appearance-based error term. One way to mitigate the problem is to include a subset of camera parameters in the optimization pipeline (similarly to [88]); more sophisticated approaches seek non-rigid correction functions for each color image [145].

- Frame-to-frame correspondences established during the first stage could be improved including a color-based error term (dense optical flow or sparse SURF features); another option to explore is the use of RGB-D flow algorithms [69].

- Even though our approach estimates a set of pose parameters per frame, computational times on the order of a few minutes make it unsuitable for real-time tracking. Whereas this is not one of our primary goals, it would be interesting to combine our technique with a discriminative approach (e.g. for pose estimation [120]), and evaluate the tradeoff between accuracy and efficiency.

7 CONCLUSIONS AND FUTURE DIRECTIONS

This chapter summarizes the contributions of our work, discusses its main limitations, and introduces interesting directions of future research.

7.1 SUMMARY

In this thesis we have proposed novel techniques to accurately register 3D real human scans, and build high-quality shape and appearance body models from registered 3D data. Our contributions can be summarized as follows:

- We have developed a novel model-based registration technique for 3D human bodies. Our approach adapts and extends the coregistration framework introduced in [68], that brings a corpus of scans into registration and learns a set of body model parameters defining a unique objective function. The main novelty introduced by our approach is the use of appearance information, in addition to geometry, during registration. While appearance has been used for 3D registration of body parts like faces, full bodies are substantially different: they self occlude and self shadow, and are much more articulated and extended than faces. Our approach estimates scene lighting and surface albedo, and uses the albedo to construct a high-resolution textured 3D model; the model is brought into alignment with multi-camera image data using a robust matching term. Appearance information complements the partial or ambiguous information provided by geometry in smooth 3D areas (like the stomach and the back), producing highly reliable alignments.

- We have proposed FAUST (Fine Alignment Using Scan Texture), a novel dataset and benchmark for 3D mesh registration. FAUST collects real scans of different people in a wide range of poses, with automatically computed dense ground-truth correspondences. To build the dataset, we leveraged our appearance-based registration technique. We defined reliable full-body correspondences between scans by painting the subjects with a high-frequency texture pattern, and placing textured markers on key anatomical locations. Then, we automatically computed

dense scan-to-scan correspondences by aligning a common template to each scan. We verified the quality of our alignments in terms of both geometry and appearance, ensuring that our correspondences are accurate within 2mm.

Previous datasets for 3D mesh registration either provide only synthetic data, or lack reliable mesh-to-mesh correspondences. In this light, they represent an inadequate testbed for real-world applications. We defined a new evaluation methodology and tested on FAUST various well-known 3D registration algorithms, revealing significant shortcomings of existing methods when tested on real data.

- We have explored possible uses of our approach for dermatological applications – namely, for the screening of melanocytic lesions. By adapting our technique and integrating it with a lesion segmentation algorithm, we have developed a fully automated pre-screening system for detecting new lesions or changes in existing ones over almost the entire surface of the body. In a pilot study using synthetic lesions, the system proved able to detect changes as small as 2–3mm. The integration of lesion segmentation with a 3D body model is a key novelty that makes the approach robust to illumination variation, changes in subject pose and shape, and presence of sparse body hair.

- We have outlined a preliminary approach for learning personalized body shape and appearance models from dynamic sequences captured with a single Kinect camera. Many approaches in the literature propose to use the Kinect as a cheap 3D body scanner: these approaches either employ multiple devices, or require the user to stand in one or more predetermined poses, at a constant distance from the device. In contrast, we try not to put any constraints on the subject's motion during capture. Our approach proceeds in two stages: first, we optimize a set of body model pose and shape parameters over short windows of consecutive depth frames; then, we use these estimates to initialize a second round of registration, in which we exploit both depth and color data. We showed preliminary results on a set of 24 sequences, capturing 6 subjects in a range of motions.

