The Stitched Puppet: A Graphical Model of 3D Human Shape and Pose

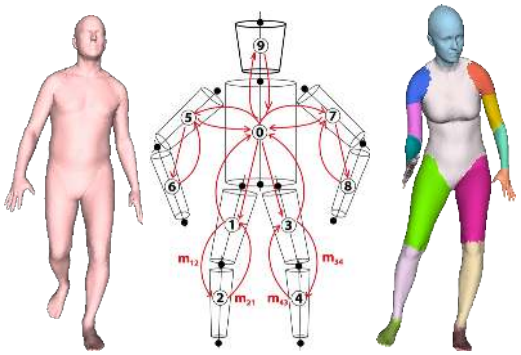
Silvia Zuffi[1,2] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]ITC - Consiglio Nazionale delle Ricerche, Milan, Italy

Abstract

We propose a new 3D model of the human body that is both realistic and part-based. The body is represented by a graphical model in which nodes of the graph correspond to body parts that can independently translate and rotate in 3D and deform to represent different body shapes and to capture pose-dependent shape variations. Pairwise potentials define a "stitching cost" for pulling the limbs apart, giving rise to the stitched puppet *(SP) model. Unlike existing realistic 3D body models, the distributed representation facilitates inference by allowing the model to more effectively explore the space of poses, much like existing 2D pictorial structures models. We infer pose and body shape using a form of particle-based max-product belief propagation. This gives SP the realism of recent 3D body models with the computational advantages of part-based models. We apply SP to two challenging problems involving estimating human shape and pose from 3D data. The first is the FAUST mesh alignment challenge, where ours is the first method to successfully align all 3D meshes with no pose prior. The second involves estimating pose and shape from crude visual hull representations of complex body movements.*

1. Introduction

Inference of human body shape and pose from images [19], depth data [28, 39], 3D scans [3, 10, 20, 32], and sparse markers [23] is of great interest. There are two main classes of 3D body models in use. The first represents 3D body shape and pose-dependent shape variation with high realism (Fig. 1(a)) [4, 7, 14, 20, 28]. Such models are described using a relatively high dimensional state space, combing shape and pose parameters, making inference computationally challenging [31]. The second class of models is based on simple geometric parts connected in a graphical model (Fig. 1(b)) [34, 35]. This approach breaks the global state space into smaller ones allowing each part's parameters to be estimated independently from data. Such models connect the parts via potential functions and inference is performed using message passing algorithms such as belief propagation (BP). These models are advantageous for inference but have a crude geometric structure that does not make it possible to recover body shape and that does not match well to image evidence.



(a) (b) (c)

Figure 1. **3D Body Models.** (a) A SCAPE body model [7] realistically represents 3D body shape and pose using a single high-dimensional state space. (b) A graphical body model composed of geometric primitives connected by pairwise potentials (image reproduced from [35]). (c) The **stitched puppet** model has the realism of (a) and the graphical structure of (b). Each body part is described by its own low-dimensional state space and the parts are connected via pairwise potentials that "stitch" the parts together.

Here we propose a new *stitched puppet* (SP) model that offers the best features of both approaches in that it is both part-based and highly realistic (Fig. 1(c)). The SP model is learned from a detailed 3D body model based on SCAPE [7]. Each body part is represented by a mean shape and two subspaces of shape deformations, learned using principal component analysis (PCA), that model deformations related to intrinsic body shape and pose-dependent shape changes. These shape variations allow SP to capture and fit a wide range of human body shapes. Each part can also undergo translation and rotation in 3D. As with other part-based models, the parts form a graph with pairwise potentials between nodes in the graph. The SP potentials represent a "stitching cost" that penalizes parts that do not fit properly together in 3D to form a coherent shape. Unlike the SCAPE model, parts can move away from each other

but with some cost. This ability of parts to separate and then be stitched back together is exploited during inference to better explore the space of solutions.

Unfortunately, the state space for each part includes continuous random variables representing the part shape and the 3D pose of the part and cannot be easily discretized to apply discrete BP as in pictorial structures models. This is similar to previous 3D part-based models [34] that use inference with continuous random variables. To deal with this, we leverage recent advantages in optimization for distributed models with continuous variables in high dimensional spaces. Namely we perform max-product belief propagation using D-PMP, a particle-based method that has been shown to work well for the similar inference problem of 2D human pose estimation [26].

We apply SP to two challenging 3D inference problems. First, we use it to infer the pose and shape of people in the FAUST dataset [10]. FAUST contains high-resolution 3D scans of people with different body shapes in a wide variety of poses. The goal is to align all the body scans so that points on each scan are in full correspondence with all other scans. The pose variation, noise, self contact, and missing data make the dataset challenging, and our approach is the first to successfully align the full dataset and report quantitative results. Second, we use SP to fit body shape and pose to low-resolution visual hulls with a wide range of poses. Such data is commonly extracted from multi-camera capture systems and most previous pose estimation methods that use detailed body shape assume that the body shape is known a priori from a 3D scan [15, 18] or is refined given a good initialization [12]. Here we show that SP can robustly fit body shape and pose given low-resolution and noisy data, without a good initialisation.

The SP code is available for research purposes [1].

2. Related Work

Representing 3D shape and pose. Geometric descriptions of 3D object shape in terms of parts represent some of the earliest models in vision [24, 25]. Simple geometric models are widely used to represent and track the human body [11, 16, 33], but richer models made from deformable parts of various types are also used [27, 29, 36, 37]. These approaches, however, are fairly crude approximations to human body shape. To address this, Corazza et al. [15] take a body scan of a known subject and chop the 3D mesh into rigid parts. They then build an articulated kinematic tree model and fit it to visual hulls via ICP [9]. We go beyond this to estimate the body part shapes directly from data and to model the interconnection between parts, including their pose-dependent shape variation. Rodgers et al. [32] also use a human-like part-based model with fixed shape for pose estimation from 3D range scan data.

