FAUST: Dataset and evaluation for 3D mesh registration

Federica Bogo[1,2] Javier Romero[1] Matthew Loper[1] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Università degli Studi di Padova, Padova, Italy

Figure 1: FAUST dataset: 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.

Abstract

New scanning technologies are increasing the importance of 3D mesh data and the need for algorithms that can reliably align it. Surface registration is important for building full 3D models from partial scans, creating statistical shape models, shape retrieval, and tracking. The problem is particularly challenging for non-rigid and articulated objects like human bodies. While the challenges of real-world data registration are not present in existing synthetic datasets, establishing ground-truth correspondences for real 3D scans is difficult. We address this with a novel mesh registration technique that combines 3D shape and appearance information to produce high-quality alignments. We define a new dataset called FAUST *that contains 300 scans of 10 people in a wide range of poses together with an evaluation methodology. To achieve accurate registration, we paint the subjects with high-frequency textures and use an extensive validation process to ensure accurate ground truth. We find that current shape registration methods have trouble with this real-world data. The dataset and evaluation website are available for research purposes at* http://faust.is.tue.mpg.de.

1. Introduction

Surface registration is a fundamental problem in computer vision. The identification of a set of dense or sparse correspondences between two surfaces is a prerequisite for common tasks like shape retrieval, registration of range data, or identification of objects in a 3D scene. The task is particularly challenging when the surfaces are those of articulated and deformable objects like human bodies. While many surface matching algorithms have been proposed, little attention has been paid to the development of adequate datasets and benchmarks [21]. This lack is mainly due to the difficulty of dealing with real data.

The popular TOSCA [9] dataset contains synthetic meshes of fixed topology with artist-defined deformations. SHREC [7] adds a variety of artificial noise to TOSCA meshes, but meshes and deformation models created by an artist cannot reproduce what we find in the reality, and artificial noise is quite different from the real thing. To advance the field, datasets and benchmarks should contain *noisy*, *realistically deforming* meshes that *vary in topology*: these are the data real-world applications deal with. The definition of dense ground-truth correspondences, and therefore of a reliable evaluation metric, on such meshes is far from trivial. In this case, common approaches like manual landmarking are time-consuming, challenging, and error-prone for humans – and provide only sparse correspondences.

The registration of human body scans is a challenging problem with many applications; *e.g.*, in tracking [14], statistical modeling [2, 11], and graphics [4]. We present a dataset of human body scans of people of different shapes in different poses, acquired with a high-accuracy 3D multi-stereo system. Ground-truth correspondences are defined

by bringing each scan into alignment with a common template mesh using a novel technique that exploits both 3D shape and surface texture information. In many applications, shape matching has to happen based on surfaces with no texture information; *e.g.*, when aligning two objects with similar shapes and very different textures. But to construct the dataset, texture plays an important role in establishing ground truth. To achieve full-body ground-truth correspondence between meshes, we paint the subjects with a high-frequency texture pattern and place textured markers on key anatomical locations (see Fig. 1). We call the dataset FAUST for Fine Alignment Using Scan Texture.

Our contribution is threefold. First, we present a novel mesh registration technique for human meshes exploiting both shape *and appearance* information. The 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. Our registration process results in highly reliable alignments. Second, on the basis of our alignments, we provide a dataset of 300 real, high-resolution human scans with automatically computed ground-truth correspondences. We verify the quality of the alignments both in terms of geometry and color so that we can ensure high accuracy. Finally, we define an evaluation methodology and test several well-known registration algorithms, revealing significant shortcomings of existing methods when used with real data. FAUST is available for research purposes together with a website for evaluation and publication of results [1].

2. Related work

The literature on surface matching is extremely rich; see [20] for a survey. We briefly review the key themes, with a particular focus on human body registration. Human body shape modeling has received a great deal of attention recently [10, 11, 12] but there is a paucity of high-quality scan data for building and evaluating such models.

One approach starts by defining an intrinsic surface representation that is invariant to bending. This representation is then 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. Common intrinsic representations include Generalized Multi-Dimensional Scaling (GMDS) [8], Möbius transformations [13, 15], and heat kernel maps [16]. 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.

