Human Body Shape Estimation Using a Multi-Resolution Manifold Forest

Frank Perbet Sam Johnson Minh-Tri Pham Björn Stenger
Toshiba Research Europe, Cambridge, UK
{frank.perbet,sam.johnson,minhtri.pham,bjorn.stenger}@crl.toshiba.co.uk

Abstract

This paper proposes a method for estimating the 3D body shape of a person with robustness to clothing. We formulate the problem as optimization over the manifold of valid depth maps of body shapes learned from synthetic training data. The manifold itself is represented using a novel data structure, a Multi-Resolution Manifold Forest (MRMF), which contains vertical edges between tree nodes as well as horizontal edges between nodes across trees that correspond to overlapping partitions. We show that this data structure allows both efficient localization *and* navigation *on the manifold for on-the-fly building of local linear models (*manifold charting*). We demonstrate shape estimation of clothed users, showing significant improvement in accuracy over global shape models and models using pre-computed clusters. We further compare the MRMF with alternative manifold charting methods on a public dataset for estimating 3D motion from noisy 2D marker observations, obtaining state-of-the-art results.*

1. Introduction

Estimating the body shape of a person offers the potential for applications in the domains of clothes fitting, fitness analysis, and digital content creation. A number of commercial full-body capture systems exist that have been deployed in a range of retail outlets. Such systems, using laser or structured light scanning [2, 3], provide accurate reconstructions, but tend to be costly and require a dedicated capture space. Consumer level depth sensors offer an inexpensive alternative, but pose a number of challenges: The first is the quality and completeness of the data. The current generation of sensors is still relatively noisy (*e.g.* the depth error standard deviation of a sensor using random dot pattern projection is approximately 3-10mm at distances of 1.5-3m where most of the body is in view). The scanned data may also be incomplete and contain holes, which need to be filled in order to obtain a watertight mesh. A method for regularization is therefore required, and typically a parametric body shape model, trained on a large database is used for this [5, 6, 31].

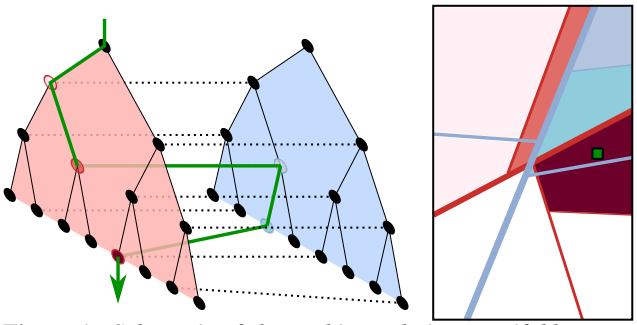


Figure 1: ***Schematic of the Multi-Resolution Manifold Forest.*** *The manifold learning data structure proposed in this paper is based on randomized decision forests. In addition to the standard vertical moves, we additionally allow horizontal traversal between tree nodes, based on a learned manifold graph. Finding the region with the closest mean to the green point, right, with a single search path requires moving between trees. Shown are tree nodes and example path (left) and corresponding search space regions (right) with matching colors.*

A second issue – in certain settings – is clothing. For an accurate measurement users may be willing to undress in the privacy of their home or a dedicated booth. However, for applications in public areas, or for passive measurement, it may be required to estimate the body shape with the user fully dressed. Previous work handling such cases uses skin color segmentation and fits a body shape model only to this partial data [8], while most other work does not address this issue explicitly.

In this paper we deal with both of these issues, and present a method that is able to estimate human body shape under clothing from a single depth map in less than one second. We formulate the task of shape estimation as that of optimizing an energy function over the manifold of human body shapes. The energy function is designed such that it is robust to clothing, leading to solutions which fit *inside* the input depth map, as a person fits inside their clothes (Figure 2). The manifold of possible, unclothed human body shapes is learned from synthetically generated depth measurements. To this manifold we attach a map of generating parameter vectors for pose and shape. Given a segmented input depth map, we first find an initial solution on the man-

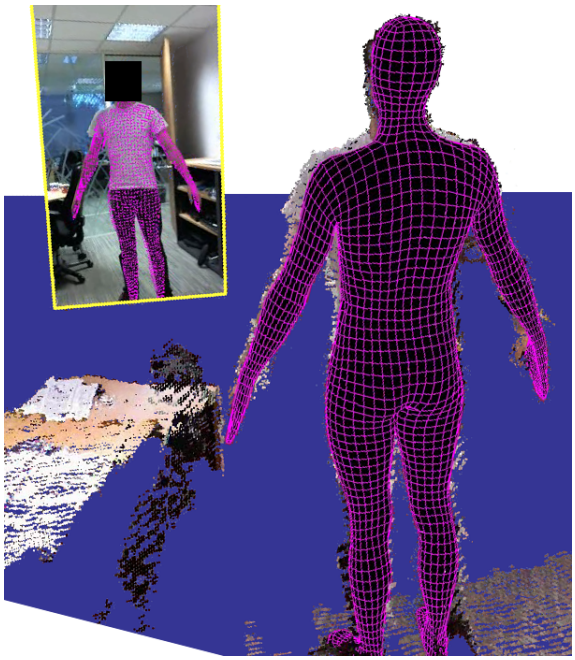


Figure 2: ***Human body shape estimation.*** *A qualitative example of our method. The user stands in front of the system and is shown a visualization of their estimated body shape in less than one second.*

ifold using a similarity measure robust to clothing. Around this location we build a parametric model over the generating parameters for pose and shape *personalized* to the current user. As a final step we use this model in an iterative closest point (ICP) framework and minimize an energy term robust to clothing.