More detailed and realistic models of body shape can be learned from training data of aligned 3D scans of people [4, 7, 12, 14, 20, 28]. For example, the SCAPE model [7] factors mesh deformations due to different body shapes from those dependent from pose. SCAPE models have been fit to multi-camera image data [12], Kinect data [39], mocap markers [23], and high resolution 3D scans [10]. Fitting such a model is difficult because search has to happen in a high-dimensional space of body shape and pose; e.g. Bălan et al. [12] fit SCAPE to image data but require a good initialization from another method.

Part-based models: 2D. Much of the motivation for part-based models comes from work on 2D pictorial structures (PS) [6, 17]. These methods dominate 2D human detection and pose estimation, but are less widely used in 3D. PS models typically represent the body parts by simple rectangles in 2D. A notable exception is the deformable structures model [41], which is a 2D part-based model with more realistic part shapes and pose-dependent deformations. The SP model is similar in spirit to this, but in 3D and with higher realism.

Part-based models: 3D. Distributed inference in 3D using part-based models is more challenging than with 2D models since it is not practical to discretize the state space of poses, making efficient discrete optimization methods inappropriate. Sigal et al. [34] introduce the loose-limbed model, a 3D version of PS, and use non-parametric BP for inference. Their body parts are represented by simple geometric primitives with six shape parameters that are set manually. We use a similar number of shape parameters, but optimize over them. We also use a more sophisticated inference method [26]. There have been several more recent attempts to extend 2D pictorial structures models to enable 3D inference [5, 8, 13], but none of the methods attempt to estimate detailed body shape.

3. Model

The Stitched Puppet (SP) model is a part-based 3D model of the human body parameterized by pose, intrinsic shape, and pose-dependent shape deformations. Intrinsic shape is the body shape that varies between people due to gender, age, height, weight, fitness, etc. Pose-dependent shape deformations capture shape changes due to muscle bulging and soft tissue motion. The model is composed of 16 body parts: head, torso, shoulders, upper arms, lower arms, upper legs, lower legs, hands and feet (see color coding in Fig. 1(c)). The SP model is a tree-structured graphical model in which each body part corresponds to a node, with the torso at the root. Each part is represented by a triangulated 3D mesh in a canonical, part-centered, coordinate system. Let i be a node index, with $i \in [0..15]$. The node variables are represented by a random vector:

$$\mathbf{x}_i = [\mathbf{o}_i^T, \mathbf{r}_i^T, \mathbf{d}_i^T, \mathbf{s}_i^T]^T, \quad (1)$$

where $\mathbf{o}_i$ is a 3D vector representing the location of the center of the part in a global frame and $\mathbf{r}_i$ is a three-dimensional Rodrigues vector representing the rotation of the part with respect to a reference pose. The reference pose is the pose of the part in the *template* mesh (Fig. 2(a)). The parts also have two vectors of linear shape coefficients, $\mathbf{d}_i$ and $\mathbf{s}_i$, that represent pose-dependent deformations and intrinsic body shape, respectively. Learning these shape deformation models, using PCA, is described below.

From model variables to meshes. Given a set of node variables $\mathbf{x}_i$, the mesh vertices for the part i are generated as follows. First, we use the intrinsic shape parameters $\mathbf{s}_i$ to generate a deformed template mesh with the desired intrinsic shape (Fig. 2(b)):

$$\mathbf{q}_i = B_{s,i}\mathbf{s}_i + \mathbf{m}_{s,i}, \tag{2}$$

where $\mathbf{m}_{s,i}$ is a vector of 3D vertices of part i in a local frame with origin in the part center corresponding to a part with mean intrinsic shape across all training body shapes. $B_{s,i}$ is the matrix of PCA basis vectors of size $3N_i \times n_s$, where $n_s = 4$. As described below, the PCA subspace for the intrinsic shape, $B_{s,i}$, is learned over meshes in a template pose, where we assume there are no pose-dependent deformations. $\mathbf{m}_{s,i}$ is a column vector of $3N_i$ vertex coordinates, where N_i is the number of vertices of the part i. Unlike SCAPE [7], where shape deformations are transformations operating on triangles, SP is much simpler[1]. Since each body part has its own coordinate system, the deformations can be applied directly to *vertices* in this coordinate frame. The resulting vector $\mathbf{q}_i$ represents the vertex coordinates in the local frame of the mean part deformed to represent the desired intrinsic body shape.

We next apply pose-dependent deformations to the modified template, $\mathbf{q}_i$ (Fig. 2(c)):

$$\mathbf{p}_i = B_{p,i}\mathbf{d}_i + \boldsymbol{\mu}_{p,i} + \mathbf{q}_i, \tag{3}$$

where $B_{p,i}$ is the matrix of PCA basis vectors for the pose-dependent deformation model of part i, $\boldsymbol{\mu}_{p,i}$ is the mean of the pose-dependent deformations in the training set with respect to the template, and the resulting $\mathbf{p}_i$ represents the local coordinates of the part after shape deformations have been applied. Now, given the part center $\mathbf{o}_i$ and the Rodrigues vector $\mathbf{r}_i$, a rigid 3D transformation is applied to the vertices in local frame $\mathbf{p}_i$ to convert them into a global coordinate system, $\tilde{\mathbf{p}}_i$ (Fig. 2(d,e)).

Learning the model. We learn SP from instances of a SCAPE model; the details of SCAPE are not important here and we refer the reader to [7]. Specifically, we use the model from [23]. What SCAPE gives us, and why we use it, is a set of training meshes in a wide range of body shapes and poses that are in complete correspondence and are segmented into parts. The SCAPE model we consider has a template mesh in a "T-pose," and is segmented into 19 parts. For SP we take the same template mesh but segment it into fewer parts; in particular, we treat the torso as a single part (Fig. 4), merging the upper torso and the pelvis.