7.2 EXTENSIONS AND OPEN PROBLEMS

As pointed out in Section 4.4, our registration algorithm still suffers from a number of limitations, and could be extended in several ways. In particular, we are interested in the following directions:

- Our approach is currently unable to adequately register hands and capture fine-scale face details (e.g. wrinkles). Nonetheless, hands and faces are often the most expressive parts of the body. We could obtain more accurate registrations integrating the BlendSCAPE body model with an accurate model for faces and hands.

- In some cases, alignments produced by our technique exhibit self penetrations; the problem could be mitigated introducing an additional error term in our objectives (e.g. taking into account signed distances between scan and template surfaces).

- Besides naked bodies, our approach should be extended to work with dressed people too. This would require the integration of our technique with an appropriate model of clothing deformations.

- We are interested in evaluating the impact of combining appearance and geometry information in different scenarios: for instance, when registering dynamic sequences captured with high-resolution 4D scanning systems. We expect that the importance of our appearance-based error term in preventing sliding errors might be even greater in dynamic sequences; this would be of value in data-driven modeling of clothing dynamics and human soft tissue deformations.

- Currently, we estimate scene lighting and surface albedo in a preprocessing stage. A more complete formulation of our appearance-based objective function would include lighting and albedo as parameters to optimize, leading to more accurate albedo estimates. Ideally, instead of using contrast normalization to obtain robustness against shading artifacts, our technique could exploit shading cues to reconstruct high-frequency geometric details that are not captured by the scanning device (as in the shape-from-shading approach proposed in [122, 138]).

The system for melanocytic lesion screening introduced in Chapter 5 needs further work to be adopted in the medical practice. In particular, we identify the following areas of research:

- Based on our preliminary results, we should pursue a longitudinal study of dermatological patients to evaluate accuracy and robustness of our approach on a wider range of subjects, and on real lesions.

- The system could be used to monitor, in addition to melanocytic lesions, other lesions (such as those from psoriasis or eczema); to this end, novel ad-hoc segmentation algorithms should be developed.

- In the long term, we would like to use less expensive scanning devices (e.g. consumer RGB-D cameras), perhaps in combination with one or more DSLR (Digital Single-Lens Reflex) cameras, to acquire 3D data and texture. This requires the development of robust algorithms for alignment of single-view data, like the one outlined in Chapter 6.

Currently, we estimate our appearance models by simply blending contributions of different views on a per-texel basis. This approach scales poorly; another limitation is that appearance for body parts that are not visible in any view is not recovered. Two important research lines addressing these problems are the following:

- Analogously to the low-dimensional shape space introduced in Chapter 6, it would be interesting to learn a low-dimensional space for appearance as well. This – perhaps combined with a piecewise smoothness prior – could help in extracting artifact-free models from incomplete/noisy data. Low-dimensional appearance spaces have already been proposed for individual body parts (like faces [32]); computing them for full bodies may present some problems (e.g. due to shading artifacts, differences in clothing), but would have a number of important applications.

- Another interesting line of research is the definition of a framework for super-resolving the appearance of deformable 3D surfaces, bringing into correspondence multiple reconstructions of the same object taken from different views with sub-pixel accuracy. Recent work [59, 125] showed encouraging results in this sense – but no super-resolution approaches proposed so far really handle surfaces undergoing significant non-rigid deformations. Whether it is possible to automatically compute accurate enough correspondences between deformable 3D surfaces still remains an open research problem.

BIBLIOGRAPHY

[1] http://bmivisualizer.com.

[2] http://microsoft.com/en-us/kinectforwindows.

[3] http://cyberware.com.

[4] http://vitronic.de.

[5] http://3dmd.com.

[6] http://ir-ltd.net.

[7] http://4dviews.com.

[8] http://asus.com/Multimedia/Xtion_PRO_LIVE.

[9] http://intel.com/content/www/us/en/architecture-and-technology/realsense-depth-technologies.html.

[10] http://faust.is.tue.mpg.de.

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

[12] http://chumpy.org.

[13] O. Alexander et al. "Digital Ira: Creating a real-time photoreal digital actor." In: *ACM SIGGRAPH 2013 Posters*. 2013, 1:1–1:1.

