Inferring 3D Shapes and Deformations from Single Views

Yu Chen, Tae-Kyun Kim, and Roberto Cipolla

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, United Kingdom
{yc301, tkk22, rc10001}@cam.ac.uk

Abstract. In this paper we propose a probabilistic framework that models shape variations and infers dense and detailed 3D shapes from a single silhouette. We model two types of shape variations, the object phenotype variation and its pose variation using two independent Gaussian Process Latent Variable Models (GPLVMs) respectively. The proposed shape variation models are learnt from 3D samples without prior knowledge about object class, e.g. object parts and skeletons, and are combined to fully span the 3D shape space. A novel probabilistic inference algorithm for 3D shape estimation is proposed by maximum likelihood estimates of the GPLVM latent variables and the camera parameters that best fit generated 3D shapes to given silhouettes. The proposed inference involves a small number of latent variables and it is computationally efficient. Experiments on both human body and shark data demonstrate the efficacy of our new approach.

1 Introduction

3D shape estimation from a single image has wide applications for graphics, surveillance, HCI and 3D object recognition. Single view reconstruction is a highly under-constrained problem and requires prior knowledge on 3D shapes of an object class. Various approaches have been investigated with different constraints. While previous methods for general scenes/object categories find it typically hard to capture complex 3D topology of objects, much of recent study has tackled estimating detailed 3D shapes of specific categories, e.g., human faces [11] and body shapes [12–15]. In this work, we propose an approach for both synthesizing and reconstructing dense 3D shapes of general object categories under articulations or deformations given a single image.

1.1 Literature Review

Below we give a brief overview of related work for general scenes/object categories and work designed specifically for human body.

Methods for general scene reconstruction have relied on primitive geometrical constraints such as symmetry and yielded a coarse pop-up reconstruction: e.g., Criminisi et al. [17] have used vanishing points and projective geometry

constraints and Hoiem et al. [2] assumed planar/ground-vertical scenes. Prasad et al. [1] have tackled reconstruction of curved objects, requiring user interactions to reduce down complexity of 3D object topology. Saxena et al. [18] have investigated to recover rough depth estimate from image features. Hassner and Basri [19] have similarly inferred depth from image appearance. 3D geometries having similar image appearance to that of a query object from a database served as the shape prior. These view based methods require an exhaustive number of samples. Some efforts have been made for 3D shape reconstruction from 2D sketches or line drawings [20], where man-made objects are represented by transparent edge-vertex graphs. Bayesian reconstruction of Han et al's [3] is limited to polyhedral objects, tree or grass only. An unified method to segment, infer 3D shapes and recognise object categories proposed in [4] is based on a voxel representation for the shape prior model and applied to object categories such as a cup, mug, plate etc, all rather simple and *rigid* objects. Torresani et al.'s [21] have attempted to recover non-rigid 3D object shape as in our work but only up to sparse reconstruction using 2D point tracks. Their work falls into a different topic, structure-from-motion.

More related study to ours is the work for estimation of detailed human body shape [13–15]. Human body is an articulated object with a number of joint angles. A fixed or deformable crude model based on skeleton, e.g. a cylinder model has been widely exploited for human body pose estimation and tracking. By fitting the model to images, joint angles and a rough 3D shape estimation are obtained, e.g. [6]. Finer body models, e.g. using volumetric representations [7] or generic deformable models [8] have been used to capture more subtle shape variations. These models, however, consider body parts independently and decouple pose from shape variations, therefore not representing shape variations around joints and pose-dependent shape deformations. Recently, a more detailed human model called SCAPE (Shape Completion and Animation for PEople) has been proposed [12]. SCAPE models 3D shape variations among different human bodies in a canonical pose by Principal Component Analysis (PCA), and different poses, i.e. articulation, by joint angles. The shape transfer from a source body to target bodies is obtained by rigid rotations of the 13 body parts manually defined and the pose-dependent deformations for subtle muscular deformation around joints. Balan et al. [13] have adopted this model for the detailed human body shape estimation from silhouettes and formulated the problem as an optimisation over the SCAPE model parameters. However, the optimisation of the SCAPE model is difficult due to uniform priors placed on a number of parameters (joint angles and eigen-coefficients). Stochastic search in [13] is computationally expensive and has initialisation problems. Sigal et al. [14] have used a regression technique to help in initialising the SCAPE model parameters prior to stochastic search and Guan et al. [15] have incorporated more visual cues, the shading cues and internal edges as well as silhouettes to facilitate fitting the SCAPE model to images. Although these methods have shown detailed shape recovery from a few silhouettes, using strong priors on a human body model, i.e.

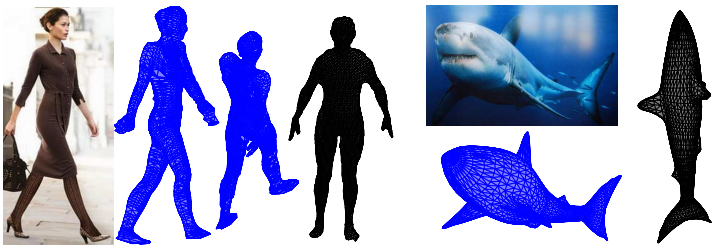

Fig. 1. 3D shape recovery (blue meshes) of a human body (left) and a shark (right) under pose change and their shapes in the canonical pose (gray meshes).

manually defined skeleton and body parts, makes it difficult to extend to other, especially, free-form object categories without redesigning the representation.

1.2 Proposed Approach

In this work, we propose a probabilistic generative model for both learning and inferring dense and detailed 3D shapes of a class of nonrigid objects from a single silhouette. In contrast to prior-arts, we learn shape priors under a challenging setting including pose variations and camera viewpoint changes, and we infer more complex and general deformable 3D shapes from a single image (see Fig. 1).

In our probabilistic framework the shape variations of objects are modeled by two separate Gaussian Process Latent Variable Models (GPLVMs) [22], named the *shape* generator and the *pose* generator. The former captures the *phenotype* variation, which refers to the shape variation between objects: tall vs short, fat vs thin, etc, while the latter captures the *pose* variation, which includes articulation or other nonrigid self-deformation. They are learnt directly from 3D samples without prior knowledge about object class. The GPLVM has been successfully applied for human pose estimation by mapping a high-dimensional parameter space, i.e., a number of joint angles, to a low dimensional manifold [9]. In our work, it nonlinearly maps the complex 3D shape data into a low-dimensional manifold, expressing detailed shape variations only by a few latent variables. With both generators, arbitrary 3D shapes can be synthesized through shape transfer [5], as shown in Fig. 2.