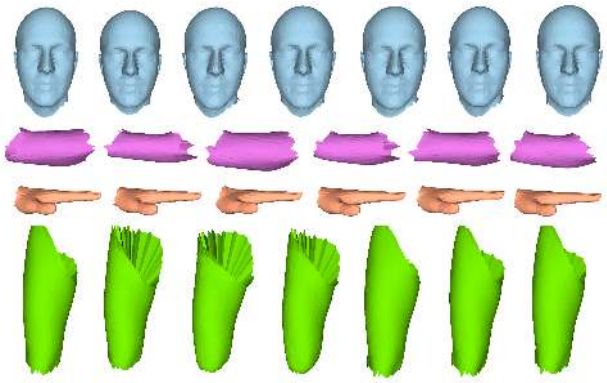


Figure 3. **SP parts**. Examples of training samples for SP for the parts head, right upper arm, left hand, right upper leg. Parts are independent meshes in local coordinate systems.

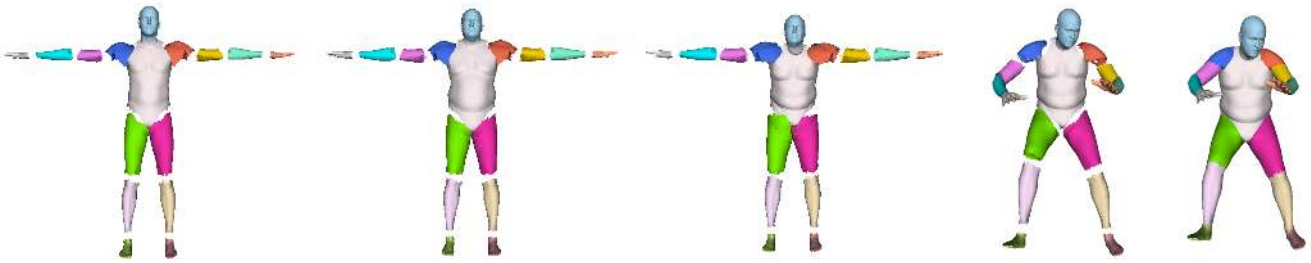
To create training data, we sample a set of 600 poses from motion capture data[2] and for each pose we generate 9 more by adding small amounts of noise to the part rotations. This gives 6000 SCAPE meshes in different poses for training. We also generate 600 samples of bodies with different intrinsic shapes in the template pose by sampling from the SCAPE body shape model. From this set of samples, where we only vary body shape, we learn the intrinsic shape model of our parts. Note that we learn separate models for men and women. Also note that in SP the intrinsic shape and pose-dependent deformations are independent as in SCAPE. We did not consider dependencies among them as the training samples come from a model that assumes independence.

We define the SP mesh topology as a "chopped" version of SCAPE, where each part is an independent mesh with a locally defined assignment of vertices to faces (Fig. 3). For neighboring body parts, we duplicate the vertices that are in common, creating a set of "interface points" that should match when the parts are stitched together.

Given the part segmentation of the training meshes, we take each part and transform it to a canonical coordinate system (Fig. 3). The vertices in this frame for each part form one training example. We compute the mean shape as the mean location of each vertex, subtract this, and perform PCA. For the pose-dependent shape deformations we learn 16 independent PCA models, $B_{p,i}$, one for each part. We

[1]Unlike SCAPE, SP has no need for a least-squares optimization to stitch triangles into a coherent mesh. This has significant computational advantages and results from the fact that part shapes are defined in their own coordinate systems.

[2]The data was obtained from http://mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.



a b c d e

Figure 2. **Stitched Puppet model**. To generate an SP we start with the template body (a), which is segmented into parts. To each body part, defined in a local frame with origin in its center, we apply (b) intrinsic shape deformations, (c) pose-dependent deformations, (d,e) part rotation and translation that bring the parts in a global frame where the 3D model is defined.

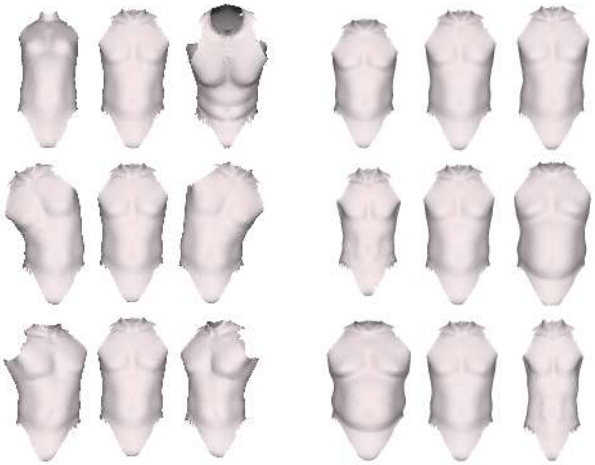


Figure 4. **Part deformation spaces**. The torso PCA models for pose-dependent deformations (left) and intrinsic shape deformations (right) are shown. In both figures: rows correspond to the first three principal components from top to bottom; the center mesh in each row is the mean shape and the left and right meshes correspond to $\pm$ 3 standard deviations from the mean.

use 5 shape basis vectors to represent the pose-dependent deformations of each part, except for the torso which has 12. The range of shapes of the torso varies much more than the other parts due to the flexibility of the spine (Fig. 4, left).

To model intrinsic shape, we apply PCA over the full body, obtaining a single matrix of PCA components, B_s, of size $3\sum_{i=0:15} N_i \times n_s$. For intrinsic shape we use 4 basis vectors for each part. Groups of rows in the shape basis matrix correspond to different body parts and define the PCA component matrix, $B_{s,i}$, for each body part. This approach means that, if each node in SP has the same intrinsic shape coefficients $\mathbf{s}_i$, this corresponds to a coherent body shape. This, however, is not enforced by SP during inference and parts can take on different shapes as needed.