[14] B. Allen, B. Curless, and Z. Popovic. "The space of human body shapes: Reconstruction and parameterization from range scans." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 22.3 (2003), pp. 587–594.

[15] B. Allen, B. Curless, Z. Popovic, and A. Hertzmann. "Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis." In: *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*. 2006, pp. 147–156.

[16] C. Allene, J. Pons, and R. Keriven. "Seamless image-based texture atlases using multi-band blending." In: *IEEE International Conference on Pattern Recognition (ICPR)*. 2008, pp. 1–4.

[17] B. Amberg. "Editing faces in videos." PhD thesis. University of Basel, 2011.

[18] B. Amberg, S. Romdhani, and T. Vetter. "Optimal step nonrigid ICP algorithms for surface registration." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007, pp. 1–8.

[19] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, H.-C. Pand, and J. Davis. "The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces." In: *Advances in Neural Information Processing Systems (NIPS)*. 2005, pp. 33–40.

[20] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. "SCAPE: Shape Completion and Animation of PEople." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 24.3 (2005), pp. 408–416.

[21] M. Aubry, U. Schlickewei, and D. Cremers. "The wave kernel signature: A quantum mechanical approach to shape analysis." In: *IEEE International Conference on Computer Vision (ICCV) Worskhops*. 2011, pp. 1626–1633.

[22] A. Balan and M. J. Black. "The naked truth: Estimating body shape under clothing." In: *European Conference on Computer Vision (ECCV)*. Vol. 5303. LNCS. 2008, pp. 15–29.

[23] A. Balan, M. J. Black, H. Haussecker, and L. Sigal. "Shining a light on human pose: On shadows, shading and the estimation of pose and shape." In: *IEEE International Conference on Computer Vision (ICCV)*. 2007, pp. 1–8.

[24] A. Balan, L. Sigal, M. J. Black, J. Davis, and H. Haussecker. "Detailed human shape and pose from images." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007, pp. 1–8.

[25] J. Barron and J. Malik. "Color constancy, intrinsic images, and shape estimation." In: *European Conference on Computer Vision (ECCV)*. Vol. 7575. LNCS. 2012, pp. 57–70.

[26] R. Basri and D. Jacobs. "Lambertian reflectance and linear subspaces." In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 25.2 (2003), pp. 218–233.

[27] A. Baumberg. "Blending images for texturing 3D models." In: *British Machine Vision Conference (BMVC)*. 2002, pp. 1–10.

[28] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. "SURF: Speeded up robust features." In: *Computer Vision and Image Understanding* 110.3 (2008), pp. 346–359.

[29] F. Bernardini, I. Martin, and H. Rushmeier. "High-quality texture reconstruction from multiple scans." In: *IEEE Trans. on Visualization and Computer Graphics* 7.4 (2001), pp. 318–332.

[30] P. Besl and N. McKay. "A method for registration of 3D shapes." In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 14.2 (1992), pp. 239–256.

[31] F. Blais. "Review of 20 years of range sensor development." In: *Journal of Electronic Imaging* 13.1 (2004), pp. 231–243.

[32] V. Blanz and T. Vetter. "A morphable model for the synthesis of 3D faces." In: *ACM SIGGRAPH*. 1999, pp. 187–194.

[33] J. Blinn. "Models of light reflection for computer synthesized pictures." In: *ACM SIGGRAPH*. 1977, pp. 192–198.

[34] F. Bogo, J. Romero, M. Loper, and M. J. Black. "FAUST: Dataset and evaluation for 3D mesh registration." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 3794–3801.

[35] F. Bogo, J. Romero, E. Peserico, and M. J. Black. "Automated detection of new or evolving melanocytic lesions using a 3D body model." In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 8673. LNCS. 2014, pp. 593–600.

[36] A. Bronstein, M. Bronstein, L. Guibas, and M. Ovsjanikov. "Shape Google: Geometric words and expressions for invariant shape retrieval." In: *ACM Trans. on Graphics* 30.1 (2011), 1:1–1:20.