We also propose a novel probabilistic inference algorithm for 3D shape estimation from silhouettes. The shape estimate is obtained by maximum-likelihood estimation of the latent variables of the shape and pose generators and camera parameters that best match generated shapes to input silhouettes. Compared to stochastic optimisation over a large parametric space, i.e. joint angles in [7, 13–15], the proposed inference is computationally efficient as the latent space has a very low dimension. Experiments on articulated human bodies and sharks demonstrate efficacy of the proposed method for reconstructing detailed shapes of general deformable object categories.

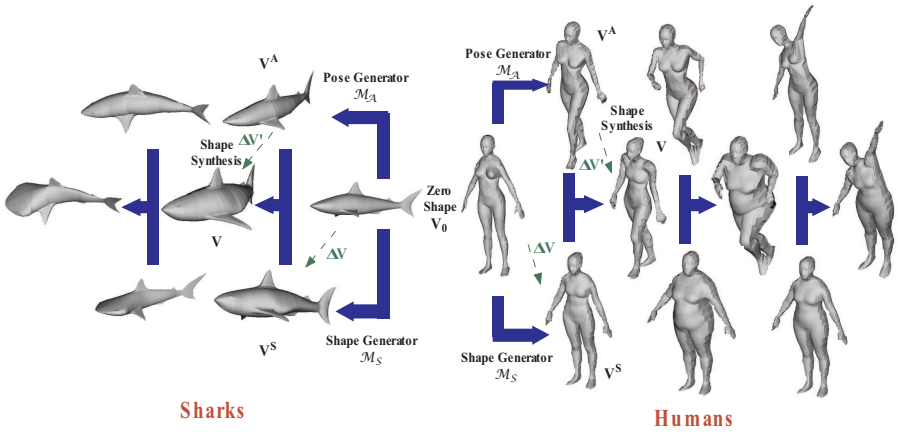

Fig. 2. Synthesizing sharks (left) and human bodies (right) by shape transfer.

The rest of this paper is structured as follows. Section 2 presents the proposed probabilistic model; Section 3 explains learning the shape and pose generator and synthesizing new shapes by the shape transfer; Section 4 presents probabilistic inference algorithm; experimental results are shown in Section 5, and discussions conclusions are drawn in Section 6 and 7 respectively.

2 Probabilistic Model for 3D Shape Estimation

The proposed shape estimation is done by: first, synthesizing 3D shapes from a shape generator $\mathcal{M}_S$ that spans the phenotype variation, and a pose generator $\mathcal{M}_A$ that spans the pose variation; and then, matching the generated shapes with the input silhouette(s). The proposed graphical model is shown in Fig. 3(a). In the formulation, we consider a more general k-views setting. Let $\mathbf{S_k}$ $(k = 1, 2, \cdots, K)$ be the observed silhouettes in K distinct views, which are given in the form of 2D point sets; $\mathbf{V} = [\mathbf{V_1}, \mathbf{V_2}, \cdots, \mathbf{V_N}]$ is a $3N$-D vector which represents the 3D shape with N sampling points on its surface; and $\mathbf{W_k}$ $(k = 1, 2, \cdots, K)$ is the silhouette of $\mathbf{V}$ in the k-th view. The joint distribution can be written as:

$$\begin{aligned} & P(\{\mathbf{S_k}, \mathbf{W_k}\}_{k=1}^K, \mathbf{V}, \mathbf{u}, \mathbf{v} | \{\gamma_\mathbf{k}\}_{k=1}^K, \mathbf{x_A}, \mathbf{x_S}, \mathcal{M}_\mathcal{A}, \mathcal{M}_\mathcal{S}) \\ = & \left(\prod_{k=1}^{K} P(\mathbf{S_k}|\mathbf{W_k}) P(\mathbf{W_k}|\mathbf{V}, \gamma_\mathbf{k}) \right) P(\mathbf{u}|\mathbf{x_A}, \mathcal{M}_\mathcal{A}) P(\mathbf{v}|\mathbf{x_S}, \mathcal{M}_\mathcal{S}) P(\mathbf{V}|\mathbf{u}, \mathbf{v}). \quad (1) \end{aligned}$$

In (1), $\mathbf{x_A}$ and $\mathbf{x_S}$ are the latent coordinates of the corresponding models; $\mathbf{u}$ and $\mathbf{v}$ are the respective latent feature vectors generated by $\mathcal{M}_\mathcal{A}$ and $\mathcal{M}_\mathcal{S}$ at $\mathbf{x_A}$

and $\mathbf{x_S}$; $\gamma_\mathbf{k} = \{\mathbf{P_k}, \mathbf{t_k}\}$ $(k = 1, 2, \cdots, K)$ are the camera parameters of K views. Here, we assume an affine camera model, $\mathbf{P_k}$ is a 3×2 projection matrix and $\mathbf{t_k}$ is a 2×1 translation vector on the image plane. The terms $P(\mathbf{S_k}|\mathbf{W_k})$ and $P(\mathbf{W_k}|\mathbf{V}, \gamma_\mathbf{k})$ $(k = 1, 2, \cdots, K)$ model the matching of 3D shapes $\mathbf{V}$ with the observed silhouettes S_k. The details of inferring shapes from silhouettes will be presented in Section 4. The last three terms $P(\mathbf{u}|\mathbf{x_A}, \mathcal{M}_\mathcal{A})$, $P(\mathbf{v}|\mathbf{x_S}, \mathcal{M}_\mathcal{S})$, and $P(\mathbf{V}|\mathbf{u}, \mathbf{v})$ of (1) model the 3D shape synthesis from the pose generator $\mathcal{M}_\mathcal{A}$ and the shape generator $\mathcal{M}_\mathcal{S}$ given the new latent coordinates $\mathbf{x_A}$ and $\mathbf{x_S}$, which will be presented in detail in Section 3.

3 Shape Generation

3.1 Data Set and Shape Registration

In our approach, the shape generator $\mathcal{M}_\mathcal{S}$ and the pose generator $\mathcal{M}_\mathcal{A}$ are modeled by two independent GPLVMs [22], and trained separately on two data sets, named *shape data set* and *pose data set*. The former contains different shape instances in the canonical pose, while the latter is comprised of various poses of a particular shape instance called *zero shape*.

In order to train the generators, we must build up vertex-wise correspondences among training instances so that we can encode the phenotype variation and pose variation in a vectorized form. For the pose data set, the correspondences are straightforward as all the pose data are generated by animating the same 3D instance in our experiment. Such correspondences are, however, not given for the shape data set and shape registration is required.