Pairwise potentials. We define the SP model as a tree-structured graphical model with Gaussian potentials. Implicit in the idea of SP are stitching potentials to glue parts

Figure 5. **Stitching parts**. Parts can be thought of as being connected by springs between the interface points. When the model fits together seamlessly, this stitching cost is zero. During inference, the parts can move apart to fit data and then the inference method tries to infer a consistent model.

together. These potentials cannot be learned from the training set, since the training parts are already stitched together. Consequently, we define them manually to allow body parts to be loosely connected (Fig. 5); cf. [34]. We define the stitching potentials as a weighted sum of squared distances between the interface points of adjacent parts: $\Psi_{ij}(\mathbf{x}_i, \mathbf{x}_j) =$

$$\exp(-\frac{1}{\|\mathbf{u}_{ij}\|_1} \sum_{k=1..N_{ij}} u_{ij}(k) \|\tilde{\mathbf{p}}_{i,I_{ij}(k)}(\mathbf{x}_i) - \tilde{\mathbf{p}}_{j,I_{ji}(k)}(\mathbf{x}_j)\|^2).$$

There are N_{ij} interface points between part i and part j. Let $I_{ij}(k)$ and $I_{ji}(k)$ denote the index of the k-th interface point on part i and part j, respectively. $\tilde{\mathbf{p}}_i(\mathbf{x})$ indicates the part points in the global frame after applying parameters $\mathbf{x}$ for each part. $u_{ij}(k)$ defines a stitching weight that is set to 0.8 for points that are allowed to stretch more, like the front of the knee or the back of the elbow, and is 1.0 otherwise; $\mathbf{u}_{ij}$ is a vector of these weights.

Generating instances of the SP model. It is useful to propose bodies during inference and for this we use a simple proposal process, illustrated in Figure 2. We define multivariate Gaussian distributions over the pose-dependent deformation variables and relative rotations of neighboring

Figure 6. **Example SP bodies.** Several bodies generated using SP. Note the realism of the 3D shapes. Note that sampling for inference creates disconnected models that are not stitched.

parts:

$$\Phi_{ij}(\mathbf{r}_{ij}, \mathbf{d}_i, \mathbf{d}_j) = \mathcal{N}(\mathbf{r}_{ij}, \mathbf{d}_i, \mathbf{d}_j; \mu_{ij}, \Sigma_{ij}), \quad (4)$$

where $\mathbf{r}_{ij}$ is the relative rotation of part j with respect to part i, and $\mathbf{d}_i$ and $\mathbf{d}_j$ are PCA coefficients for the pose-dependent deformation models of part i and j, respectively. We learn these functions from the training set with pose variations, for each combination i, j of connected parts.

To generate an instance of the SP model, we first sample a vector of intrinsic shape variables $\mathbf{s}_i$ (we sample with a Gaussian distribution over the PCA coefficients given the variance estimated by PCA). The intrinsic shape variables are replicated for each node and are used to generate body parts for the template mesh with the desired intrinsic shape (Fig. 2(b)). We then sample a vector of pose-dependent deformation variables for the torso. These define the pose of the torso: since the SP torso includes the pelvis, poses in which the torso is bent or twisted with respect to the pelvis are modeled as pose-dependent deformations (Fig. 4, left). We then assign a global rotation and generate the torso mesh in the global frame. Recursively in the tree, starting at torso, for each node i: we get the pose-dependent deformation variables of the parent, $\mathbf{d}_{pa(i)}$; we condition the pairwise Gaussian $\Phi_{pa(i)i}$ with $\mathbf{d}_{pa(i)}$, and marginalize the relative rotation vector $\mathbf{r}_{pa(i)i}$. This gives a Gaussian distribution over $\mathbf{d}_i$; we sample this conditional distribution to get part deformations, and generate the part mesh in the local frame. The effect of the part deformations applied to each body part is shown in Figure 2(c). We finally compute the rotation and translation that stitch the parts together at their interface (Fig. 2(d,e)) using the orthogonal Procrustes algorithm. Figure 6 shows samples of bodies generated using this procedure. Note, this process does not prevent penetration of parts. During inference, when we generate samples, we add noise to the part locations, creating disconnected bodies.

4. Method

Consider the task of aligning SP to a 3D mesh S. We optimize the following energy:

$$E(\mathbf{x}, S) = E_{\text{stitch}}(\mathbf{x}) + E_{\text{data}}(\mathbf{x}, S), \quad (5)$$

where $\mathbf{x} = [\mathbf{x}_0, ..\mathbf{x}_{15}]$ are the model's variables. The energy is the sum of a stitching term and a data term.

Stitching term. The stitching term is the cost for disconnected parts, plus a penalty for penetration, $E_{\text{stitch}}(\mathbf{x}) =$

$$\sum_{i=0..15} \sum_{j \in \Gamma(i)} \alpha_{ij}(-\log(\Psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)) + Q_{ij}(\mathbf{x}_i, \mathbf{x}_j)),$$

where $\Psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ is the stitching potential and Q_{ij} is a penalty for intersecting parts. For the latter, we use a simple test on the location of the part centers, setting Q_{ij} to zero if $||\mathbf{o}_i - \mathbf{o}_j||_2$ is more than 0.05, and 1.0 otherwise. This prevents body parts of similar shape overlapping in 3D. This is important so that two parts do not try to explain the same data. More sophisticated penalty terms could be also used.