Many practical applications require fitting a common template to noisy scans [3, 11]. Often the template is of lower resolution. Classic approaches employ nonrigid ICP

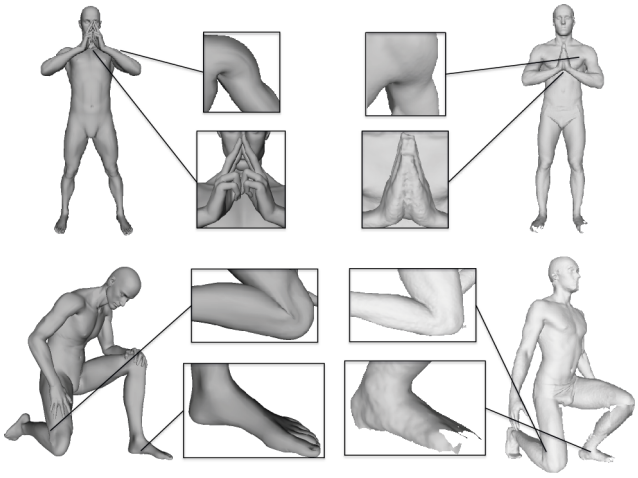

Figure 2: Comparison between TOSCA (left) and FAUST (right). Unrealistic deformations, plus the absence of noise and missing data, make synthetic datasets not representative of real-world meshes.

in conjunction with simple regularization terms favoring surface smoothness [2, 11] or deformations that are as rigid as possible [14]. Since nonrigid ICP is sensitive to local optima, the registration is often initialized by identifying (manually or automatically) a set of corresponding landmarks on both surfaces [21]. The introduction of shape priors, by coupling the template to a learned model during the alignment [10, 12], can increase accuracy and robustness.

These approaches rely only on 3D shape information. Shape alone may not prevent vertices of one mesh from being positioned inconsistently (i.e., sliding) across smooth areas of another. While many regularization methods have been proposed, without ground truth it is unclear how well they work at preventing this sliding.

While texture has been used for 3D model-based alignment of body parts like faces [6], full bodies are substantially different. Their articulated structure is too complex to represent with the cylindrical 2D parameterization in [6]; they self occlude and self shadow; they are too extended to assume a simple lighting model; the size of the body typically means lower-resolution texture as compared with face scans. We are aware of no full-body 3D mesh alignment method that uses texture.

Despite the rich literature on surface matching techniques, the availability of datasets for evaluation is limited. The CAESAR dataset [17] contains several thousand laser scans of bodies with texture information and hand-placed landmarks. It is widely used for alignment – though the only ground truth is for sparse landmarks. CAESAR contains 3 poses but all published work appears to register only the standing "A" pose. Hasler et al. [11] provide a dataset of more than 500 laser scans of 114 different subjects, each

captured in a subset of 34 poses. Each scan has been fitted to a template, and these alignments are publicly available. Unfortunately the quality of the alignments is not quantified, so they cannot be considered ground truth.

TOSCA [9] is a synthetic dataset that is widely used for evaluation of mesh registration methods. It provides 80 artificially created meshes of animals and people (with 3 subjects in a dozen different poses each). Meshes in the same class share the same topology, so ground-truth correspondences are immediately defined. The meshes and the deformations however are unrealistic and there is no noise or missing data. Figure 2 illustrates some differences between TOSCA bodies and real bodies in FAUST.

The SCAPE dataset [4] contains 71 registered meshes of a single subject in different poses. Since the meshes are reconstructed from real data, they are more realistic (*e.g.*, they do not have exactly the same local shape features). The meshes were registered using only geometric information which, as we will show, is unreliable. Hence it is unclear how accurate the deformations in this dataset are.

3. Appearance-based registration

We adapt the coregistration framework of [12], which simultaneously builds a model of the object and its deformations while registering the scans using the model. This approach does not leverage texture information; we add this and introduce a number of other improvements.

3.1. Technique overview

We register a corpus of scans of multiple subjects in multiple poses 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. The registration is performed in two steps: first, we roughly register each scan and learn the parameters of our body model; then, we refine our alignments by introducing a novel appearance-based error term.

The common template T^* is segmented into 31 parts, connected in a kinematic tree structure. Following [12], our body model parameterizes the deformations that fit T^* to a given scan into a set of pose parameters $\boldsymbol{\theta}$ and a set of shape parameters D: $\boldsymbol{\theta}$ collects the relative rotations between neighboring parts, while D defines subject-specific deformations corresponding to the person's body shape. During alignment, T^* is first unstitched into disconnected triangles T_f^*; each triangle is then fit according to a sequence of pose- and shape-dependent deformations:

$$T_f = B_f(\boldsymbol{\theta}) D_f Q_f(\boldsymbol{\theta}) T_f^* \tag{1}$$

where $B_f(\boldsymbol{\theta}) \equiv \sum_i w_{fi} R^i(\boldsymbol{\theta})$ is a linear blend of rigid rotations $R^i(\boldsymbol{\theta})$ of body parts i, and D_f and $Q_f(\boldsymbol{\theta})$ account for deformations dependent on the subject identity and on the pose, respectively. After deformation, the disconnected triangles are stitched into a watertight mesh, T, by solving for vertex positions via least-squares (cf. [3]). While in [12] the blending weights w_{fi} are fixed, we optimize them together with D and Q.