Our work is inspired by work on manifold learning [13, 29]. Unlike global methods that 'unwrap' the manifold, we use a local charting approach, *i.e*. a local linear approximation to parameterize the manifold [20, 24]. At the center of this approach is a novel data structure: the Multi-Resolution Manifold Forest (MRMF). Similar to the manifold forests in [16] it defines multiple random partitions of the space. Crucially, in addition to the standard vertical moves down the trees, the MRMF allows *horizontal* moves between different trees, see Figure 1. We show that these horizontal moves lead to a significant improvement in performance.

Our key contributions are (1) introducing the Multi-Resolution Manifold Forest (MRMF) for representing a manifold and its application to on-the-fly charting and manifold-constrained optimization, (2) evaluating the approach on two different problems: clothed human body shape estimation and 3D motion reconstruction from 2D markers.

1.1. Related work

In this section we briefly review prior methods for body shape modeling and relevant work on manifold learning.

Model-based Body Shape Estimation. A number of methods have been proposed to estimate body shape from multiple images or depth measurements. Early work [5], inspired by the 3D Morphable Models of Blanz and Vetter [12], learned separate models of pose and shape combined with linear blend skinning (LBS) (A discussion of which can be found in [22]). One draw-back of the separation of pose and shape is the inability to model *pose-dependent* deformations such as muscle bulges. This was a motivation for the SCAPE model [6], where the deformation of each mesh triangle is composed of three transformations: (i) the rigid transformation of a single relevant bone, (ii) a pose-dependent transformation learned from the mesh of a single person in multiple poses, and (iii) a shape-dependent transformation learned from meshes of multiple people in a neutral pose. The per-triangle transformations are smoothed in a post-processing step to provide the final, watertight mesh. Previous work fitted the SCAPE model to multiple synchronized images [9], a single minimally clothed image [18], and multiple un-synchronized depth images [31]. Hasler *et al*. semi-automatically fit a human model to a full-body scan of a clothed person by iteratively deforming the mesh and projecting the result back into the space of valid body shapes [19]. These methods provide accurate results, but are still computationally expensive (in the order of minutes).

Recent work extends the concept of pose-dependent shape deformations, proposing a tensor-based body model (TenBo), which conditions the non-rigid triangle deformations on both pose *and* shape, thus capturing, for example, differences between male and female deformation [15]. The method was demonstrated on single depth map fitting with tight clothing and took approximately 1-2 seconds. Handling clothing, *i.e*. estimating the true body shape under clothing, remains a research challenge. One avenue is explicit clothes segmentation and modeling [32], however the number of possible styles and materials makes this a formidable task. This motivates the idea of being as robust to clothing as possible. Previous work relies on two constraints: any body shape estimate must lie *inside* any present clothing, and as close as possible to unclothed skin regions found by color segmentation [8]. We develop this idea further and remove the need for skin segmentation.

Manifold Learning. Segmented, non-clothed human depth images lie on a low-dimensional manifold embedded in the *ambient* space of all possible depth images. Given a novel input – which contains clothing – we wish to localize it on the manifold and use the generating parameters of the local neighborhood to learn a statistical model for optimization. There exists a large body of research on discovering and parameterizing a manifold. The majority of methods seek a map from the ambient data space to a low-dimensional global parameter space – effectively *un-*

wrapping the data, while preserving certain statistical properties of the neighborhood graph [14]. Global methods, such as ISOMAP [29] unwrap the manifold by preserving all geodesic distances in the neighborhood graph. Local, neighborhood preserving methods include Locally Linear Embedding (LLE) [25], which preserves the approximation of a point by a linear combination of its neighbors, and Laplacian Eigenmap [10] and Hessian LLE [17], which preserve the Laplacian and Hessian derivatives of the neighborhood graph.

Global mapping methods fail to adequately model closed manifolds, which are commonplace in vision tasks, *e.g.* the cyclical manifold of human walking poses. To unwrap such a manifold, one has a few options. Firstly, an arbitrary location on the manifold can be selected to apply a 'cut' - thus allowing the manifold to unwrap along its intrinsic dimensions, but losing the continuity. Secondly, a multi-level neighbor graph [21] can be built to support easy navigation, but at the cost of an expensive construction stage. Alternatively, the manifold can be embedded into a higher dimensional space in which closed loops are preserved. Pitelis *et al*. [24] show that these approaches perform poorly on closed manifolds. They propose a piecewise linear model of a manifold, learning an *atlas* of overlapping linear *charts*. In contrast with previous manifold charting approaches of Roweis *et al*. [26] and Brand [13] they do not attempt to unwrap the charts – thereby avoiding loop cutting or the need for spurious extra dimensions. We propose to extend this idea further, and build charts as and when required around a point of interest, approximating the tangent space around this point and maximizing the accuracy of the linear approximation. All previously discussed methods require a two-stage approach: (i) construction of a neighbor graph, and (ii) learning the manifold from this graph. Our MRMF combines the two into a single structure.