Data term. The data term varies depending on the problem and here we consider fitting SP to 3D data. Specifically we fit it to high-resolution scan data and low-resolution visual hull data. We define a matching cost between the 3D model and the 3D data as the distance between model vertices and data vertices, S, plus a penalty for differences in the normal directions:

$$E_{\text{data}}(\mathbf{x}) = \sum_{i=0..15} \beta_i (D_i(\mathbf{x}_i, S) + R_{ij}(\mathbf{x}_i, S)), \quad (6)$$

where

$$D_i(\mathbf{x}_i, S) = \frac{1}{N_i} \sum_{k=1..N_i} (d_{i,k}(S)^2 + b)^\gamma \quad (7)$$

and

$$d_{i,k}(S) = \min_{\mathbf{v}_s \in S} \|\tilde{\mathbf{p}}_{i,k}(\mathbf{x}_i) - \mathbf{v}_s\|_2 \quad (8)$$

is the distance from the model's point $\tilde{\mathbf{p}}_{i,k}(\mathbf{x}_i)$ to the data. We take $b = 0.001$ and $\gamma = 0.45$. The term $R_{ij}(\mathbf{x}_i, S)$ penalizes cases where the normal at a point on the model and the normal at its closest data point have opposite direction. We define $R_{ij}(\mathbf{x}_i, S) = \eta \sum_{k=1..N_i} \mathbb{I}(\theta_{i,k} > \frac{3}{4}\pi)$. Here $\mathbb{I}$ is the indicator function, $\theta_{i,k}$ is the angle between the normal at $\tilde{\mathbf{p}}_{i,k}$ and the normal at $\mathbf{v}_{i,k}$, where $\mathbf{v}_{i,k}$ is the minimizer for $d_{i,k}(S)$ and $\eta = 0.005$.

Note that the energy (Eq. 5) does not include a regularization term over pose parameters. We did not find this necessary for our experiments, and indeed we consider the absence of a pose prior an advantage for estimating unlikely poses. If the data is noisy or highly ambiguous a pose prior could be added to the graphical model.

Optimization. To minimize the energy we use the D-PMP algorithm [26], a particle-based method for MAP estimation in graphical models with pairwise potentials. In contrast to Markov Chain Monte Carlo methods, where particles represent distributions, in D-PMP particles represent locations of modes of the posterior. This implies that, even if the model is very high dimensional, it is not necessary

to use a large number of particles. D-PMP is an iterative method where BP is applied at each iteration over the set of particles. At each iteration, particles are resampled with the aim of creating new particles in better locations. A key component of the algorithm is the selection step. During resampling, the number of particles is doubled in order to place particles in new locations without removing any of the current ones. Then, a selection step based on preserving the BP messages is applied. During resampling, different strategies can be considered. Typically new particles are created by sampling the prior, with random walks from the current particles, or exploiting data-driven proposals [26].

We initialize particles by generating SP sample bodies with mean intrinsic shape. Each particle represents a body part, and is a vector of node variables. To place the samples in the global frame, we set the position of the torso at the origin, in an upright posture, but with random orientation about the vertical axis. To provide a rough alignment with the input 3D data, we also align it to the origin of the global frame by computing the mean of the input data points. We add a small amount of random noise to the location of each particle, obtaining disconnected sets of body parts. Figure 7 (left) shows the set of initial particles in an example where the optimization uses 30 particles. During optimization, we use an adaptive scheme for assigning the weights α and β in the energy. In a first stage we set the weights in a way that lowers the influence of the distal parts (lower limbs, hands and feet), to which we assign small weights for the stitching and the data terms. In a second stage we increase these weights to bring in more influence from the distal parts. At a final stage we apply a greedy PMP algorithm (also used in [26]), where at each iteration all the particles are resampled with random noise around the current best location. This has the effect of refining the solution.

Resampling and refinement. At each iteration, for each node i in the graphical model (body part), and for each particle $\mathbf{x}_i^{(s)}$, we resample particles as follows. With probability 0.5 we sample a new particle $\hat{\mathbf{x}}_i^{(s)}$ with a random walk from $\mathbf{x}_i^{(s)}$. The sampling is performed over all the node variables or only over the pose-dependent deformation variables $\mathbf{d}_i^{(s)}$, with equal probability. Alternatively, we generate a new particle as a proposal from a neighbor node. First, we select a neighbor j for the node i, then a random particle from node j, $\mathbf{x}_j^{(t)}$. We use $\mathbf{x}_j^{(t)}$ to condition the pairwise Gaussian between node j and node i, $\Phi_{ji}(\mathbf{r}_{ji}, \mathbf{d}_j, \mathbf{d}_i)$. With probability 0.5 the conditioning variables are the pose deformation variables $\mathbf{d}_j^{(t)}$, otherwise we also condition the pairwise Gaussian with a random relative angle uniformly sampled within joints limits. We sample pose-dependent deformation variables from the conditional Gaussian, and obtain $\hat{\mathbf{d}}_i^{(s)}$. We then set the intrinsic shape parameters $\hat{\mathbf{s}}_i^{(s)} = \mathbf{s}_j^{(t)}$. The location and orientation are computed as those that stitch the mesh of $\hat{\mathbf{x}}_i^{(s)}$ to the mesh of the neighbor's particle $\mathbf{x}_j^{(t)}$. After each particle is resampled, we run a few steps of Levenberg-Marquardt (LM) optimization over the location, rotation and pose-deformation parameters to locally improve alignment to the scan data. Since this local optimization is applied only to the generated particle, it has the effect of making it disconnected from its neighbor. The local LM optimization aids convergence and improves accuracy. Figure 8 shows the average alignment error over a subset of the FAUST training set for different numbers of LM iterations; we use 4 iterations below.

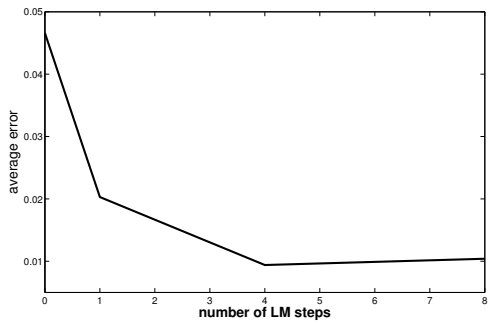


Figure 8. **Local Levenberg-Marquardt (LM) optimization**. Average alignment error for a subset of the FAUST training set plotted for different numbers of local LM optimization steps. We use 4 in our experiments.