Given a corpus $\{S^k\}$ of scans of different people, p, we compute a preliminary alignment for each scan and simultaneously learn a preliminary model of shape-dependent and pose-dependent deformations by minimizing the following *shape-based* error function E_{shape}:

$$\begin{aligned} &E_{shape}(\{T^k\}, \{\boldsymbol{\theta}^k\}, \{B^k\}, \{D^p\}, Q; \{S^k\}) = \\ &\sum_{\text{scans } k} [E_S(T^k; S^k) + \lambda_C (E_C(T^k, \boldsymbol{\theta}^k, D^{p_k}, Q) + \lambda_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{\theta}^k))] \\ &+ \lambda_C [\lambda_Q E_Q(Q) + \lambda_D \sum_{\text{subjects } p} E_D(D^p)] \end{aligned} \tag{2}$$

where E_S is a data term evaluating the 3D distance between scan and template, E_Q is a regularization term damping the pose-dependent deformations, E_D a smoothness term for the shape space, E_C a regularization term coupling the template to the model, $E_{\boldsymbol{\theta}}$ a pose prior, and $\lambda_C, \lambda_{\boldsymbol{\theta}}, \lambda_D, \lambda_Q$ are weights for the different terms (see [12] for details).

In (2) nothing, apart from the coupling term E_C, prevents the template from sliding along the scan surface where no high-frequency shape information is available; in flat areas, detailed deformations are determined only by the model. In [12], they address this 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.

3.2. Appearance error term

Optimizing (2) provides us with initial alignments $\{T^k\}$ of all the scans in a corpus. These alignments are sufficient to build an initial subject-specific appearance model. To that end, we assume that the albedo of a subject is *consistent* across scans [5] – as is their 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. As we do for pose and shape, we learn an appearance model and (re)align a template to each scan simultaneously.

Synchronized with each 3D scan, S^k, are 22 color cameras, capturing images of the body, I_j^k, from different views j (see [1] for example camera views). Since the calibration parameters, $\boldsymbol{c}_j^k$, of each camera are known, we can project any 3D surface point $\boldsymbol{x}$ onto a 2D point $\pi_j^k(\boldsymbol{x})$ in the image plane of camera j; $I_j^k[\pi_j^k(\boldsymbol{x})]$ returns $\boldsymbol{x}$'s color if $\boldsymbol{x}$ is visible in I_j^k.

We preprocess the original images to discriminate between albedo and irradiance. We assume the illumination can be captured by a Spherical Harmonics (SH) model [18].

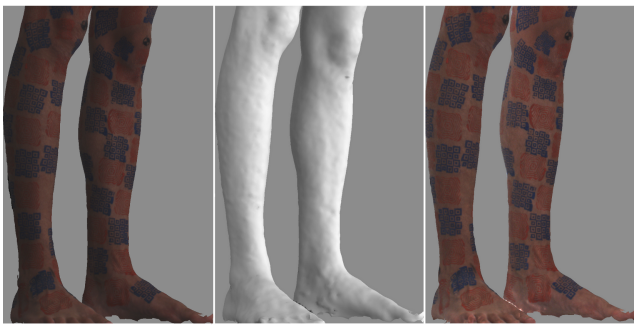

Figure 3: Light and shading (middle) and albedo (right) estimation in one camera image (left).

Since human bodies are extended and articulated, it is critical to model self-casting shadows. We work on each RGB channel independently. For each channel, we represent the light as a 9-dimensional vector $\boldsymbol{l}_{SH}$ (i.e. a 3^{rd} order projection on the SH basis). We assume Lambertian reflectance, and introduce a shadowed diffuse transfer as a 9-dimensional vector, $\boldsymbol{\tau}$, depending only on scan geometry (see [18]). Given a generic scan surface point $\boldsymbol{x}$, its color $i_{\boldsymbol{x}}$ and its albedo $a_{\boldsymbol{x}}$ are related as:

$$i_{\boldsymbol{x}} \quad = \quad (\boldsymbol{\tau}_{\boldsymbol{x}} \cdot \boldsymbol{l}_{SH}) a_{\boldsymbol{x}}.$$