[37] A. Bronstein, M. Bronstein, and R. Kimmel. "Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching." In: *Proc. of the National Academy of Sciences (PNAS)* 103.5 (2006), pp. 1168–1172.

[38] A. Bronstein, M. Bronstein, and R. Kimmel. *Numerical geometry of non-rigid shapes*. Springer, 2008.

[39] A. Bronstein, M. Bronstein, R. Kimmel, M. Mahmoudi, and G. Sapiro. "A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching." In: *International Journal of Computer Vision* 89.2–3 (2010), pp. 266–286.

[40] A. Bronstein et al. "SHREC 2010: Robust correspondence benchmark." In: *Eurographics Workshop on* 3D *Object Retrieval (*3D*OR)*. 2010.

[41] B. Brown and S. Rusinkiewicz. "Global non-rigid alignment of 3-D scans." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 26.3 (2007), 148:1–148:10.

[42] C. Cagniart, E. Boyer, and S. Ilic. "Probabilistic deformable surface tracking from multiple videos." In: *European Conference on Computer Vision (ECCV)*. Vol. 6314. LNCS. 2010, pp. 326–339.

[43] E. Catmull. "Subdivision algorithm for computer display of curved surfaces." PhD thesis. University of Utah, 1974.

[44] Y. Chen, Z. Liu, and Z. Zhang. "Tensor-Based Human Body Modeling." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 105–112.

[45] Y. Chen, D. Robertson, and R. Cipolla. "A practical system for modelling body shapes from single view measurements." In: *British Machine Vision Conference (BMVC)*. 2011, pp. 82–91.

[46] R. Coifman and S. Lafon. "Diffusion maps." In: *Applied and Computational Harmonic Analysis* 21.1 (2006), pp. 5–30.

[47] Y. Cui, W. Chang, T. Nöll, and D. Stricker. "KinectAvatar: Fully automatic body capture using a single Kinect." In: *Asian Conference in Computer Vision (ACCV) Workshops*. Vol. 7729. LNCS. 2012, pp. 133–147.

[48] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. "Performance capture from sparse multi-view video." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 27.3 (2008), 98:1–98:10.

[49] E. De Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. "Marker-less deformable mesh tracking for human shape and motion capture." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007, pp. 1–8.

[50] M. Dou, H. Fuchs, and J. Frahm. "Scanning and tracking dynamic objects with commodity depth cameras." In: *International Symposium on Mixed and Augmented Reality (ISMAR)*. 2013, pp. 99–106.

[51] R. Drugge, C. Nguyen, L. Gliga, and E. Drugge. "Clinical pathway for melanoma detection using comprehensive cutaneous analysis with Melanoscan." In: *Dermatology Online Journal* 16.8 (2010), p. 1.

[52] A. Dubrovina and R. Kimmel. "Matching shapes by eigendecomposition of the Laplace-Beltrami operator." In: *International Symposium on* 3D *Data Processing, Visualization and Transmission (*3D*PVT)*. 2010.

[53] E. Dunki-Jacobs, G. Callender, and K. McMasters. "Current management of melanoma." In: *Current Problems in Surgery* 50.8 (2013), pp. 351–382.

[54] M. Eisemann et al. "Floating textures." In: *Computer Graphics Forum (Proc. Eurographics)* 27.2 (2008), pp. 409–418.

[55] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard. "3-D mapping with an RGB-D camera." In: *IEEE Trans. on Robotics* 30.1 (2014), pp. 177–187.

[56] D. Filiberti, P. Bellutta, P. Ngan, and D. Perednia. "Efficient segmentation of large-area skin images: An overview of image processing." In: *Skin Research and Technology* 1.4 (1995), pp. 200–208.

[57] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. "Motion capture using joint skeleton tracking and surface estimation." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 1746–1753.

[58] S. Geman and D. McClure. "Statistical methods for tomographic image reconstruction." In: *Bulletin of the International Statistical Institute* 52.4 (1987), pp. 5–21.

[59] B. Goldlücke, M. Aubry, K. Kolev, and D. Cremers. "A super-resolution framework for high-accuracy multiview reconstruction." In: *International Journal of Computer VIsion* 106.2 (2014), pp. 172–191.