5. Experiments

We apply SP to a problem of mesh alignment, specifically to the FAUST challenge [10], and to the problem of estimating shape and pose from 3D meshes reconstructed with shape-from-silhouette. For both experiments we use the same set of parameters for the energy.

Mesh alignment in the FAUST challenge. Mesh registration involves putting two 3D meshes into correspondence, and is a fundamental problem that has received a lot of attention [38]. Previous work focuses on the generic case of two meshes of any object type. Here we consider model-based registration of humans, which is an important problem for building 3D models of humans like the one we develop here. Much of the previous work in human mesh registration assumes known or estimated landmarks [40] or a known pose [4]. In FAUST, ground-truth scan-registration data is generated by accurate texture matching. The goal of the challenge is to find correspondences between scans of people of different gender and shape, in various poses. There are two sub-challenges: the intra-subject requires the alignment of scans of the same person in two different poses. The inter-subject requires alignment across different people, who may be of different gender. Our approach is model-based: we first align all the scans to SP, and

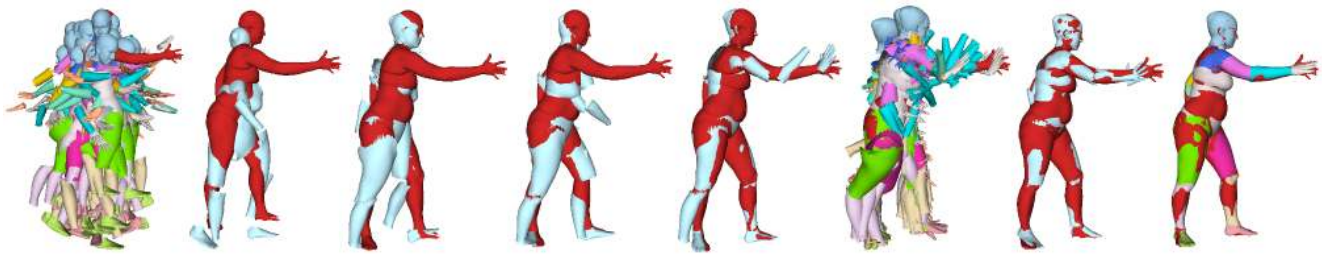


Figure 7. **D-PMP optimization**. Example of inference with 30 particles for 60 iterations. From left to right: initial particles; scan (red) and current best solution (light blue) at iteration 1, 2, 3, 5; the particles at iteration 11; the scan and best solution at iteration 16 and the final set of particles. Note that at the end the particles are all very similar as we run a greedy algorithm that resamples all the particles around the current best solution.

then exploit the model to link vertices across scans, creating the required correspondences. We run our method with 160 particles for 240 iterations, for 3 different random initializations, among which we select the result with lowest energy. Our final results give an average error of 1.57 cm for the intra-challenge and 3.13 cm for the inter-challenge. The authors of the benchmark report average errors on the intra-subject test of 28.3 cm for MÖBius voting [22] and 12.0 cm for Blended Intrinsic Maps (BIM) [21], which are two model-free registration techniques. The methods actually performed worse than this since they did not return any results for 6 and 12 cases, respectively. At the time of this publication no performance numbers are available for model-based methods, apart from the average errors for the method used to build the ground truth data (without the appearance term that was used to create the ground truth). These errors are 0.07 cm and 0.11 cm for the intra-subject and inter-subject tasks, respectively. Figure 11 shows example results, and Figure 10 illustrates an example that produced some of the highest errors. We found that the major source of error is due to self contact. For example, when the hands are touching and vertices of one hand are assigned to the other hand (Fig. 10), this creates large errors where the mesh to be aligned has hands very far away. More examples can be seen on the FAUST webpage [2]. Figure 9 shows the estimated intrinsic shape for different subjects.

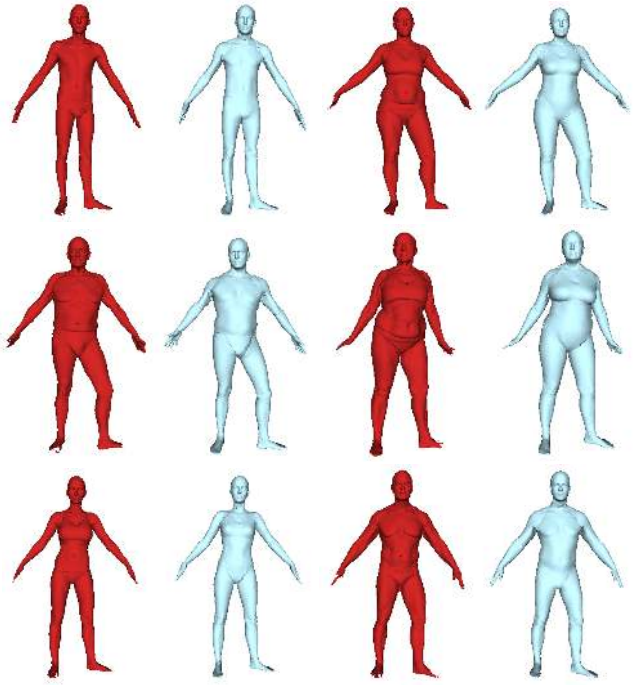


Figure 9. **Intrinsic body shape estimation.** Comparison between scan (red) and estimated intrinsic shape (light blue).