2. Building a Manifold Forest

The proposed Multi-Resolution Manifold Forest (MRMF) is an ensemble of randomized space partitioning trees which are connected to each other. During training we learn a graph including the tree edges and edges *between* trees.

2.1. Learning the trees

The aim is to learn an ensemble of trees that are balanced while still maintaining randomization between them. Essentially, the trees can be viewed as defining an adaptive grid on the ambient space similar to *k*-d trees.

The MRMF is a set, $\mathcal{T}$, of binary trees $t_i \in \mathcal{T}$ which hierarchically partition the ambient data space $\mathbb{R}^D$. We train each tree with the same dataset $\mathcal{X} = \{\mathbf{x}_i\}$, $\mathbf{x}_i \in \mathbb{R}^D$, *i.e.* we do not use bagging [16]. In our applications we assume

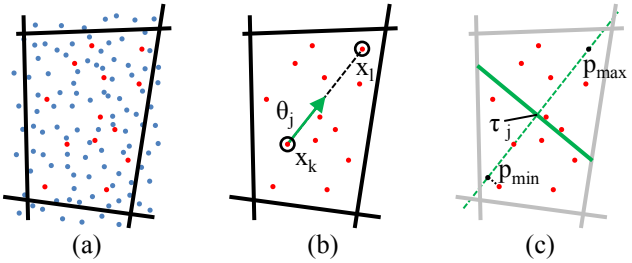


Figure 3: ***Splitting a region.*** *Split function parameters are selected based on a random subset of points (red, (a)) within a region corresponding to a tree node (delimited by the black lines). (b) A random point $\mathbf{x}_k$ and its most distant point $\mathbf{x}_l$ within this subset define θ_j – the normal to a separating hyperplane. (c) The hyperplane splits the region halfway between the minimum and the maximum projected values.*

the samples $\mathbf{x}_i$ to lie on a d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$ with $d < D$.

The parameters $\Theta_j = (\theta_j, \tau_j)$ for each node j define a separating hyperplane in the ambient space $\mathbb{R}^D$ by its unit normal $\theta_j \in \mathbb{R}^D$ and a threshold $\tau_j \in \mathbb{R}$. The data assigned to each node, $\mathcal{X}_j$, is partitioned into two subsets: $\mathcal{X}_j^L$ and $\mathcal{X}_j^R$, depending on the value of the split function $h(\mathbf{x}, \Theta_j) \in \{0, 1\}$. The split functions take the form:

$$h(\mathbf{x}, \Theta_j) = \mathtt{I}(\mathbf{x}^\top \theta_j > \tau_j), \tag{1}$$

where $\mathtt{I}(\cdot)$ is the indicator function. The set $\mathcal{X}_j^L$ contains samples $\mathbf{x} \in \mathcal{X}$ for which $h(\mathbf{x}, \Theta_j) = 0$, the set $\mathcal{X}_j^R$ those with $h(\mathbf{x}, \Theta_j) = 1$. To find Θ_j, we sample a random subset, $\mathcal{D}_j \subset \mathcal{X}_j$, sample a point $\mathbf{x}_k \in \mathcal{D}_j$ and find the most distant point to it in $\mathcal{D}_j$:

$$\mathbf{x}_l = \underset{\mathbf{x} \in \mathcal{D}_j}{\arg\max} \, \|\mathbf{x}_k - \mathbf{x}\| \,. \tag{2}$$

The normal θ_j to the hyperplane is the unit vector between these two points: $\theta_j = (\mathbf{x}_l - \mathbf{x}_k)/(\|\mathbf{x}_l - \mathbf{x}_k\|)$. Figure 3 illustrates the parameter selection process within a region.

In contrast to standard methods [16], the trees are learned in an unsupervised manner without optimizing for a classification or regression objective. Instead, the goal of an MRMF is to define a space partitioning adapted to data located on an unknown manifold. To keep the tree aproximately balanced, the threshold is set to $\tau_j = (p_{max} - p_{min})/2$ where p_{max} (resp. p_{min}) is the maximum (resp. minimum) value of $p_i = \mathbf{x}_i^\top \theta_j$ for all $\mathbf{x} \in \mathcal{D}_j$.