We estimate $\boldsymbol{l}_{SH}$ by minimizing $E_{\boldsymbol{l}}(\boldsymbol{l}_{SH}; \{S^k\}) =$

$$\sum_{\text{scans } k} \sum_{\text{cams } j} \sum_{\text{verts } h} V(\boldsymbol{c}_j^k, \boldsymbol{v}_h^k)(I_j^k[\pi_j^k(\boldsymbol{v}_h^k)] - (\boldsymbol{\tau}_{\boldsymbol{v}_h^k} \cdot \boldsymbol{l}_{SH}) i_{avg})^2$$

where i_{avg} is the average color over the vertices of all the scans and $V(\boldsymbol{c}_j^k, \boldsymbol{v}_h^k)$ is a visibility function returning 1 if $\boldsymbol{v}_h^k$ is visible from a camera with parameters $\boldsymbol{c}_j^k$, 0 otherwise. Given $\boldsymbol{l}_{SH}$, we calculate the irradiance at vertex $\boldsymbol{v}_h^k$ as $(\boldsymbol{\tau}_{\boldsymbol{v}_h^k} \cdot \boldsymbol{l}_{SH})$; at a generic scan surface point $\boldsymbol{x}$ this is given by interpolating between vertices belonging to the same triangle. An albedo image A_j^k is then computed, for any pixel $\boldsymbol{y}$ with corresponding surface point $\boldsymbol{x}$ such that $\boldsymbol{y} = \pi_j^k(\boldsymbol{x})$, as $A_j^k[\boldsymbol{y}] = I_j^k[\boldsymbol{y}]/(\boldsymbol{\tau}_{\boldsymbol{x}} \cdot \boldsymbol{l}_{SH})$. See Fig. 3.

Given the albedo images for each scan, we seek a per-subject albedo model represented as a UV map U^p that is consistent with all scans of that particular subject (see Fig. 4). For any template surface point $\boldsymbol{x}$, we denote by $uv(\boldsymbol{x})$ its mapping from 3D to UV space and by $uv'(\boldsymbol{y})$ its inverse, from the UV map to the surface. We initialize U^p by averaging over the set of maps $\{U^{p_k}\}$ corresponding to subject p; U^{p_k} is obtained from alignment T^k as $U^{p_k}[\boldsymbol{y}] =$

$$\frac{\sum_{\text{cams } j} V(\boldsymbol{c}_j^k, uv'(\boldsymbol{y})) A_j^k[\pi_j^k(uv'(\boldsymbol{y}))] \max(\boldsymbol{\zeta}_{\boldsymbol{c}_j^k} \cdot \boldsymbol{n}_{uv'(\boldsymbol{y})}, 0)}{\sum_{\text{cams } j} V(\boldsymbol{c}_j^k, uv'(\boldsymbol{y})) \max(\boldsymbol{\zeta}_{\boldsymbol{c}_j^k} \cdot \boldsymbol{n}_{uv'(\boldsymbol{y})}, 0)} \tag{3}$$

where $\boldsymbol{n}_{uv'(\boldsymbol{y})}$ is the surface normal at $uv'(\boldsymbol{y})$ and $\boldsymbol{\zeta}_{\boldsymbol{c}_j^k}$ denotes the ray from $uv'(\boldsymbol{y})$ to $\boldsymbol{c}_j^k$'s center.

Per-subject maps are usually noisy and incomplete, since no single pose can provide full-body coverage. Our approach integrates information over multiple per-subject poses, refining each alignment and simultaneously learning an appearance model U^p. We therefore define a data term E_U, penalizing appearance errors, and a regularization term E_{C_U}, penalizing difference from the learned model. Our data term compares real albedo images against a set of synthetic ones rendered from the model. An alignment T^k, in conjunction with a UV map and a set of camera calibration parameters $\boldsymbol{c}_j^k$, renders a synthetic image $\bar{A}_j^k$. For simplicity, we do not model image background. Let $F_j^k(T^k)$ be the intersection between the foreground masks of A_j^k and $\bar{A}_j^k$; the residual image G_j^k evaluates the discrepancy between A_j^k and $\bar{A}_j^k$; $G_j^k[\boldsymbol{y}] =$

$$\begin{cases} (\Gamma_{\sigma_1,\sigma_2}(A_j^k)[\boldsymbol{y}] - \Gamma_{\sigma_1,\sigma_2}(\bar{A}_j^k)[\boldsymbol{y}])^2 & \text{if } \boldsymbol{y} \in F_j^k(T^k) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\Gamma_{\sigma_1,\sigma_2}$ defines a Ratio of Gaussians (RoG) with parameters σ_1 and σ_2. RoG filtering is a form of divisive contrast normalization, invariant to homogeneous light modification; in our multi-camera environment, it provides robustness against differences in color calibration or brightness. Summing over multiple residual images and over the RGB channels, we obtain the error term $E_U(T^k, U^p; \{\boldsymbol{c}_j^k, A_j^k\}) =$

$$\sum_{\text{channels}} \sum_{\text{cams } j} \sum_{\text{pixels } \boldsymbol{y}} G_j^k[\boldsymbol{y}]. \tag{5}$$