Pose and shape from visual hull. We perform a further experiment on a different type of 3D data to illustrate that our method can deal also with very noisy, approximate, 3D information. We consider a set of frames from the MPI08 dataset [30], from which we extracted the visual hull. We then aligned SP, independently on each frame. We used the same set of parameters we defined for the initial stage of the optimization for the previous experiment, with the difference that we only perform the first stage, as we found that our settings for the refinement stages were decreasing the quality of the solution. This is no surprise given the different quality of the input data. In the data term we did not use the penalty for mismatched normals. We run our method with 200 particles for 120 iterations, for 3 different random initializations, independently for each frame. Figure 12 shows a subset of the results. We show the sequence up to the frame where our model performed correctly. After the last frame the actor turns upside down and our algorithm failed to align to the data, giving preference to a standing position. This is due to our initialization procedure, where we only initialize torso parts in vertical positions. A straightforward solution is to generate samples in any orientation, but this would require a significant increase in the number of particles, with a waste of computation time. One solution would be to track the pose over time. For this experiment there is no ground truth.

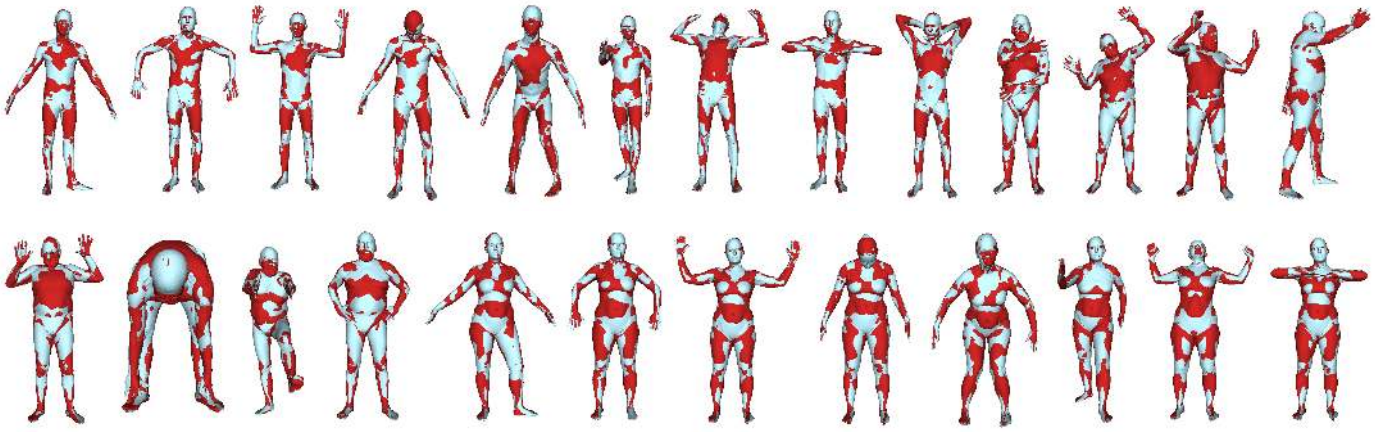


Figure 11. **Alignment on FAUST**. We show the test scan in red and SP in light blue.

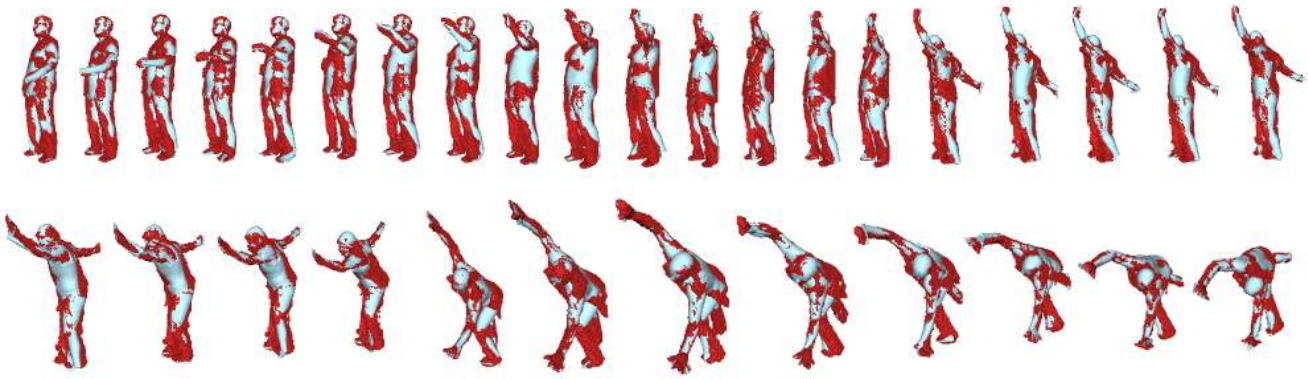


Figure 12. **Alignment to visual hull data**. We show the visual hull data in red and SP in light blue.

Figure 10. **FAUST errors.** One of the worst results, on the intra-subject challenge, is due to a mistake in associating scan points to the model when there is self-contact. (left and middle) Color-coded correspondences between the two poses; note that in the pose on the left, points on the left and right hands are confused. For the pose in the middle, the hands are far apart, thus the resulting correspondence errors are high (right, where red is high error).

6. Conclusions and Future Work

The stitched puppet combines benefits of highly realistic body models like SCAPE with those of part-based graphical models. As such, the SP model provides a bridge between the literature on part-based human pose inference and graphics-like models of human body shape.

Estimating accurate 3D human shape and pose is challenging, even with high-resolution 3D scan data. We demonstrate that SP effectively explores the space of human poses and shapes using a form of non-parametric belief propagation without needing a pose prior. Our results on FAUST and on noisy voxel data suggest that SP can be applied to problems in human pose estimation. To estimate pose and shape from depth data, a loopy version of the model might be necessary to account for body self-occlusion. Note that the solution we find using distributed inference could be used to initialize a refinement stage using the SCAPE model. While we have described SP in the context of human body modeling, the idea can be more widely applied to modeling other animals or man-made shapes. Part deformations need not be learned and adding affine deformations would allow parts to independently "stretch" during inference, fitting a wider range of shapes.