[60] M. de la Gorce, D. Fleet, and N. Paragios. "Model-based 3D hand pose estimation from monocular video." In: *IEEE TPAMI* 33.9 (2011).

[61] P. Guan, L. Reiss, D. Hirshberg, A. Weiss, and M. J. Black. "DRAPE: DRessing Any PErson." In: 31.4 (2012), 35:1–35:10.

[62] P. Guan, A. Weiss, A. Balan, and M. J. Black. "Estimating human shape and pose from a single image." In: *IEEE International Conference on Computer Vision (ICCV)*. 2009, pp. 1381–1388.

[63] D. Hähnel, S. Thrun, and W. Burgard. "An extension of the ICP alogrithm for modeling nonrigid objects with moble robots." In: *International Joint Conference on Artificial Intelligence (IJCAI)*. 2003, pp. 915–920.

[64] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H.-P. Seidel. "Multilinear pose and body shape estimation of dressed subjects from image sets." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2010, pp. 1823–1830.

[65] 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* 28.2 (2009), pp. 337–346.

[66] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2009.

[67] T. Helten, A. Baak, G. Bharaj, M. Müller, H.-P. Seidel, and C. Theobalt. "Personalization and evaluation of a real-time depth-based full body tracker." In: *IEEE International Conference on* 3D *Vision (*3D*V)*. 2013, pp. 279–286.

[68] D. Hirshberg, M. Loper, E. Rachlin, and M. J. Black. "Coregistration: Simultaneous alignment and modeling of articulated 3D shape." In: *European Conference on Computer Vision (ECCV)*. Vol. 7577. LNCS. 2012, pp. 242–255.

[69] M. Hornacek, A. Fitzgibbon, and C. Rother. "SphereFlow: 6 DoF scene flow from RGB-D pairs." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 3526–3533.

[70] H. Huang and P. Bergstresser. "A new hybrid technique for dermatological image registration." In: *IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*. 2007, pp. 1163–1167.

[71] Q. Huang, B. Adams, M. Wicke, and L. Guibas. "Non-rigid registration under isometric deformations." In: *Symposium on Geometry Processing (SGP)*. 2008, pp. 1449–1457.

[72] S. Izadi et al. "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera." In: *ACM Symposium on User Interface Software and Technology (UIST)*. 2011, pp. 559–568.

[73] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. "MovieReshape: Tracking and reshaping of humans in videos." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 29.6 (2010), 148:1–148:10.

[74] Z. Janko and J. Pons. "Spatio-temporal image-based texture atlases for dynamic 3-D models." In: *IEEE International Conference on Computer Vision (ICCV) Worskhops*. 2009, pp. 1646–1653.

[75] I. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[76] O. van Kaick, H. Zhang, G. Hamarneh, and D. Cohen-Or. "A survey on shape correspondence." In: *Computer Graphics Forum* 30.6 (2011), pp. 1681–1707.

[77] M. Kazhdan, M. Bolitho, and H. Hoppe. "Poisson surface reconstruction." In: *Symposium on Geometry Processing (SGP)*. 2006, pp. 61–70.

[78] V. Kim, Y. Lipman, and T. Funkhouser. "Blended intrinsic maps." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 30.4 (2011), 79:1–79:12.

[79] A. Lehrmann, P. Gehler, and S. Nowozin. "Efficient nonlinear Markov Models for human motion." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1314–1321.

[80] V. Lempitsky and D. Ivanov. "Seamless mosaicing of image-based texture maps." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007, pp. 1–6.

[81] H. Lensch, W. Heidrich, and H.-P. Seidel. "A silhouette-based algorithm for texture registration and stitching." In: *Graphical Models* 63.4 (2001), pp. 245–262.

[82] H. Li, B. Adams, L. Guibas, and M. Pauly. "Robust single-view geometry and motion reconstruction." In: *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* 28.5 (2009), 175:1–175:10.