2.2. Learning the graph

The set of nodes, $\mathcal{V}$, of the MRMF graph are simply the nodes of the trees, denoted as $v_i^t \in \mathcal{V}$ with t the tree, and i the node index respectively. The set of edges, $\mathcal{E}$, which are all directed, is composed of all parent-child edges $\mathcal{E}_t$ as well as edges *between* trees $\mathcal{E}_{s,t}$. Formally, the set of edges is defined as:

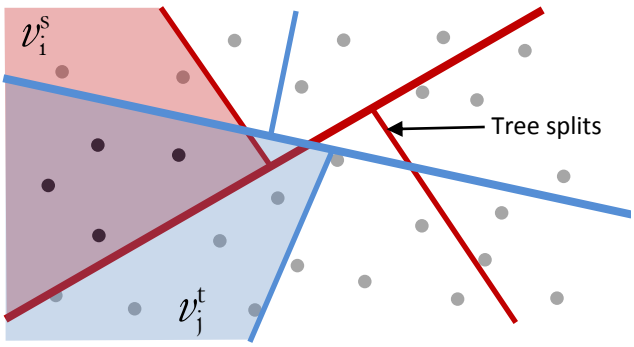


Figure 4: ***Connecting tree nodes with overlap.*** *The red and blue trees define a hierarchical partitioning of the data space. The MRMF connects nodes across trees whose regions intersect, such as red tree node v_i^s and blue tree node v_j^t. The intersection is detected by data samples belonging to both regions (in black). Such horizontal edges allow a search to move between trees.*

$$\mathcal{E} = (\bigcup_{t \in \mathcal{T}} \mathcal{E}_t) \cup (\bigcup_{(s,t) \in \mathcal{T}} \mathcal{E}_{s,t}) \,. \tag{3}$$

While $\mathcal{E}_t$ is defined as part of the tree learning process, learning the inter-tree edges $\mathcal{E}_{s,t}$, is more involved. The idea, as shown in Figure 4, is that two nodes v_i^s and v_j^t in trees s and t are connected if the regions they define *intersect*. Exact computation of these intersections is expensive in high dimensions, even in the case of linear splits [7]. Instead, we use the data samples to estimate intersections, and connect nodes v_i^s and v_j^t if the intersection of their sample sets $\mathcal{D}_i^s$ and $\mathcal{D}_j^t$ is non-empty.

Not all pairs of regions are connected: two regions are connected only if they are tree leaves at the same stage during the training process (trees are grown breadth-first). Doing so ensures that connected regions are of similar volume, resulting in a coarse-to-fine structure. Note that exact computation of region volume is expensive in high dimensions, and many regions are open with infinite volume.

For compactness our implementation of the MRMF graph is pointer-free, as in [28]. The graph is stored in an array, requiring setting a maximum number of edges per node.

3. Optimization on the Manifold

Our aim is to optimize a function f defined on points that lie on a manifold $\mathcal{M}$. We first *locate* an initial solution by traversing, both horizontally and vertically, the trained MRMF. Upon reaching a leaf node we are able to efficiently *navigate* the local neighborhood with the horizontal connections and build a local chart.

3.1. Finding approximate initial solutions

In order to minimize a function f with an MRMF, we first need to find good initial solutions. These solutions are points on the manifold $\mathcal{M}$ which we hope are close to the global solution. Note that we are not restricted to finding a single initial solution. Rather we are looking for k candidates from which horizontal searches are initiated. In the case where $f = d(\mathbf{x}, \mathbf{y})$ and $d()$ is a metric, existing methods such as optimized k-d trees offer an efficient solution [23, 28]. However, these methods do not necessarily extend to the general case, where no assumptions about f can be made. Inspired by the multi-scale approach of graduated optimization [11], we propose a coarse-to-fine graph walk over the MRMF. During tree construction, we keep track of the arithmetic mean $\bar{\mathbf{x}}_j^t$ of all the samples in node v_j^t. During testing we evaluate the function at those points and choose further nodes to explore by evaluating neighbors on the MRMF graph. Since the $\bar{\mathbf{x}}_j^t$ are arithmetic means, they can lie outside the manifold: the only assumption we make on the function f is that it can be evaluated everywhere in the ambient space, that is, for every $\mathbf{x} \in \mathbb{R}^D$.

The search strategy is as follows: a priority queue is initialized with all tree roots giving higher priority to nodes with lower cost function values. The current best candidate is removed from the queue and its children added; if they are leaves they are stored as potential results. The method iterates until reaching its budget of function evaluations. Using horizontal moves increases the chance of finding the leaf node in the graph which minimizes f, allowing to correct for choices made during early tree traversal (Figure 8).

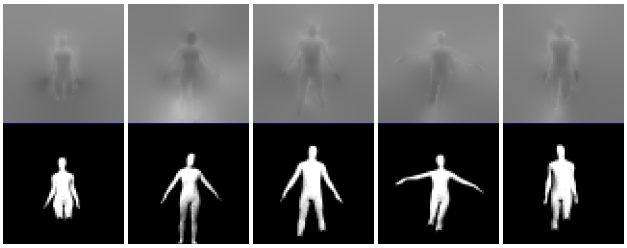
The output of this approximate discrete function minimization method is given as a list of leaf averages $\bar{\mathbf{x}}_j^t$, serving as seed points to compute a local chart on the manifold $\mathcal{M}$. Generally, this can be seen as another characteristic of our method: no graph is explicitly built over the individual training data samples, as in [16, 30] for instance. We believe this is an advantage as the graph representation is smaller and no information is lost.