Acknowledgements. We thank Oren Freifeld, Javier Romero and Matt Loper for help with the SCAPE model and Gerard Pons-Moll for the visual hull code.

References

[1] http://stitch.is.tue.mpg.de. 2

[2] http://faust.is.tue.mpg.de. 7

[3] B. Allen, B. Curless, and Z. Popović. Articulated body deformation from range scan data. *ACM Trans. Graph.*, 21(3):612–619, July 2002. 1

[4] B. Allen, B. Curless, and Z. Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Trans. Graph.*, 22(3):587–594, July 2003. 1, 2, 6

[5] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3D human pose estimation. In *Proc. Brit. Mach. Vis. Conf.*, BMVC, Sept. 2013. 2

[6] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 1014–1021, June 2009. 2

[7] 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, July 2005. 1, 2, 3

[8] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures for multiple human pose estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 1669–1679, June 2014. 2

[9] P. Besl and N. D. McKay. A method for registration of 3-d shapes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 14(2):239–256, Feb 1992. 2

[10] F. Bogo, J. Romero, M. Loper, and M. J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 3794 –3801, June 2014. 1, 2, 6

[11] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 8–15, Jun 1998. 2

[12] A. O. Bălan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 1–8, June 2007. 2

[13] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 3618–3625, June 2013. 2

[14] Y. Chen, Z. Liu, and Z. Zhang. Tensor-based human body modeling. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 105–112, June 2013. 1, 2

[15] S. Corazza, L. Mndermann, A. Chaudhari, T. Demattio, C. Cobelli, and T. Andriacchi. A markerless motion capture system to study musculoskeletal biomechanics: Visual hull and simulated annealing approach. *Annals of Biomedical Engineering*, 34(6):1019–1029, 2006. 2

[16] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *Int. J. Comput. Vision*, 61(2):185–205, Feb. 2005. 2

[17] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, Jan. 2005. 2

[18] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 1746–1753, June 2009. 2

[19] P. Guan, A. Weiss, A. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *Proc. IEEE Int. Conf. Comp. Vis.*, ICCV, pages 1381–1388, Oct. 2009. 1

[20] 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, Mar. 2009. 1, 2

[21] V. G. Kim, Y. Lipman, and T. Funkhouser. Blended intrinsic maps. In *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, SIGGRAPH '11, pages 79:1–79:12, New York, NY, USA, 2011. ACM. 7

[22] Y. Lipman and T. Funkhouser. Mobius voting for surface correspondence. In *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, SIGGRAPH '09, pages 72:1–72:12, New York, NY, USA, 2009. ACM. 7

[23] M. M. Loper, N. Mahmood, and M. J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH ASIA)*, 33(6):220:1–220:13, Nov. 2014. 1, 2, 3

[24] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 200(1140):pp. 269–294, 1978. 2

[25] K. Nevatia and T. O. Binford. Structured descriptions of complex objects. In *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, IJCAI'73, pages 641–647, San Francisco, CA, USA, 1973. Morgan Kaufmann Publishers Inc. 2

[26] J. Pacheco, S. Zuffi, M. J. Black, and E. Sudderth. Preserving modes and messages via diverse particle selection. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, volume 32(1), pages 1152–1160, Beijing, China, June 2014. 2, 5, 6

[27] A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(7):730–742, Jul 1991. 2

[28] F. Perbet, S. Johnson, M.-T. Pham, and B. Stenger. Human body shape estimation using a multi-resolution manifold forest. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 668–675, June 2014. 1, 2

[29] R. Plänkers and P. Fua. Articulated soft objects for video-based body modeling. In *Proc. IEEE Int. Conf. Comp. Vis.*, volume 1, pages 394–401, 2001. 2

[30] G. Pons-Moll, A. Baak, T. Helten, M. Müller, H.-P. Seidel, and B. Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 663–670, June 2010. 7

[31] G. Pons-Moll and B. Rosenhahn. *Model-Based Pose Estimation*. Springer, June 2011. 1

[32] J. Rodgers, D. Anguelov, H.-C. Pang, and D. Koller. Object pose detection in range scan data. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR'06, pages 2445–2452, June 2006. 1, 2

[33] H. Sidenbladh, M. J. Black, , and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *Proc. IEEE Euro. Conf. Comp. Vis.*, volume 1843 of *LNCS*, pages 702–718, Dublin, Ireland, June 2000. Springer Verlag. 2

[34] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision*, 98:15–48, 2011. 1, 2, 4

[35] L. Sigal, M. I. Isard, B. H. Sigelman, and M. J. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *Adv. Neur. Inf. Proc. Sys.*, NIPS, pages 1539–1546, Dec. 2003. 1

[36] C. Sminchisescu and A. Telea. Human Pose Estimation from Silhouettes. A Consistent Approach Using Distance Level Sets. In *WSCG International Conference for Computer Graphics, Visualization and Computer Vision*, Czech Republic, 2002. 2

[37] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *Proc. IEEE Int. Conf. Comp. Vis.*, ICCV '11, pages 951–958, Washington, DC, USA, 2011. IEEE Computer Society. 2

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

[39] A. Weiss, D. Hirshberg, and M. Black. Home 3D body scans from noisy image and range data. In *Proc. IEEE Int. Conf. Comp. Vis.*, ICCV, pages 1951–1958, Barcelona, Nov. 2011. IEEE. 1, 2

[40] S. Wuhrer, P. Xi, and C. Shu. Human shape correspondence with automatically predicted landmarks. *Machine Vision and Applications*, 23(4):821–830, 2012. 6

[41] S. Zuffi, O. Freifeld, and M. J. Black. From pictorial structures to deformable structures. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, CVPR, pages 3546–3553. IEEE, June 2012. 2