[83] H. Li, R. Sumner, and M. Pauly. "Global correspondence optimization for non-rigid registration of depth scans." In: *Symposium on Geometry Processing (SGP)*. 2008, pp. 1421–1430.

[84] H. Li, E. Vouga, A. Gudym, L. Luo, J. Barron, and G. Gusev. "3D self-portraits." In: *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* 32.6 (2013), 187:1–187:9.

[85] H. Li et al. "Temporally coherent completion of dynamic shapes." In: *ACM Trans. on Graphics* 31.1 (2012), 2:1–2:11.

[86] M. Liao, Q. Zhang, H. Wang, R. Yang, and M. Gong. "Modeling deformable objects from a single depth camera." In: *IEEE International Conference on Computer Vision (ICCV)*. 2009, pp. 167–174.

[87] Y. Lipman and T. Funkhouser. "Möbius voting for surface correspondence." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 28.3 (2009), 72:1–72:12.

[88] M. Loper and M. J. Black. "OpenDR: An approximate differentiable renderer." In: *European Conference on Computer Vision (ECCV)*. Vol. 8695. LNCS. 2014, pp. 154–169.

[89] M. Loper, N. Mahmood, and M. J. Black. "MoSh: Motion and Shape capture from sparse markers." In: *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)*. Vol. 33. 6. 2014, 220:1–220:13.

[90] D. Mateus, R. Horaud, D. Knossow, F. Cuzzolin, and E. Boyer. "Articulated shape matching using Laplacian eigenfunctions and unsupervised point registration." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2008, pp. 1–8.

[91] H. Mirzaalian, G. Hamarneh, and T. Lee. "A graph-based approach to skin mole matching incorporating template-normalized coordinates." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 2152–2159.

[92] H. Mirzaalian, T. Lee, and G. Hamarneh. "Uncertainty-based feature learning for skin lesion matching using a high order MRF optimization framework." In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 7511. LNCS. 2012, pp. 98–105.

[93] L. Mundermann, S. Corazza, and T. Andriacchi. "Accurately measuring human movement using articulated ICP with soft-joint constraints and a repository of articulated models." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007, pp. 1–6.

[94] T. Neumann, K. Varanasi, N. Hasler, M. Wacker, M. Magnor, and C. Theobalt. "Capture and statistical modeling of arm-muscle deformations." In: *Computer Graphics Forum* 32.2 (2013), pp. 285–294.

[95] R. Newcombe et al. "KinectFusion: Real-time dense surface mapping and tracking." In: *International Symposium on Mixed and Augmented Reality (ISMAR)*. 2011, pp. 127–136.

[96] J. Nocedal and S. Wright. *Numerical optimization*. Springer, 2006.

[97] I. Oikonomidis, N. Kyriazis, and A. Argyros. "Efficient model-based 3D tracking of hand articulations using Kinect." In: *British Machine Vision Conference (BMVC)*. 2011, pp. 122–147.

[98] M. Ovsjanikov, Q. Merigot, Q. Memoli, and L. Guibas. "One point isometric matching with the heat kernel." In: *Computer Graphics Forum* 29.5 (2010), pp. 1555–1564.

[99] S. Park and J. Hodgins. "Capturing and animating skin deformation in human motion." In: 25.3 (2006), pp. 881–889.

[100] M. Pauly, N. Mitra, J. Giesen, M. Gross, and L. Guibas. "Example-based 3D scan completion." In: *Symposium on Geometry Processing (SGP)*. 2005, pp. 23–32.

[101] D. Perednia, R. White, and R. Schowengerdt. "Automated feature detection in digital images of skin." In: *Computer Methods and Programs in Biomedicine* 34.1 (1991), pp. 41–60.

[102] D. Perednia, R. White, and R. Schowengerdt. "Automatic registration of multiple skin lesions by use of point pattern matching." In: *Computerized Medical Imaging and Graphics* 16.3 (1992), pp. 205–216.

[103] J. Pierrard and T. Vetter. "Skin detail analysis for face recognition." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007, pp. 1–8.