In our implementation, every instance in the shape data set is registered with the zero shape in the canonical pose. Firstly, we compute hybrid distances as weighted averages of the spatial distance [24] and the χ^2 distance of the 3D shape contexts [23] between every paired sample points of two shapes, and then use Hungarian algorithm to find the minimal cost matching. Secondly, we use the thin-plate spline (TPS) model to recover point-wise displacements between the pair of shapes using the correspondences established. After this, Principal Component Analysis (PCA) is applied to reduce the dimension of input data before training the pose and shape generators. We use the first $m = 30$ principal components as the pose feature $\mathbf{u}$ and shape features $\mathbf{v}$ for training the GPLVMs.

3.2 Synthesizing New Shapes and Poses from GP

Given the new latent coordinates $\mathbf{x_A}$ and $\mathbf{x_S}$, generating the pose vector $\mathbf{u}$ of the zero shape and the shape vector $\mathbf{v}$ of the canonical pose from $\mathcal{M}_\mathcal{A}$ and $\mathcal{M}_\mathcal{S}$ can be formulated as the following Gaussian predictive likelihoods:

$$\begin{aligned} P(\mathbf{u}|\mathbf{x_A}, \mathcal{M}_\mathcal{A}) &= \mathcal{N}\big(\mathbf{u}; \mathbf{k_U^T}(\mathbf{x_A})\mathbf{K_U^{-1}}\mathbf{Y_A}, (k_U(\mathbf{x_A}, \mathbf{x_A}) - \mathbf{k_U^T}(\mathbf{x_A})\mathbf{K_U^{-1}}\mathbf{k_U}(\mathbf{x_A}))\mathbf{I}\big) \\ &= N\big(\mathbf{u}; \bar{\mathbf{u}}(\mathbf{x_A}), \sigma_A^2(\mathbf{x_A})\mathbf{I}\big) \qquad (2) \\ P(\mathbf{v}|\mathbf{x_S}, \mathcal{M}_\mathcal{S}) &= \mathcal{N}\big(\mathbf{v}; \mathbf{k_V^T}(\mathbf{x_S})\mathbf{K_V^{-1}}\mathbf{Y_S}, (k_V(\mathbf{x_S}, \mathbf{x_S}) - \mathbf{k_V^T}(\mathbf{x_S})\mathbf{K_V^{-1}}\mathbf{k_V}(\mathbf{x_S}))\mathbf{I}\big) \\ &= N\big(\mathbf{v}; \bar{\mathbf{v}}(\mathbf{x_S}), \sigma_S^2(\mathbf{x_S})\mathbf{I}\big). \qquad (3) \end{aligned}$$

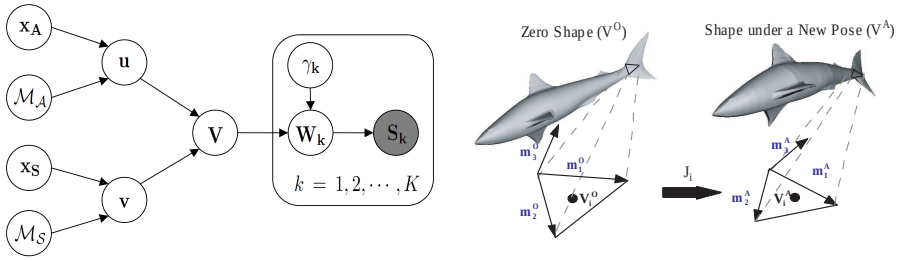

Fig. 3. (a) The graphical model for the 3D shape inference. (b) Transforming local triangle meshes during pose change.

In (2) and (3), $\mathbf{Y_A} = [\mathbf{u_i}]_{i=1}^{N_A}$ and $\mathbf{Y_S} = [\mathbf{v_i}]_{i=1}^{N_S}$ are matrices which contain N_A and N_S training instances in columns for learning $\mathcal{M}_\mathcal{A}$ and $\mathcal{M}_\mathcal{S}$, respectively; $\mathbf{K_U} = [k_U(\mathbf{x_{A,i}}, \mathbf{x_{A,j}})]_{1 \le i \le N_A, 1 \le j \le N_A}$, $\mathbf{K_V} = [k_V(\mathbf{x_{S,i}}, \mathbf{x_{S,j}})]_{1 \le i \le N_S, 1 \le j \le N_S}$, $\mathbf{k_U}(\mathbf{x_A}) = [k_U(\mathbf{x_A}, \mathbf{x_{A,i}})]_{1 \le i \le N_A}$, $\mathbf{k_V}(\mathbf{x_S}) = [k_V(\mathbf{x_S}, \mathbf{x_{S,i}})]_{1 \le i \le N_S}$ are the corresponding non-linear kernel matrices/vectors. In this paper, k_U and k_V are defined as the RBF+linear kernels [9].

3.3 Shape Transfer using Jacobian Matrices

$\mathcal{M}_\mathcal{A}$ or $\mathcal{M}_\mathcal{S}$ only models the shape variation along one of two axes in the shape space. To fully span the shape space, we present a shape synthesis method based on shape transfer in this section.

For the convenience of formulation, we introduce two auxiliary variables $\mathbf{V^A}$ and $\mathbf{V^S}$ to represent the shapes with only the pose variation/phenotype variation imposed, respectively. See Fig. 2. Both of them are $3N$-D vectors, which contain the 3D spatial positions of N sampling vertices of the shape. $\mathbf{V^A}$ and $\mathbf{V^S}$ are recovered from the m-D features $\mathbf{u}$ and $\mathbf{v}$ through linear combinations of the PCA eigen-vectors as: $\mathbf{V^A} = \mathbf{G^A} + \mathbf{A^A u}$ and $\mathbf{V^S} = \mathbf{G^S} + \mathbf{A^S v}$, where $\mathbf{G^A}$ and $\mathbf{G^S}$ are the mean vectors, and $\mathbf{A^A}$ and $\mathbf{A^S}$ are $3N \times m$ matrices containing the first m eigen-vectors of the pose and shape data set, respectively; $\mathbf{V^O}$ denotes the zero-shape in the canonical pose.