The coupling term E_{C_U} enforces consistency across per-subject maps, penalizing deviations from the current model:

$$E_{C_U}(T^k, U^p) = \sum_{\text{pixels } \boldsymbol{y}} (U^{p_k}[\boldsymbol{y}] - U^p[\boldsymbol{y}])^2. \tag{6}$$

Combining (5) and (6) with the 3D shape term already defined in (2), we obtain our global objective – that registers a corpus of scans to a common template and, simultaneously, learns a model of shape, pose and appearance:

$$\begin{aligned} E(\{T^k\}, \{\boldsymbol{\theta}^k\}, \{B^k\}, \{D^p\}, \{U^p\}, Q; \{S^k, \boldsymbol{c}_j^k, A_j^k\}) = \\ E_{shape} + \sum_{\text{scans } k} [\lambda_U E_U(T^k, U^p; \{\boldsymbol{c}_j^k, A_j^k\}) + \\ \lambda_{C_U} E_{C_U}(T^k, U^p)] \end{aligned} \tag{7}$$

where λ_U and λ_{C_U} are weights for the appearance data and coupling term, respectively.

Figure 5 illustrates the benefits of the appearance error term. Texture information adjusts vertex placement mostly in smooth 3D areas (like the stomach and back), complementing the partial or ambiguous information provided by the shape. Using a learned appearance model improves intra-subject correspondences between scans, resulting in sharper estimated albedo texture.

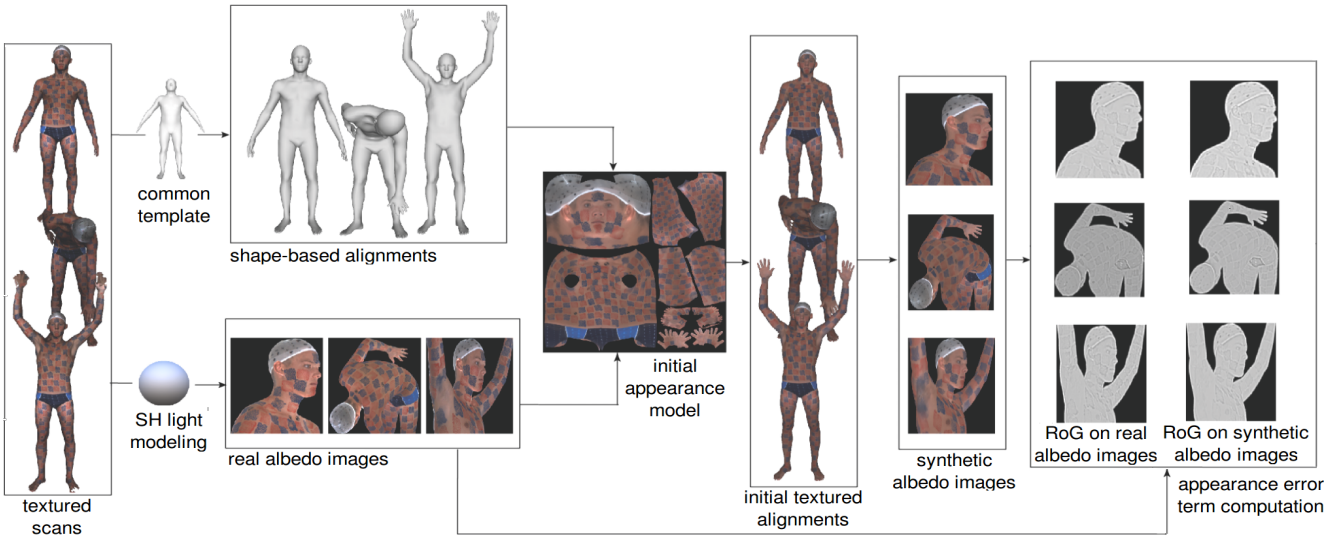

Figure 4: **Overview.** Our registration method proceeds in two steps. 1) a common template is roughly aligned to each scan, using only shape information. The resulting alignments are used to build a subject-specific appearance model. 2) alignments are refined, being brought into registration with a set of pre-processed multi-camera images using a robust matching term.

3.3. Optimization

Objectives (2) and (7) are nonlinear and exhibit a high-dimensional space of solutions; we optimize them in an alternating fashion. For the first phase, our approach is similar to that proposed in [12]. We consider two separate sub-problems, optimizing for $\{T^k\}$ and $\{\boldsymbol{\theta}^k\}$ first, and then for $\{D^p\}$ and Q. In our technique, linearly blended rotations $\{B^k\}$ are optimized together with $\{D^p\}$ and Q.