[104] I. Piryankova, J. Stefanucci, J. Romero, S. de la Rosa, M. J. Black, and B. Mohler. "Can I recognize my body's weight? The influence of shape and texture on the perception of self." In: *ACM Trans. on Applied Perception for the Symposium on Applied Perception* 11.3 (2014), 3:1–13:18.

[105] M. Reuter, F. Wolter, and N. Peinecke. "Laplace-Beltrami spectra as 'Shape-DNA' of surfaces and solids." In: *Computer Aided Design* 38.4 (2006), pp. 342–366.

[106] D. Rigel, J. Russak, and R. Friedman. "The evolution of melanoma diagnosis: 25 years beyond the ABCDs." In: *CA: A Cancer Journal for Clinicians* 60.5 (2010), pp. 301–316.

[107] K. Robinette, H. Daanen, and E. Paquet. "The CAESAR project: A 3-D surface anthropometry survey." In: *International Conference on* 3-D *Digital Imaging and Modeling*. 1999, pp. 380–386.

[108] E. Rodolà, S. Rota Bulò, T. Windheuser, M. Vestner, and D. Cremers. "Dense non-rigid shape correspondence using random forests." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 4177–4184.

[109] G. Sansoni, M. Trebeschi, and F. Docchio. "State-of-the-art and applications of 3D imaging sensors in industry, cultural heritage, medicine, and criminal investigation." In: *Sensors* 9.1 (2009), pp. 568–601.

[110] H. Seo and N. Magnenat-Thalmann. "An example-based approach to human body manipulation." In: *Graphical Models* 66.1 (2004), pp. 1–23.

[111] A. Shapiro et al. "Rapid avatar capture and simulation using commodity depth sensors." In: *Computer Animation and Virtual Worlds* 25.3–4 (2014), pp. 201–211.

[112] L. Sigal, A. Balan, and M. J. Black. "Combined discriminative and generative articulated pose and non-rigid shape estimation." In: *Advances in Neural Information Processing Systems (NIPS)*. 2007, pp. 1337–1344.

[113] P. Sloan, J. Kautz, and J. Snyder. "Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 21.3 (2002), pp. 527–536.

[114] O. Sorkine and M. Alexa. "As-rigid-as-possible surface modeling." In: *Symposium on Geometry Processing (SGP)*. 2007, pp. 109–116.

[115] J. Starck and A. Hilton. "Surface capture for performance-based animation." In: *Computer Graphics and Applications* 27.3 (2007), pp. 21–31.

[116] D. Sun, S. Roth, and M. J. Black. "A quantitative analysis of current practices in optical flow estimation and the principles behind them." In: *International Journal of Computer Vision* 106.2 (2014), pp. 115–137.

[117] J. Sun, M. Ovsjanikov, and L. Guibas. "A concise and provably informative multi-scale signature based on heat diffusion." In: *Computer Graphics Forum* 28.5 (2009), pp. 1383–1392.

[118] S. Taeg, W. Freeman, and H. Tsao. "A reliable skin mole localization scheme." In: *IEEE International Conference on Computer Vision (ICCV)*. 2007, pp. 1–8.

[119] G. Tam et al. "Registration of 3D point clouds and meshes: A survey from rigid to nonrigid." In: *IEEE Trans. on Visualization and Computer Graphics* 19.7 (2013), pp. 1199–1217.

[120] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. "The Vitruvian manidold: Inferring dense correspondences for one-shot human pose estimation." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012, pp. 103–110.

[121] J. Taylor et al. "User-specific hand modeling from monocular depth sequences." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 644–651.

[122] C. Theobalt, N. Ahmed, H. Lensch, M. Magnor, and H.-P. Seidel. "Seeing people in different light – joint shape, motion, and reflectance capture." In: *IEEE Trans. on Visualization and Computer Graphics* 13.4 (2007), pp. 663–674.

[123] N. Thorstensen and R. Keriven. "Non-rigid shape matching using geometry and photometry." In: *Asian Conference in Computer Vision (ACCV) Workshops*. Vol. 5996. LNCS. 2009, pp. 644–654.