The concept of transferring deformation from a source object to target objects has been investigated in the previous work [5]. In our problem, an arbitrary shape $\mathbf{V}$ is synthesized by applying the phenotype variation on the posed zero-shape $\mathbf{V^A}$ locally as follows:

$$\mathbf{V} = \mathbf{V^A} + \mathbf{\Delta V'} + \mathbf{n_V}, \tag{4}$$

where $\mathbf{\Delta V'} = [\mathbf{\Delta V_i}']_{i=1}^{N}$ is a $3N$-D concatenating displacement vector that represents the pose-dependent local shape variation from $\mathbf{V^A}$, and $\mathbf{n}_V$ is an additional random variable modeled by the white Gaussian noise subjected to $\mathcal{N}(0, \sigma_n^2 \mathbf{I_{3N \times 3N}})$. We assume that the vertex-wise phenotype variations $\mathbf{\Delta V_i}$ and $\mathbf{\Delta V_i'}$ before and after the pose change are locally linear transforms as $\mathbf{\Delta V_i} = \mathbf{V_i^S} - \mathbf{V_i^O}$ and $\mathbf{\Delta V_i'} = \mathbf{V_i} - \mathbf{V_i^A}$ (refer to Fig. 2) and they can be related by the

3×3 local Jacobian matrix $\mathbf{J_i}$, similarly to [5]:

$$\mathbf{\Delta V'_i} = \mathbf{J_i} \mathbf{\Delta V_i}. \tag{5}$$

We calculate the local Jacobian matrix at each single sampling vertex approximately from the mesh triangle it belongs to. Given a sampling vertex $\mathbf{V_i^O}$ on the canonical-posed zero-shape (and its corresponding vertex $\mathbf{V_i^A}$ in the new pose), we can find their corresponding the mesh triangles as shown in Fig. 3(b). Two in-plane vectors $\mathbf{m_i^{O,1}}, \mathbf{m_i^{O,2}}$ and one normal vector perpendicular to the triangle plane $\mathbf{m_i^{O,3}}$ are computed for the mesh in the canonical pose and the same $\mathbf{m_i^{A,1}}, \mathbf{m_i^{A,2}}, \mathbf{m_i^{A,3}}$ for the mesh in the new pose. The local Jacobian matrix $\mathbf{J_i}$ can then be computed as:

$$\mathbf{J_i} = [\mathbf{m_i^{A,1}}, \mathbf{m_i^{A,2}}, \mathbf{m_i^{A,3}}][\mathbf{m_i^{O,1}}, \mathbf{m_i^{O,2}}, \mathbf{m_i^{O,3}}]^{-1}. \tag{6}$$

In the training stage, we compute the Jacobian matrix at every sampling point for all the instances of the data set using the method described above. A weighted average filtering over 8 nearest-neighbor sampling points is applied to Jacobian matrices for smoothness. Finally, these matrices are vectorized and used to learn the pose generator $\mathcal{M}_\mathcal{A}$ in junction with the vertex displacements. In the prediction, the elements of Jacobian matrices can thus also be recovered from the pose feature $\mathbf{u}$ using PCA mean $\mathbf{G^J}$ and eigen-vectors $\mathbf{A^J}$ as

$$\mathbf{vec(J)} = \mathbf{G^J} + \mathbf{A^J u}, \tag{7}$$

where $9N$-D vector $\mathbf{vec(J)} = [\mathbf{vec(J_1)}, \mathbf{vec(J_2)}, \cdots, \mathbf{vec(J_N)}]$ is the vectorized-form of matrix $\mathbf{J}$.

3.4 A Probabilistic Model for the Shape Synthesis

The last term $P(\mathbf{V}|\mathbf{u}, \mathbf{v})$ of (1) models the synthesis of new 3D shapes from the pose feature $\mathbf{u}$ and shape feature $\mathbf{v}$, which are generated by GPLVMs in Section 3.2. By combining (4), (5), and (7) the shape synthesis can therefore be formulated as the following equation:

$$\begin{aligned} \mathbf{V} &= \mathbf{V^A} + \mathbf{J} \cdot (\mathbf{V^S} - \mathbf{V^O}) + \mathbf{n_V} \\ &= \mathbf{G^A} + \mathbf{A^A u} + \mathbf{mat}\big(\mathbf{G^J} + \mathbf{A^J u}\big) \cdot \big(\mathbf{G^S} + \mathbf{A^S v} - \mathbf{V^O}\big) + \mathbf{n_V}, \end{aligned} \tag{8}$$

where $\mathbf{J} = diag(\mathbf{J_1}, \mathbf{J_2}, \cdots, \mathbf{J_N})$ is a $3N \times 3N$ matrix, and $\mathbf{mat}(\cdot)$ is an operator which reshapes the $9N \times 1$ vector into a $3N \times 3N$ block diagonal matrix.

We hope to formulate the posterior distribution of the synthesized shape $\mathbf{V}$ explicitly given the latent coordinates $\mathbf{x_A}$ and $\mathbf{x_S}$ of the pose and shape generators $\mathcal{M}_\mathcal{A}$ and $\mathcal{M}_\mathcal{S}$. From the previous subsection, we know that the distributions of $\mathbf{V^A}$, $\mathbf{V^S}$, and $\mathbf{vec(J)}$ have Gaussian form, since they are linearly generated from Gaussian-Process predictions $\mathbf{u}$ and $\mathbf{v}$.

$$\mathbf{V^A}|\mathbf{x_A}, \mathcal{M}_\mathcal{A} \sim \mathcal{N}(\mathbf{V^A}; \mu_{\mathbf{V^A}}(\mathbf{x_A}), \mathbf{\Sigma_{V^A}}(\mathbf{x_A})), \tag{9}$$

$$\mathbf{V^S}|\mathbf{x_S}, \mathcal{M}_\mathcal{S} \sim \mathcal{N}(\mathbf{V^S}; \mu_{\mathbf{V^S}}(\mathbf{x_S}), \mathbf{\Sigma_{V^S}}(\mathbf{x_S})), \tag{10}$$

$$\mathbf{vec(J)}|\mathbf{x_A}, \mathcal{M}_\mathcal{A} \sim \mathcal{N}(\mathbf{vec(J)}; \mu_\mathbf{J}(\mathbf{x_A}), \mathbf{\Sigma_J}(\mathbf{x_A})). \tag{11}$$

where

$$\mu_{\mathbf{V^A}}(\mathbf{x_A}) = \mathbf{G^A} + \mathbf{A^A}\mu_{\mathbf{u}}, \quad \mu_{\mathbf{V^S}}(\mathbf{x_S}) = \mathbf{G^S} + \mathbf{A^S}\mu_{\mathbf{v}}, \quad \mu_{\mathbf{J}}(\mathbf{x_A}) = \mathbf{G^J} + \mathbf{A^J}\mu_{\mathbf{u}},$$
$$\mathbf{\Sigma_{V^A}}(\mathbf{x_A}) = \sigma_u^2\mathbf{A^A}\mathbf{A^{A}}^T, \quad \mathbf{\Sigma_{V^S}}(\mathbf{x_S}) = \sigma_v^2\mathbf{A^S}\mathbf{A^{S}}^T, \quad \mathbf{\Sigma_J}(\mathbf{x_A}) = \sigma_u^2\mathbf{A^J}\mathbf{A^{J}}^T.$$

According to (8), the synthesized shape $\mathbf{V}$ is the product of multi-variate Gaussian $\mathbf{V^S}$ and $\mathbf{J}$, and it is non-Gaussian. However, we find its Gaussian projection $\hat{\mathbf{V}}$ with the same mean and covariance is very good approximation to the true distribution of $\mathbf{V}$, and this projection greatly helps the computation.