The second phase adopts a similar approach. After obtaining a set of initial alignments $\{T^k\}$, we keep fixed all the parameters but $\{U^p\}$ and obtain a set of subject-specific appearance models. We then refine each alignment T^k by minimizing (5), for each scan separately. A single alignment, optimizing simultaneously over 22 images (of size 612×512 each, see Sec. 4.1 for details), took less than 5 minutes on a desktop machine equipped with a Quad-core Intel processor and 64GB RAM. A coarse-to-fine approach, in which the variance of both Gaussians in (4) becomes progressively narrower, leads to more accurate alignments. In our experiments, we ran two iterations; σ_1 and σ_2 ranged from 4 to 2 and from 8 to 4, respectively. The ratio between λ_C and λ_U 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 landmarks, by relying on a strong pose prior term $E_{\boldsymbol{\theta}}$. 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.

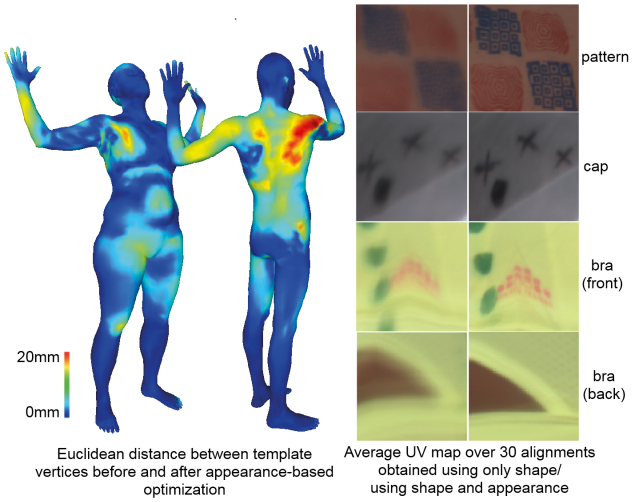

Figure 5: Example comparison between results obtained with and without appearance error minimization. Appearance information prevents sliding effects in smooth areas, providing sharper estimated albedo texture.

4. Building the FAUST dataset

4.1. Acquisition of scans

Our acquisition system is a full-body 3D stereo capture system (3dMD, Atlanta, GA). It is composed by 22 scanning units; each unit contains a pair of stereo cameras for 3D shape computation, 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

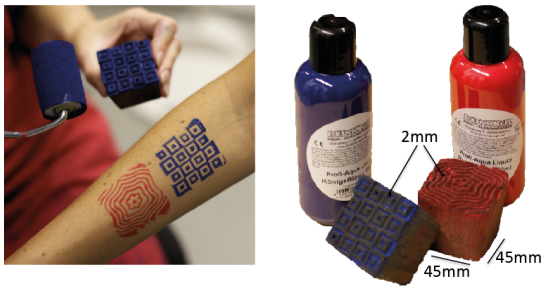

Figure 6: Colored patterns applied to subjects' skin.

Figure 7: Sliding analysis using optical flow. We compute the optical flow between real images (first column) and synthetic ones (second column). Vertices mapped to pixels with high flow magnitude are deemed misaligned.

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 2ms.

The dataset includes 300 triangulated, non-watertight meshes of 10 different subjects (5 male and 5 female), each scanned in 30 different poses. 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.

We provide a training set collecting 10 ground-truth alignments for each subject. All the alignments are watertight meshes with identical topology (with resolution of 6890 vertices). We withhold the alignments for the scans in the test set.

4.2. High-frequency textured bodies

It is 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. 6). Each stamp has a surface of 45×45mm and pattern details up to 2mm in width.

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 subjects. And in general, a correspondence 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 this is 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 for sparse correspondence. 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.

4.3. Ground-truth scan-to-scan correspondences

Our alignments implicitly define a set of *scan-to-scan* correspondences – dense if both scans are of the same subject, sparse otherwise. 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 that we have "ground truth", we identify vertices that are not aligned to an accuracy of 2mm using two techniques.

1: Scan-to-scan distance. Since all scans are in alignment with a common template, we can compute the scan-to-scan correspondence between two scans, S^j and S^k, as follows. For any vertex $\boldsymbol{v}_h^j$ on S^j, find the closest point on the *surface* of the aligned template mesh, T^j. Call this point $T^j(\boldsymbol{v}_h^j)$. If the distance between $\boldsymbol{v}_h^j$ and $T^j(\boldsymbol{v}_h^j)$ is greater than a threshold, t_{shape}, we say that we are not able to provide any correspondence for $\boldsymbol{v}_h^j$. Otherwise, we can uniquely identify $T^j(\boldsymbol{v}_h^j)$ by a face index and a triplet of barycentric coordinates. Since T^j and T^k share the same topology, the same face and barycentric coordinates identify a point $T^k(\boldsymbol{v}_h^j)$ on T^k. Given this point, we find the closest *point*, $S^k(\boldsymbol{v}_h^j)$, on the *surface* of scan S^k. Note our emphasis that this does not compute point-to-point correspondence but point-to-surface (mesh) correspondence.