3.2. Building a local chart

All the leaf averages $\bar{\mathbf{x}}_j^t$ act as the seeds from which the local chart of the manifold $\mathcal{M}$ is computed. Each seed is the starting node of a random walk of horizontal moves in its neighborhood. The walk continues until a given number of nodes has been reached. This parameter controls the local chart size and must therefore be chosen carefully, depending on a given optimization problem.

A local chart of the manifold $\mathcal{M}$ is then computed using Principal Component Analysis (PCA) over the set of nodes reached by the walks (*c.f.* in differential geometry the tangent space is used to compute a local chart). Our mapping is linear and is given by the transformation from the PCA space to the ambient space $c(\mathbf{y}) = \mathbf{x}$ where $\mathbf{y}$ is the vector of coefficients for the first principal components. The chart provides a locally linear parameterization of the space, in which standard methods like gradient descent can be used

(a) Training meshes



(b) Sub-sampled measurement images

Figure 5: ***Training data used for body shape estimation.*** *We learn the manifold of segmented human depth measurements from synthetically rendered samples. The first row shows synthetic 3D meshes generated during training, the second row shows the smoothed and sub-sampled virtual depth measurement and silhouette images from which we learn a manifold. The samples shown here are representative of the pose variation which we train upon.*

to minimize the function $f \circ c$ w.r.t. $\mathbf{y}$. To account for local curvature, using the chart is restricted within a given range of the PCA components. Outside this range, a new chart is recomputed around the new initial solution [4].

4. Body Shape Estimation

We formulate the estimation of human bodies obscured by clothing as optimization over the manifold of *unclothed* body shapes. The function we wish to optimize is asymmetric – we wish to find a solution on the manifold, *i.e.* a nude body shape, which lies *inside* the clothed input. The MRMF allows efficient optimization of such asymmetric functions.

Our model is learned from synthetic depth measurement images (vectorized as $\mathbf{x}$) which are smoothed and sub-sampled (Figure 5(b)). Every element $x \in \mathbf{x}$ is defined as $x = (x^\alpha, x^d)$, with x^α representing the amount of valid information at each pixel, computed from the blurred and sub-sampled silhouette image, and x^d the depth value, computed from the silhouette and depth images using the *premultiplied alpha* compositing method. Our dissimilarity measure is defined between input $\mathbf{x}$ and points on the manifold $\mathbf{y}$ as

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{x}|} k\left(v(x_i^\alpha, y_i^\alpha)\right) + k\left(v(x_i^\alpha x_i^d, y_i^\alpha y_i^d)\right) \quad (4)$$

with the inside function v being defined as

$$v(x, y) = \left|(x - y)\left(1 + \mathtt{I}(x < y)\beta\right)\right| , \quad (5)$$

and where $k(\cdot)$ is a kernel function reducing the influence of outliers. This function induces a penalty of β for manifold points that are greater than input points in either α or depth, *i.e.* they either lie outside the input or in front of it.[1]

We compute initial solutions on the manifold using the approach described in Section 3.1. We then perform a random walk to find a neighborhood within which to build a parametric body model.

For the final estimation of body shape we revert to a standard ICP approach between the original high-resolution point cloud and our parametric body model. The parametric body model is built from the vector field of generating parameters attached to the manifold neighborhood found previously. Our ICP optimization minimizes the following energy function:

$$E(\mathbf{\Phi}) = E_d(\mathbf{\Phi}, \mathbf{q}) + \gamma E_r(\mathbf{\Phi}) , \quad (6)$$

where $\mathbf{\Phi} = (\mathbf{\Phi}_s, \mathbf{\Phi}_p)$ are the parameters for the shape and pose respectively, and $\mathbf{q}$ are the corresponding points in the input depth map to each vertex of our model. The data term is defined as

$$E_d(\mathbf{\Phi}, \mathbf{q}) = \sum_{i=1}^{|\mathbf{q}|} k\left(d(m(\mathbf{\Phi})_i, \mathbf{q}_i)/\sigma\right) , \quad (7)$$

where $m(\mathbf{\Phi})_i$ generates the model vertex in correspondence with $\mathbf{q}_i$, $d(.)$ is a distance function defined below, σ is the noise level, and k a kernel function which increases robustness to outliers. The distance we use is a modified point-to-plane distance of the form

$$d(\mathbf{p}, \mathbf{q}) = inside\left((\mathbf{p} - \mathbf{q})^\top \mathbf{n}_q\right) , \quad (8)$$

where $\mathbf{n}_q$ is the normal at point $\mathbf{q}$. Our clothing-robust inside term,