$$P(\hat{\mathbf{V}}|\mathbf{x_A}, \mathbf{x_S}, \mathcal{M}_{\mathcal{A}}\mathcal{M}_{\mathcal{S}}) \approx \mathcal{N}\big(\hat{\mathbf{V}}; \mu_{\mathbf{V}}(\mathbf{x_A}, \mathbf{x_S}), \mathbf{\Sigma_V}(\mathbf{x_A}, \mathbf{x_S})\big), \tag{12}$$

where

$$\mu_{\mathbf{V}} = \mu_{\mathbf{V^A}} + \hat{\mu}_{\mathbf{J}}(\mu_{\mathbf{V^S}} - \mathbf{V^O})$$
$$\mathbf{\Sigma_V} = \sigma_n^2\mathbf{I} + \mathbf{\Sigma_{V^A}} + \hat{\mu}_{\mathbf{J}}\mathbf{\Sigma_{V^S}}\hat{\mu}_{\mathbf{J}}^{\mathbf{T}} + \left[\left[Tr\big(\mathbf{\Sigma_J}_{mn}^{ij}\mathbf{S}_{ij}\big)\right]_{m,n=0,1,2}\right]_{i,j=0,1,\cdots,N-1}$$

where $\hat{\mu}_{\mathbf{J}} = \mathbf{mat}(\mu_{\mathbf{J}})$ represents $3N \times 3N$ matrix shape of $\mu_{\mathbf{J}}$ [1]; $\mathbf{S_{ij}} = \mathbf{S}(3i+1 : 3i+3, 3j+1 : 3j+3)$ is the 3×3 sub-matrix of the $3N \times 3N$ matrix $\mathbf{S} = \mathbf{\Sigma_{V^S}} + (\mu_{\mathbf{V^S}} - \mathbf{V^O})(\mu_{\mathbf{V^S}} - \mathbf{V^O})^T$; and $\mathbf{\Sigma_J}_{mn}^{ij} = \mathbf{\Sigma_J}_{(9i+3m+1:9i+3m+3, 9j+3n+1:9j+3n+3)}$ is the 3×3 sub-matrix of the $9N \times 9N$ matrix $\mathbf{\Sigma_J}$.

4 Inferring 3D Shapes from Silhouettes

The matching between the synthesized 3D shapes and input silhouettes is formulated as a two-stage process in our approach. The first stage is the projection stage, which models the procedure of projecting the 3D shape $\mathbf{V}$ into a silhouette $\mathbf{W_k}$ in the k-th view, as shown in (13).

$$P(\mathbf{W_k}|\mathbf{V}, \gamma_{\mathbf{k}}) = \mathcal{N}(\mathbf{W_k}; \tilde{\mathbf{P}}_{\mathbf{k}}\mathbf{V} + \tilde{\mathbf{t}}_k, \sigma_w^2\mathbf{I}), \tag{13}$$

where $\tilde{\mathbf{P}}_{\mathbf{k}} = \mathbf{P_k} \otimes \mathbf{M_k}$ and $\tilde{\mathbf{t}}_k = \mathbf{t}_k \otimes \mathbf{1_{N'}}$ are the expanded version of projection matrix and the offset vector in the k-th view, respectively. Here, $\mathbf{M_k} = [m_{k,ij}]_{1 \le i \le N', 1 \le jleqN}$ is a $N' \times N$ binary masking matrix with element $m_{k,ij} = 1$ if the projection of the i-th 3D sample points is on the boundary and $m_{k,ij} = 0$ otherwise. $\mathbf{M_k}$ selects the N' silhouette points of the projection in the k-th view and it is fully determined by $\mathbf{P_k}$.

The second stage is the matching stage, which models how well the input silhouette $\mathbf{S_k}$ fits the corresponding boundary projection $\mathbf{W_k}$ of the generated shape in the k-th view. The observation likelihood is defined on the basis of Chamfer matching, which provides more robustness to errors and outliers in the input silhouettes as

$$P(\mathbf{S_k}|\mathbf{W_k}) = \frac{1}{Z}\exp\left(-\frac{1}{2\sigma_s^2}DT_{\mathbf{S_k}}^2(\mathbf{W_k})\right), \tag{14}$$

[1] For the convenience of notation, we sometimes omit the parameters of the mean and covariance in the formulation. E.g., $\mu_{\mathbf{J}} = \mu_{\mathbf{J}}(\mathbf{x_A})$

where $DT^2_{\mathbf{S}}(\cdot)$ refers to the squared L2-distance transform of the silhouette $\mathbf{S} = \{\mathbf{s_i}\}_{i=1}^{|\mathbf{S}|}$. For an arbitrary point set $\mathbf{W} = \{\mathbf{w_i}\}_{i=1}^{|\mathbf{W}|}$, it is defined as $DT^2_{\mathbf{S}}(\mathbf{W}) = \frac{1}{2|\mathbf{W}|}\sum_{i=1}^{|\mathbf{W}|}\min_{\mathbf{s_i}\in S}\|\mathbf{w_i}-\mathbf{s_i}\|^2 + \frac{1}{2|\mathbf{S}|}\sum_{j=1}^{|\mathbf{S}|}\min_{\mathbf{w_j}\in\mathbf{W}}\|\mathbf{w_j}-\mathbf{s_j}\|^2$. To simplify the computation, the normalization factor Z is approximated by a constant here.

As stated in the previous section, generating the 3D shapes $\mathbf{V}$ from $\mathcal{M}_{\mathcal{S}}$ and $\mathcal{M}_{\mathcal{A}}$ can be approximately formulated as a Gaussian Process (12). It follows that the silhouette likelihood $P(\mathbf{W_k}|\mathbf{x_A}, \mathbf{x_S}, \mathcal{M}_{\mathcal{A}}, \mathcal{M}_{\mathcal{S}}, \gamma_\mathbf{k})$ also has the Gaussian form by combining (12) with (13):

$$P(\mathbf{W_k}|\mathbf{x_A}, \mathbf{x_S}, \mathcal{M}_{\mathcal{A}}, \mathcal{M}_{\mathcal{S}}, \gamma_\mathbf{k}) = \mathcal{N}\big(\mathbf{W_k}; \mu_{\mathbf{W_k}}(\mathbf{x_A}, \mathbf{x_S}, \gamma_\mathbf{k}), \mathbf{\Sigma}_{\mathbf{W_k}}(\mathbf{x_A}, \mathbf{x_S}, \gamma_\mathbf{k})\big) \quad (15)$$

where $\mu_{\mathbf{W_k}} = \tilde{\mathbf{P}}_\mathbf{k}\mu_\mathbf{V} + \tilde{\mathbf{t}}_\mathbf{k}$ and $\mathbf{\Sigma}_{\mathbf{W_k}} = \tilde{\mathbf{P}}_\mathbf{k}\mathbf{\Sigma}_\mathbf{V}\tilde{\mathbf{P}}_\mathbf{k}^\mathbf{T} + \sigma_w^2\mathbf{I}$.