[124] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. "Scanning 3D full human bodies using Kinects." In: *IEEE Trans. on Visualization and Computer Graphics* 18.4 (2012), pp. 643–650.

[125] V. Tsiminaki, J. Franco, and E. Boyer. "High resoultion 3D shape texture from multiple videos." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1502–1509.

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

[127] A. Tsoli, N. Mahmood, and M. J. Black. "Breathing life into shape: Capturing, modeling and animating 3D human breathing." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 33.4 (2014), 52:1–52:11.

[128] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. "Face transfer with multilinear models." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 24.3 (2005), pp. 426–433.

[129] D. Vlasic et al. "Dynamic shape capture using multi-view photometric stereo." In: *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* 28.5 (2009), 174:1–174:11.

[130] J. Vogel et al. "Towards robust identification and tracking of nevi in sparse photographic time series." In: *SPIE*. 2014.

[131] H. Voigt and R. Classen. "Topodermatographic image analysis for melanoma screening and the quantitative assessment of tumor dimension parameters of the skin." In: *Cancer* 75.4 (1995), pp. 981–988.

[132] M. Volino, D. Casas, J. Collomosse, and A. Hilton. "Optimal representation of multi-view video." In: *British Machine Vision Conference (BMVC)*. 2014, pp. 105–112.

[133] T. Weise, B. Leibe, and L. Van Gool. "Fast 3D scanning with automatic motion compensation." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007, pp. 1–8.

[134] A. Weiss, D. Hirshberg, and M. J. Black. "Home 3D body scans from noisy image and range data." In: *IEEE International Conference on Computer Vision (ICCV)*. 2011, pp. 1951–1958.

[135] R. Weyrich et al. "Analysis of human faces using a measurement-based skin reflectance model." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)*. Vol. 25. 3. 2006, pp. 1013–1024.

[136] C. Wu, K. Varanasi, Y. Liu, H.-P. Seidel, and C. Theobalt. "Shading-based dynamic shape refinement from multi-view video under general illumination." In: *IEEE International Conference on Computer Vision (ICCV)*. 2011, pp. 1108–1115.

[137] C. Wu, K. Varanasi, and C. Theobalt. "Full-body performance capture under uncontrolled and varying illumination: A shading-based approach." In: *European Conference on Computer Vision (ECCV)*. Vol. 7575. LNCS. 2012, pp. 757–770.

[138] C. Wu, M. Zollhöfer, M. Nießner, M. Stamminger, S. Izadi, and C. Theobalt. "Real-time shading-based refinement for consumer depth cameras." In: *ACM Trans. on Graphics* 33.6 (2014), 200:1–200:10.

[139] S. Wuhrer, C. Shu, and P. Xi. "Landmark-free posture invariant human shape correspondence." In: *The Visual Computer* 27.9 (2011), pp. 843–852.

[140] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt. "Performance capture of interacting characters with handheld Kinects." In: *European Conference on Computer Vision (ECCV)*. Vol. 7573. LNCS. 2012, pp. 828–841.

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

[142] M. Zeng, J. Zheng, and X. Liu. "Templateless quasi-rigid shape modeling with implicit loop-closure." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 145–152.

[143] Y. Zeng, C. Wang, Y. Wang, X. Gu, F. Samaras, and N. Paragios. "Intrinsic dense 3D surface tracking." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011, pp. 1225–1232.

[144] L. Zhang, N. Snavely, B. Curless, and S. Seitz. "Spacetime faces: High resolution capture for modeling and animation." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 23.3 (2004), pp. 548–558.

[145] Q. Zhou and V. Koltun. "Color map optimization for 3D reconstruction with consumer depth cameras." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 33.4 (2014), 155:1–155:10.

[146] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. "Parametric reshaping of human bodies in images." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 29.4 (2010), 126:1–126:10.

[147] M. Zollhöfer et al. "Real-time non-rigid reconstruction using an RGB-D camera." In: *ACM Trans. on Graphics (Proc. SIGGRAPH)* 33.4 (2014), 156:1–156:12.