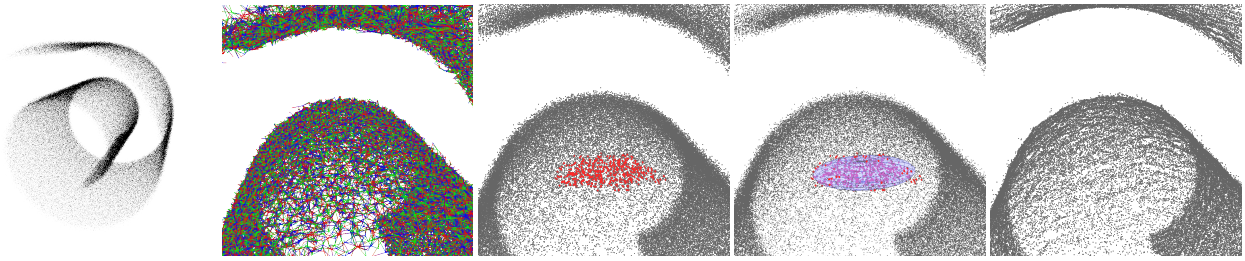
$$inside(y) = y\left(1 + \mathtt{I}(y < 0)\tau_{inside}\right) , \quad (9)$$

gives preference to models beyond the measured depth, *i.e.* the naked shape is within the clothed shape. We iterate between minimizing Equation 6 with Levenberg-Marquardt and finding correspondences. In the correspondence stage we restrict point-to-model matches based on normal directions to improve accuracy.

5. Experiments

In this section we demonstrate the efficacy of our proposed data structure, the MRMF, for both optimization of 3D human body shape and 3D reconstruction of articulated motion. We show that our method is able to optimize asymmetric similarity measures between input points and the learned manifold and handle noisy observations. In all non-toy experiments we outperform the state of the art.

[1] A property of our camera model is that depth values are negated.



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

Figure 6: ***De-noising the Swiss roll.*** *The MRMF lends itself well to de-noising. Given noisy data (a), we learn the MRMF and associated graph, shown magnified in (b). For each point we walk this graph, (c) and build a linear model using PCA (d). To obtain the denoised results each point is projected onto the first two principal components (e).*

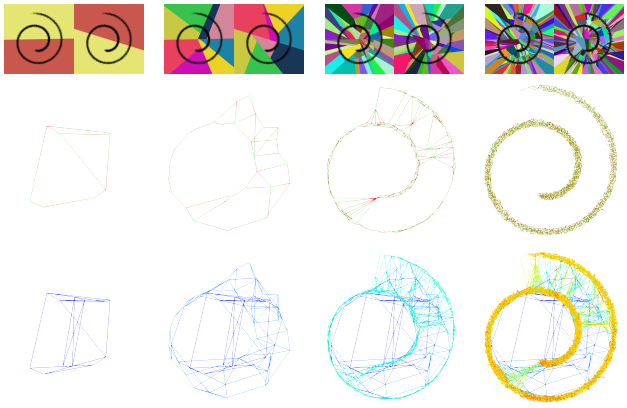


Figure 7: ***Simultaneous tree and graph growing*** *during learning the MRMF on the Swiss roll dataset. The first row shows the regions associated with the leaf nodes of two trees. The second row shows the current inter-node graph for the leaf nodes with node colors representing the tree index. The final row shows the full inter-node graph with colors encoding the depth of each node (from blue to orange).*

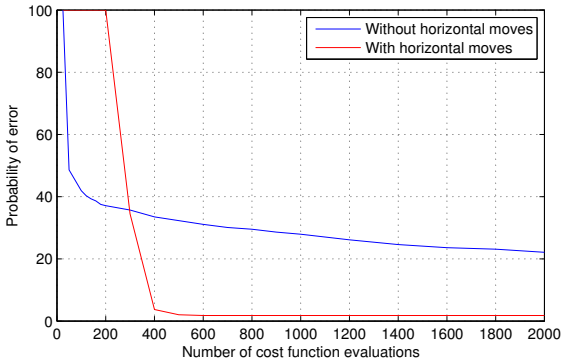


Figure 8: ***Probability of*** **not** ***selecting the global minimum with and without horizontal moves.*** *For a given budget of function evaluations the probability of finding the global minimum increases when allowing horizontal moves. In this experiment an asymmetric similarity measure is used on* $200,000$ *samples from a noisy Swiss roll dataset. Shown are the mean values of* 1000 *queries.*

5.1. Toy examples

First we illustrate key features of the MRMF. Figure 6 shows qualitative de-noising results on a Swiss roll dataset. Building a locally linear chart around *every* point allows efficient de-noising with linear models given a suitable neighborhood size.

Figure 7 visualizes the graph growing process. Connecting tree nodes by their overlap leads to a detailed structure which captures the shape of the manifold. A few edges which 'jump gaps' remain, allowing coarser moves away from local minima during optimization. Figure 8 demonstrates this quantitatively for an asymmetric similarity measure. When minimizing a general function with standard forests, one is forced into a greedy approach, leading to poor final results. The figure shows the probability of selecting a solution other than the global minimum computed over many random trials.

5.2. Human body shape estimation under clothing

Given a noisy, incomplete depth sensor input of a clothed person we estimate their body shape by learning a manifold of depth maps rendered from unclothed human body shapes. Given a clothed input image we use an asymmetric similarity function robust to clothing to optimize on the manifold.