If the distance between $T^k(\boldsymbol{v}_h^j)$ and $S^k(\boldsymbol{v}_h^j)$ is larger than t_{shape}, then we say that the vertex $\boldsymbol{v}_h^j$ on S^j does not have a corresponding point on S^k. We take $t_{shape} = 2$mm.

2: Sliding. Even scan vertices that are "near enough" to the alignment's surface can still suffer from sliding. *This point 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 synthetic images $\bar{A}_j^k$ rendered by our final model and the real albedo images A_j^k. This is illustrated in Fig. 7.

We compute the optical flow between real and rendered

images using Classic+NL [19] with the default settings. This does quite well with homogeneous differences in lighting between the images. For any vertex $\boldsymbol{v}_h^k$ that is sufficiently visible (i.e. $\boldsymbol{n}_{\boldsymbol{v}_h^k} \cdot \boldsymbol{\zeta}_{\mathbf{c}_j^k} > t_{vis}$, where $t_{vis} = 0.7$), we evaluate the flow magnitude at the image pixel $\pi_j^k(\boldsymbol{v}_h^k)$. We set a threshold t_{app} to 1 pixel. We adopt a conservative approach: vertices mapped to pixels with flow magnitude higher than t_{app} 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 the test 1 alone excludes 10%.

Inter-subject, sparse ground-truth correspondences are obtained from landmarks manually drawn on subjects' skin (see Sec. 4.2). We easily detect the position of each landmark in camera images, and back project identified 2D points to *scan* surface points. For completeness, we evaluated also 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, was 3mm.

4.4. FAUST Benchmark definition

The FAUST benchmark evaluates surface matching algorithms on real scans, on the basis of the ground-truth correspondences defined by the alignments described above. The website is available at http://faust.is.tue.mpg.de. It includes information about data, the file formats and the evaluation metric.

FAUST is split 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 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^j, S^k), the user must submit a 3D point on the surface of S^k for every vertex on S^j. If the matching point is not a surface point of S^k, we compute the closest point on the surface and use this.

We compute 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 report average error over all correspondences and the maximal error.

5. Experimental evaluation

We evaluate the performance of different state-of-the-art registration methods on FAUST, partitioning them into model-free and model-based methods.

5.1. Model-free registration

We test three embedding techniques, focusing on methods with publicly available code: Generalized Multi-Dimensional Scaling (GMDS) [8], Möbius voting [15] and Blended Intrinsic Maps (BIM) [13]. The first method achieves good results on TOSCA, while the last two perform well on both TOSCA and SCAPE [13, 15].

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 convert our original scans to watertight meshes via Poisson reconstruction, keeping them at a fairly high resolution.

The algorithms returned as output a set of sparse (GMDS and Möbius voting) or dense (BIM) correspondences. We compute the Euclidean distance between the returned correspondences and the ground truth; to compare our results with those reported in [13], 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 violates our benchmark, we report errors for the successful scans to get a sense of how FAUST compares in difficulty to previous datasets. We were not able to run GMDS at all because the method does not handle meshes with more than 4000 vertices.

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, [13] 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 identify four principal challenges for these algorithms: missing data, differing mesh topologies between scans, high resolution, and self contact. The algorithms return correspondence with high error even for similar poses when meshes have missing parts (*e.g.* truncated hands or feet) or self contact. Pose variance had in general minor (although not negligible) impact (see Fig. 8).

This evaluation points to one key benefit of FAUST – to evaluate on the dataset, methods will need to be much more robust to real scan data. This should drive the field in a useful direction.

5.2. Model-based registration

We are aware of no publicly available code for model-based registration so we removed the texture-based component of our method resulting in a method similar to [12].

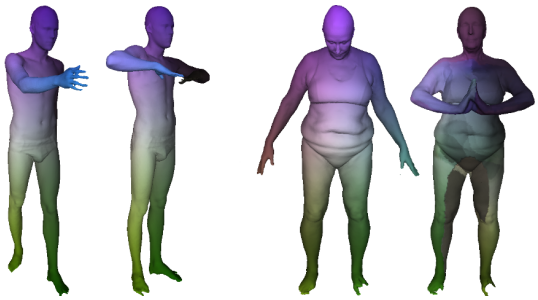

Figure 8: BIM algorithm [13] evaluated on two pairs of FAUST meshes made watertight. Correspondences rendered with identical color. BIM handles pose variation (left pair), but fails to match similar meshes with self contact (right pair).