Our target is to find the 3D shape which best fits all the image evidences $\mathbf{S_k}$ $(k = 1, 2, \cdots, K)$ in K views, or equivalently, to find such latent positions $\mathbf{x_A}$, $\mathbf{x_S}$ and the parameters $\gamma_\mathbf{k}$ of K cameras. This can be done by finding the maximum of the overall likelihood $P(\{S_k\}_{k=1}^K|\mathbf{x_A}, \mathbf{x_S}, \mathcal{M}_{\mathcal{A}}, \mathcal{M}_{\mathcal{S}}, \{\gamma_\mathbf{k}\}_{k=1}^K)$ $(k = 1, 2, \cdots, K)$. The likelihood has no closed form since the direct integral over the terms with distance transform is not tractable, but it can be efficiently optimised by the closed-form lower bound Q [16]:

$$\begin{aligned} &P(\{\mathbf{S_k}\}_{k=1}^K|\mathbf{x_A}, \mathbf{x_S}, \mathcal{M}_{\mathcal{A}}, \mathcal{M}_{\mathcal{S}}, \{\gamma_\mathbf{k}\}_{k=1}^K) \geq Q(\mathbf{x_A}, \mathbf{x_S}, \{\gamma_\mathbf{k}\}_{k=1}^K) \\ &= \prod_{k=1}^{K} \frac{1}{Z_k\sqrt{\det\left(\mathbf{I} + \frac{1}{\sigma_s^2}\mathbf{\Sigma}_{\mathbf{W_k}}\right)}} \exp\big(-\frac{1}{2\sigma_s^2}DT^2_{\mathbf{S_k}}(\mu_{\mathbf{W_k}})\big). \end{aligned} \quad (16)$$

Maximizing the lower bound Q, or equivalently, minimizing $-\log Q$, gives a good approximated maximum-likelihood estimate of the latent coordinate $\mathbf{x_A^{ML}}$, $\mathbf{x_S^{ML}}$, and camera parameters $\gamma_\mathbf{k}^\mathbf{ML}$ $(k = 1, 2, \cdots, K)$:

$$(\mathbf{x_A^{ML}}, \mathbf{x_S^{ML}}, \{\gamma_\mathbf{k}^\mathbf{ML}\}_{k=1}^K) \approx \arg\min_{\mathbf{x_A}, \mathbf{x_S}, \{\gamma_\mathbf{k}\}_{k=1}^K} -\log Q(\mathbf{x_A}, \mathbf{x_S}, \{\gamma_\mathbf{k}\}_{k=1}^K). \quad (17)$$

In our implementation, we minimize $-\log Q$ by adaptive-scale line search and use multiple initializations to avoid local minima. The optimization alternates between finding the latent coordinate ($\mathbf{x_A}$, $\mathbf{x_S}$) and correcting the camera parameters $\{\gamma_\mathbf{k}\}_{k=1}^K$ (and hence the masking matrices $\{\mathbf{M_k}\}_{k=1}^K$). The convergence usually comes fast, as the latent dimensions of GPLVMs are low. Consequently, the corresponding maximum likelihood estimate of the 3D shape can be approximately given as:

$$P(\mathbf{V^{ML}}|\mathbf{x_A^{ML}}, \mathbf{x_S^{ML}}, \mathcal{M}_{\mathcal{A}}\mathcal{M}_{\mathcal{S}}) \approx \mathcal{N}\big(\mathbf{V^{ML}}; \mu_{\hat{\mathbf{V}}}(\mathbf{x_A^{ML}}, \mathbf{x_S^{ML}}), \mathbf{\Sigma}_{\hat{\mathbf{V}}}(\mathbf{x_A^{ML}}, \mathbf{x_S^{ML}})\big) \quad (18)$$

which gives the mean shape $\mu_{\hat{\mathbf{V}}}$ and the uncertainty measurement $\mathbf{\Sigma}_{\hat{\mathbf{V}}}$.

5 Experimental Results

We have investigated two shape categories in the experiments: human bodies and sharks. For the human data, we used Civilian American and European

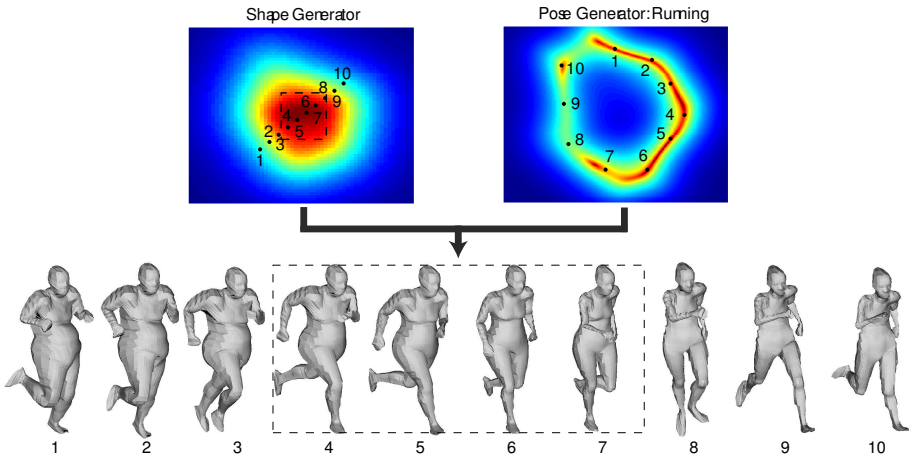

Fig. 4. Generation of new human body shapes in running pose. The shape and pose latent spaces are shown in their first two dimensions. Shapes are spanned by the paired coordinates.

Surface Anthropometry Resource (CAESAR) database as the shape data set, which contains over 2000 different body shapes of North American and European adults in the canonical pose. The pose data set was obtained by synthesizing animations of different 3D poses, e.g, running (150 frames), walking (150 frames), arm stretching and torso movements (250 frames), etc., using the 3D female human model Sydney in Poser 7. For the shark data, the shape data set contains eleven 3D shark models of different shark species available from Internet [19]. For the pose data set, we used an animatable 3D MEX shark model to generate an 11-frame sequence of shark tail-waving motion. The mesh resolution of the zero-shapes are: 3678 vertices/7356 faces for the human data, and 1840 vertices/3676 faces for shark data, respectively. To train $\mathcal{M}_{\mathcal{A}}$ and $\mathcal{M}_{\mathcal{S}}$, we empirically set the latent space dimension $d_S = 6$ for the human shape generator, $d_S = 3$ for the shark shape generator, and $d_A = 2$ for the pose generator for both data sets.