We evaluate the accuracy from four physical measurements taken from eight subjects – height, waist circumference, chest circumference, and shoulder width. From the eight subjects we capture ten depth measurements with varying pose. We define paths for the same physical measurements on the mesh model. This allows us to predict the measurements from an estimation result. All model parameters were estimated on a separate validation dataset of different people.

Training the body shape manifold. To collect training data we register a 3D model to 4,281 scans from the CAESAR dataset [27], obtaining a set of registered human

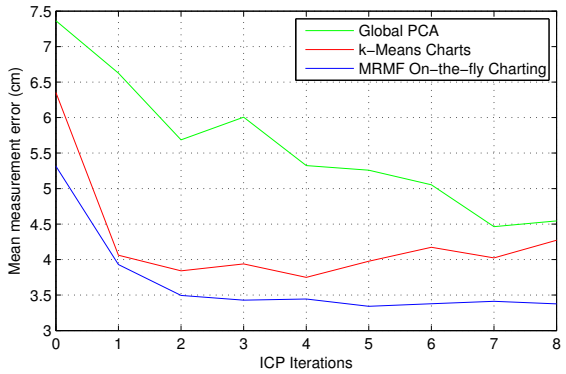


Figure 9: ***Mean body measurement error per iteration for different models.*** *We evaluate the ICP performance per iteration of three different methods of building statistical models. MRMF based on-the-fly charting outperforms both approaches both for initialization accuracy and subsequent optimization due to using all available data for initialization, and models which are precisely localized.*

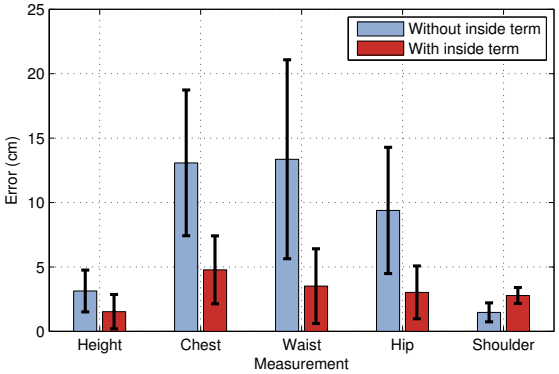


Figure 10: ***Mean errors for five physical measurements over eight different clothing types on a single person.*** *Plotted is the mean absolute error after optimization with, and without the inside term. Note that the inside term reduces both the size and variance of the error.*

meshes with corresponding pose skeletons. We perform an inversion of LBS to 'unskin' each registered mesh into the mean pose. From this set of normalized meshes we generate ten million virtual samples via interpolation (applied to shapes belonging to the same gender). These are perturbed locally by sampling from a learned pose model. To generate virtual depth images we render each model using a virtual camera setup matching the physical setup. During both capture and rendering the depth images are normalized such that the first two directions of maximum variance in 3D space lie parallel to the imaging plane. The normalized depth images are smoothed and down-sampled to 64×64 (Figure 5). The MRMF consists of ten trees of depth 18.

Evaluation. We evaluate the use of the inside term (Equation 4) over simple Euclidean distance between the input depth map and the manifold. We find that the error of the initialization decreases from 10.16cm to 5.31cm, demonstrating the benefit of the inside term. We further evaluate the effects of using different parametric models in the second stage of the fitting, measuring the accuracy per ICP iteration of 3 different approaches: (1) a single global model, (2) pre-computed *local* models, and (3) on-the-fly charting using the MRMF. The results of this are shown in Figure 9. The initialization accuracy of on-the-fly charting is higher than either the local chart initialization or the simple global mean initialization. This is due to the fine sampling of the manifold captured by the MRMF. Furthermore the convergence speed with the on-the-fly models is higher compared to those of the k-Means and global charts. To evaluate the robustness of our cost function to different clothing, we measured the shape estimation accuracy over eight clothing types worn by the same person. Figure 10 shows the results, demonstrating the benefit of employing the inside term in our cost function.

5.3. 3D human motion reconstruction

To compare our approach to methods for manifold learning we carry out the same 2D human motion reconstruction experiment presented by Pitelis *et al*. [24]. The goal is to reconstruct the 3D positions of 31 markers from noisy 2D observations. The data is taken from walking and running sequences in the CMU mocap database [1]. We learn a manifold from the vectorized 3D marker locations of training subjects, find the closest point on the manifold to the back-projected 2D marker locations, and build a chart around this point. This chart is used to reconstruct the 3D marker locations. In Table 1 we present the results of this experiment for two different datasets and two viewpoints each: a side viewpoint, and a generic viewpoint from above. In all experiments we train and test on different subjects and add 3D Gaussian noise with standard deviation of approximately 5cm [2]. Our MRMF on average outperforms a global PCA model and the Atlas method in [24] where numbers are available [3].

6. Conclusion

We have presented a novel data structure, the Multi-Resolution Manifold Forest, for the modeling of manifolds and demonstrated its efficacy in two challenging applications. Our approach to human body fitting, optimizing over the manifold of possible unclothed human body depth maps, was shown to increase robustness to clothing, estimating the

[2]The precise amount of noise added is 4.96cm for the walking only dataset and 5.33cm for the walking + running dataset. Values obtained from correspondence with Pitelis *et al*. [24].