We also used no landmarks during the alignment process.

On the full FAUST test set, the intra-subject error averaged 7mm; the maximal error was 926mm. When matching different subjects, the average error was 11mm, while the maximal error was 74mm. This is interesting because it quantifies the error one can expect due to sliding of surface points during mesh registration.

6. Conclusion

We presented FAUST, a dataset for evaluation of 3D mesh registration techniques, and a new benchmarking methodology. The 300 human scans in FAUST represent the first set of high-resolution, real human meshes with ground-truth correspondences. We show that registration of real data is substantially more difficult than existing synthetic datasets.

We define ground-truth scan-to-scan correspondences by introducing a novel technique, that registers a corpus of scans to a common template by exploiting both shape and appearance information. With heavily textured subjects, the FAUST scan-to-scan correspondences are accurate to within 2mm. In addition to its value for benchmarking, the FAUST training set, with high-quality alignments, can be used for learning non-rigid shape models.

FAUST is freely available to the research community.

Acknowledgments. FB was supported in part by a Ph.D. fellowship from Univ. Padova, by MIUR (Italy) project AMANDA, and by "Fondazione Ing. A. Gini" (Padova, Italy). We thank E. Holderness for her help with data acquisition.

References

[1] http://faust.is.tue.mpg.de. 2, 3

[2] B. Allen, B. Curless, and Z. Popovic. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Trans. Graph.*, 22(3):587–594, 2003. 1, 2

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, H.-C. Pand, and J. Davis. The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces. *NIPS*, pp. 441–448, 2004. 2, 3

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape Completion and Animation of PEople. *ACM Trans. Graph.*, 24(3):408–416, 2005. 1, 3

[5] 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. *ICCV*, pp. 1–8, 2007. 3

[6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *SIGGRAPH*, pp. 187–194. ACM, 1999. 2

[7] A. Bronstein, M. Bronstein, U. Castellani, A. Dubrovina, L. Guibas, R. Horaud, R. Kimmel, D. Knossow, E. von Lavante, D. Mateus, M. Ovsjanikov, and A. Sharma. SHREC 2010: Robust correspondence benchmark. *3DOR*, 2010. 1

[8] A. Bronstein, M. Bronstein, and R. Kimmel. Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching. *PNAS*, 103(5):1168–1172, 2006. 2, 7

[9] A. Bronstein, M. Bronstein, and R. Kimmel. *Numerical geometry of non-rigid shapes*. Springer, 2008. 1, 3

[10] Y. Chen, Z. Liu, and Z. Zhang. Tensor-based human body modeling. *CVPR*, pp. 105–112, 2013. 2

[11] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H. P. Seidel. A statistical model of human pose and body shape. *Computer Graphics Forum*, 28(2):337–346, 2009. 1, 2

[12] D. A. Hirshberg, M. Loper, E. Rachlin, and M. J. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. *ECCV*, pp. 242–255, 2012. 2, 3, 5, 7

[13] V. G. Kim, Y. Lipman, and T. Funkhouser. Blended intrinsic maps. *ACM Trans. Graph.*, 30(4):79:1–79:12, 2011. 2, 7, 8

[14] H. Li, B. Adams, L. J. Guibas, and M. Pauly. Robust single-view geometry and motion reconstruction. *ACM Trans. Graph.*, 28(5):175:1–175:10, 2009. 1, 2

[15] Y. Lipman and T. Funkhouser. Möbius voting for surface correspondence. *ACM Trans. Graph.*, 28(3):72:1–72:12, 2009. 2, 7

[16] M. Ovsjanikov, Q. Merigot, Q. Memoli, and L. Guibas. One point isometric matching with the heat kernel. *Computer Graphics Forum*, 29(5):1555–1564, 2010. 2

[17] K. Robinette, H. Dannen, and E. Paquet. The CAESAR project: A 3-D surface anthropometry survey. *Conf. 3D Digital Imaging and Modeling*, pp. 380–386, 1999. 2

[18] P. Sloan, J. Kautz, and J. Snyderk. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. *ACM Trans. Graph.*, 21(3):527–536, 2002. 3, 4

[19] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *IJCV*, 106(2):115–137, 2014. 7

[20] O. van Kaick, H. Zhang, G. Hamarneh, and D. Cohen-Or. A survey on shape correspondence. *Computer Graphics Forum*, 30(6):1681–1707, 2011. 2

[21] S. Wuhrer, C. Shu, and P. Xi. Human shape correspondence with automatically predicted landmarks. *MVA*, 23(4):821–830, 2012. 1, 2