5.1 Shape Synthesis

A direct and important application of our framework is to synthesize a variety of shapes in the category from the shape generator and the pose generator. We visualize the process of synthesizing human shapes in running pose for the latent coordinates of the pose and shape generators in Fig. 4. To examine the synthesis quality, we sampled 10 positions in both the shape and pose latent spaces along the trajectories shown by numbers, and generated the human shapes by pairing up the corresponding shape and pose coordinates. As shown in Fig. 4, a wide-range of body shapes and different stages in the running pose were synthesized. We have also observed that the predictive variances (low variance indicated by red in Fig. 4) imply the quality of shape synthesis. The higher-quality shapes

Data	Precision	Recall
22 sharks	0.8996 ± 0.0481	0.9308 ± 0.0380
20 human bodies	0.7801 ± 0.0689	0.8952 ± 0.0995

Fig. 5. (a) An example of variance estimates of the shark reconstruction; (b) Precision-Recall ratios of the predicted shapes.

(shapes $4-7$ marked by the rectangle) were generated from the low variance area of the shape latent space, where more training samples were presented.

5.2 3D Shape Reconstruction from Images

To verify the efficacy of our 3D shape inference framework, we have tested our approach over 20 human images in tight-fitting clothes and 22 shark images which were collected from Internet. These images involve different camera poses and various object motions, including human running, walking, arm stretching, and shark tail movement. We adopted GrabCut [25] to roughly segment the foreground and extract the corresponding silhouettes. The goal is to infer the reasonable 3D shapes implied by the pictures given the foreground region.

It is worth mentioning that the single-view reconstruction problem is inherently ambiguous. The single silhouette often corresponds to multiple possible 3D shapes mainly due to symmetry and viewpoint changes. Our software generates multiple shape candidates to the silhouette and provides estimate variances for each prediction (Fig 5(a)). For each image, the running time to predict 10 candidates was about 10 - 15 minutes by our unoptimized c++ codes in 2.8GHz PC. In the implementation, we randomly initialised the latent positions of the shape and pose generators. However, we find it helpful to roughly initialise the camera viewpoint. This will speed up the algorithm and greatly increase the possibility of obtaining desired results.

We have evaluated the performance of the approach qualitatively (see Fig. 6 and 7), and quantitatively by the Precision-Recall (P-R) ratios as given in Fig 5(b). Here, the precision and recall are defined as: $Precision = \frac{|S_F \cup S_R|}{S_R}$, and $Recall = \frac{|S_F \cup S_R|}{S_F}$, where S_F denotes the ground-truth foreground and S_R represents the projection of our prediction. All the 3D results provided in Fig. 6 and 7 correspond to the highest likelihood values given the input silhouettes and the shape priors. It shows that our approach captures both phenotype and pose variations and gives accurate estimates on the camera viewpoint. Also, P-R ratios on human data are of reasonable accuracy in comparison with those generated by the human specific model [13], although it is not straightforward to compare quantitatively due to different data sets and number of silhouettes. The reconstructed human bodies are comparatively worse in both visual quality and the P-R ratios than those of sharks because the more complex articulation structure makes exact pose fitting difficult. For example, the pose generator fails to explicitly model the closing hands in the first example of Fig. 7, although the arm and torso poses are well fit (see Section 6 for more discussions).

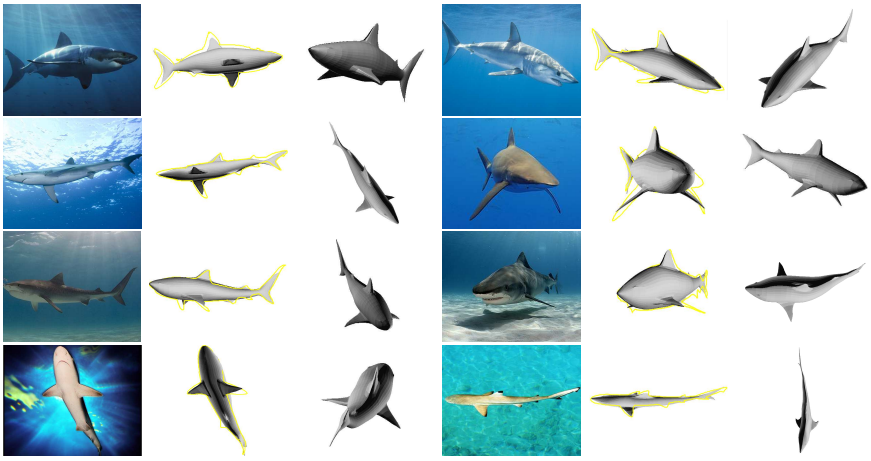

Fig. 6. The qualitative results on shark images. Column 1, 4: input images; Column 2 and 5: the reconstructed shapes in contrast with the input silhouettes; Column 3 and 6: the reconstructed shapes at another viewpoint.

6 Discussion

Compared to previous parametric models [12, 13], the proposed method has both advantages and disadvantages. The benefits include: 1) requiring no strong class-specific priors (parts and skeletons), which facilitates modeling general categories, 2) estimating a much smaller number of model parameters and thus being more efficient, and 3) providing a probabilistic intuition on the uncertainty of shape generation and inference. However, the second benefit could be the drawback at the same time. E.g. whereas the SCAPE allows all possible body configurations by joint angles, our method generates poses similar to those in the pose data set. When training instances are insufficient, the pose generator can be limited in descriptive power (see the first example of Fig. 7). However, the pose generator is easily extendable by more pose data sets and is able to span sufficient local pose variations (the same advocated for pose estimation in [9]).

It is interesting to compare the shape transfer stage in our approach with that in parametric models. In the SCAPE, part-wise rigid rotations matrices and pose-dependent deformation matrices together serve similar functions as Jacobian matrices in our method do but incorporate joint angles. The shape transfer in our method can also benefit when structure priors are available, e.g. Jacobian matrices can be more reliably computed by enforcing part-wise smoothness constraints.

Although our method exploits only silhouettes in the experiments, more visual cues such as shading and internal edges could be used to improve matching accuracy [15]. More direct mapping from silhouettes to shapes could be learnt by regression techniques [14] from the new shapes of new poses synthesized by the proposed model. This would help initialising the proposed inference.