[3]Results of Pitelis *et al*. [24] are omitted for the generic viewpoint as we could not equate our PCA baseline results to theirs.

	Walking Sequences					Walking + Running Sequences				
	Side View			Generic View		Side View			Generic View	
dims	PCA	Atlas	MRMF	PCA	MRMF	PCA	Atlas	MRMF	PCA	MRMF
1	3.50	2.99	**2.65**	5.43	**2.86**	5.88	3.75	**3.43**	7.06	**3.67**
5	2.78	2.64	**2.41**	5.22	**2.72**	3.50	3.02	**2.96**	5.67	**3.38**
10	2.50	2.69	**2.45**	5.32	**2.74**	3.08	2.96	**2.89**	5.78	**3.33**
15	2.58	2.75	**2.56**	5.63	**2.81**	3.24	2.99	**2.95**	6.12	**3.36**
20	2.63	2.79	**2.57**	6.06	**2.82**	3.33	3.03	**2.95**	6.43	**3.36**
21-NN	2.59			2.81		3.32			3.57	
k-NN	ϵ 2.58		k 12	ϵ 2.81	k 15	ϵ 3.24		k 5	ϵ 3.52	k 8

Table 1: ***3D human motion capture reconstruction results.*** *We reconstruct 3D human mocap data from orthographically projected 2D input with noise. The results of our MRMF approach consistently out-perform those using a global PCA and those of the recently proposed Atlas [24], along with those produced by averaging k-nearest-neighbors. We indicate the best score per model dimensionality d in bold, errors are given in cm.*

user's body shape in under one second. In the future we plan to investigate better heuristics to search for initial solutions and to adapt the size of the chart to the local curvature of the manifold.

References

[1] Carnegie Mellon University Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/. 7

[2] TC2 body scanner. www.tc2.com/index_3dbodyscan.html. accessed: 21 March 2014. 1

[3] Vitus 3D body scanner. www.vitronic.de/en/body-scanning. accessed: 21 March 2014. 1

[4] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2007. 5

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

[6] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. *ACM Trans. Graph. (SIGGRAPH)*, 24(3):408–416, July 2005. 1, 2

[7] D. Avis, D. Bremner, and R. Seidel. How good are convex hull algorithms? *Comp. Geometry*, 7(5):265–301, 1997. 4

[8] A. Balan and M. Black. The naked truth: Estimating body shape under clothing. In *ECCV*, pages 15–29, 2008. 1, 2

[9] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *CVPR*, 2007. 2

[10] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2002. 3

[11] A. Blake and A. Zisserman. *Visual reconstruction*, volume 2. MIT press Cambridge, 1987. 4

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

[13] M. Brand. Charting a manifold. In *NIPS*, pages 961–968, 2003. 2, 3

[14] L. Cayton. Algorithms for manifold learning. *UC San Diego, TR CS2008-0923*, 2005. 3

[15] Y. Chen, Z. Liu, and Z. Zhang. Tensor-Based Human Body Modeling. In *CVPR*, 2013. 2

[16] A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013. 2, 3, 4

[17] D. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. National Academy of Sciences*, 100(10):5591–5596, 2003. 3

[18] P. Guan, A. Weiss, A. Balan, and M. Black. Estimating human shape and pose from a single image. In *ICCV*, 2009. 2

[19] N. Hasler, C. Stoll, T. Thormählen, B. Rosenhahn, and H.-P. Seidel. Estimating body shape of dressed humans. *Computer Graphics*, 33(3):211–216, June 2009. 2

[20] X. Huo and J. Chen. Local linear projection (LLP). In *First Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, 2002. 2

[21] R. Krauthgamer and J. R. Lee. Navigating nets: Simple algorithms for proximity search. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pages 798–807, 2004. 3

[22] J. Lewis, M. Cordner, and N. Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *SIGGRAPH*, pages 165–172, 2000. 2

[23] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP (1)*, pages 331–340, 2009. 4

[24] N. Pitelis, C. Russell, and L. Agapito. Learning a Manifold as an Atlas. In *CVPR*, 2013. 2, 3, 7, 8

[25] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000. 3

[26] S. Roweis, L. Saul, and G. Hinton. Global coordination of local linear models. In *NIPS*, pages 889–896, 2002. 3

[27] SAE International. Civilian American and European Surface Anthropometry Resource Project, April 2002. 6

[28] C. Silpa-Anan and R. Hartley. Optimised kd-trees for fast image descriptor matching. In *CVPR*, pages 1–8. IEEE, 2008. 4

[29] J. Tenenbaum, V. de Silva, and J. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000. 2, 3

[30] J. Wang, J. Wang, G. Zeng, Z. Tu, R. Gan, and S. Li. Scalable k-nn graph construction for visual descriptors. In *CVPR*, pages 1106–1113, 2012. 4

[31] A. Weiss, D. Hirshberg, and M. Black. Home 3D body scans from noisy image and range data. In *ICCV*, pages 1951–1958, Nov. 2011. 1, 2

[32] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 2