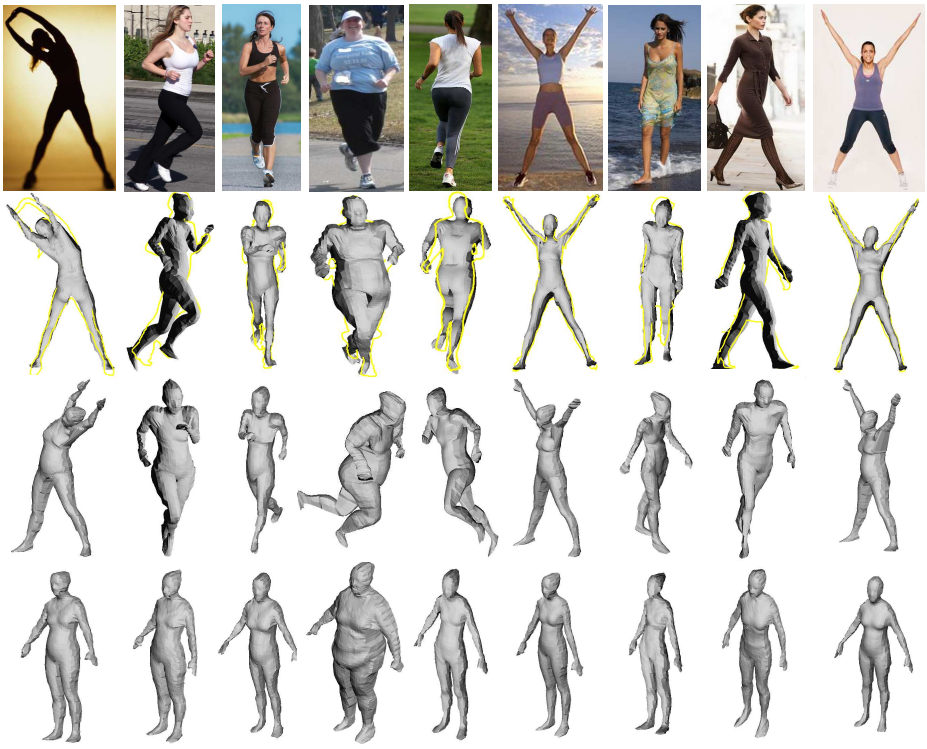

Fig. 7. The qualitative results on human images: Row 1: input images; Row 2: the reconstructed shapes in contrast with the input silhouettes; Row 3: the reconstructed shapes at another viewpoint; Row 4: the body shapes in the canonical pose.

7 Conclusions

In this paper, we have proposed a probabilistic generative method that models 3D deformable shape variations and infers 3D shapes from a single silhouette. The inference in the proposed framework is computationally efficient as it involves a small number of latent variables to estimate. The method is easy to extend to general object categories. It learns and recovers dense and detailed 3D shapes as well as camera parameters from a single image with a little interaction for segmentation. The proposed method can also serve as a good substitution or approximation of a detailed parametric model especially when physical structure of a category is not available.

As future work we shall perform experiments using multiple silhouette inputs for higher precision and extend the framework to incorporate dynamic models for inferring shapes from video sequences. Also, 3D object recognition or action recognition can also be done by the pose-free 3D shape or shape-free pose recovered by the proposed method respectively.

References

1. M. Prasad, A. Zisserman, and A. Fitzgibbon, Single view reconstruction of curved surfaces, CVPR (2006) 1345–1354.
2. D. Hoiem, A. Efros, and M. Hebert, Automatic photo pop-up, SIGGRAPH (2005).
3. F. Han and S. Zhu, Bayesian reconstruction of 3D shapes and scenes from a single image, In Proc. IEEE Int. Workshop on Higher-Level Knowledge (2003).
4. D. Rother and G. Sapiro, Seeing 3D objects in a single 2D image, ICCV (2009).
5. R. Sumner and J. Popovic, Deformation Transfer for Triangle Meshes, SIGGRAPH (2004) 399–405.
6. J. Deutscher and I. Reid, Articulated body motion capture by stochastic search, IJCV, **61** (2) (2004) 185–205.
7. S. Corazza, L. Mundermann, 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 Biomed. Eng., **34** (6) (2006).
8. I. Kakadiaris and D. Metaxas, 3D human model acquisition from multiple views, IJCV, **30** (3) (1998) 191–218.
9. R. Navaratnam, A. Fitzgibbon, and R. Cipolla. Semi-supervised Joint Manifold Learning for Multi-valued Regression, ICCV (2007).
10. M. Salzmann, R. Urtasun, and P. Fua, Local deformation models for monocular 3D shape recovery. CVPR (2008).
11. V. Blanz and T. Vetter. A Morphable model for the synthesis of 3D faces, SIGGRAPH (1999) 187–194.
12. D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, and J. Rodgers, SCAPE: Shape completion and animation of people, SIGGRAPH (2005) 408–416.
13. A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker, Detailed Human Shape and Pose from Images, CVPR (2007).
14. L. Sigal, A. Balan, and M. Black, Combined discriminative and generative articulated pose and non-rigid shape estimation, NIPS (2007).
15. P. Guan, A. Weiss, A. Balan, and M. Black, Estimating human shape and pose from a single image, ICCV (2009).
16. Y. Chen, T-K. Kim, and R. Cipolla, Inferring 3D Shapes and Deformations from Single Views, Technical Report, CUED/F-INFENG/TR.654 (2010).
17. A. Criminisi, I. Reid, and A. Zisserman, Single view metrology, IJCV, **40** (2) (2000) 123–148.
18. A. Saxena, S. Chung, and A. Ng, 3-D depth reconstruction from a single still image, IJCV **76** (1) (2008) 53–69.
19. T. Hassner and R. Basri, Example based 3D reconstruction from single 2D images, Beyond Patches Workshop at CVPR (2006) 15.
20. J. Liu, L. Cao, Z. Li, and X. Tang, Plane-based optimization for 3D object reconstruction from single line drawings, IEEE Trans. PAMI **30** (2) (2008) 315–327.
21. L. Torresani, A. Hertzmann, and C. Bregier, Nonrigid structure-from-motion, Estimating shape and motion with hierachical priors, IEEE Trans. PAMI **30** (5) (2008).
22. N. Lawrence, Gaussian process latent variable models for visualisation of high dimensional data, NIPS **16** (2004) 329–336.
23. S. Belongie, J. Malik, and J. Puzicha, Shape matching and object recognition using shape contexts. IEEE Trans. PAMI **24** (24) (2002) 509–522.
24. P. Besl and N. Mckey, A method for registration of 3-D shapes. IEEE Trans. PAMI **14** (2) (1992) 239–256.
25. C. Rother, V. Kolmogorov, and A. Blake, "GrabCut" – interactive foreground extraction using iterated graph cuts, SIGGRAPH (2004) 309–314.